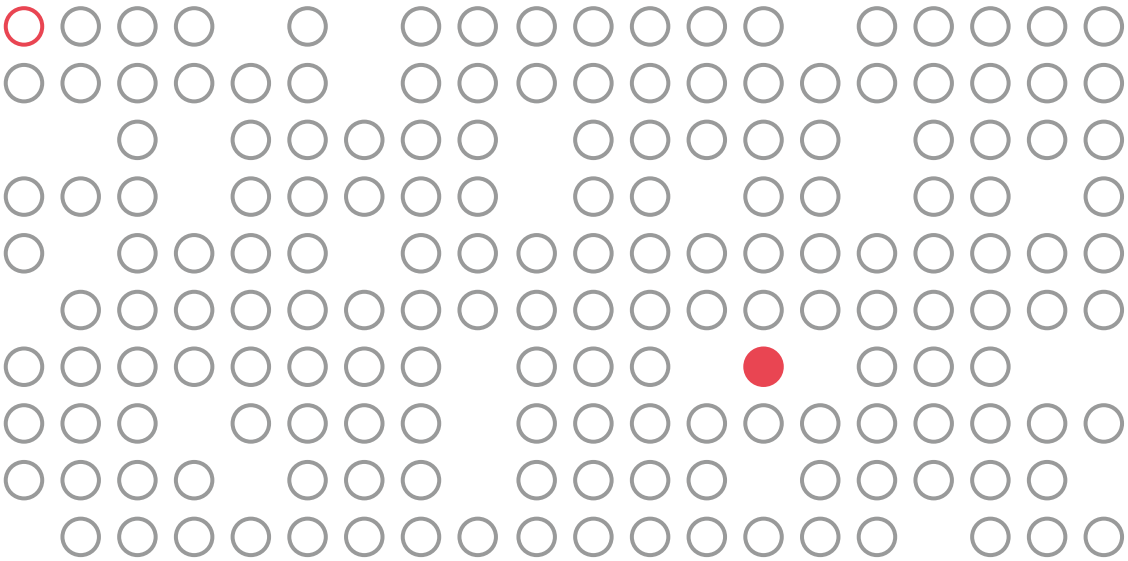
INAUGURAL DISSERTATION 2022

The Impact of Teachers on Children's Human Capital Accumulation

EVIDENCE FROM A DEVELOPING AND A DEVELOPED ECONOMY

María Daniela Araujo Piedra
University of Bamberg



BAMBERG
GRADUATE SCHOOL
OF SOCIAL SCIENCES

The Impact of Teachers on Children's Human Capital Accumulation

Evidence from a Developing and a Developed Economy

Inaugural Dissertation

María Daniela Araujo Piedra

Otto-Friedrich-Universität Bamberg

2022

BAMBERG
GRADUATE SCHOOL
OF SOCIAL SCIENCES

This manuscript has been submitted to the Faculty of Social Sciences, Economics and Business Administration of the Otto-Friedrich-University of Bamberg as a dissertation

First Supervisor:	Prof. Dr. Guido Heineck
Second Supervisor:	Prof. Dr. Bernd Süßmuth
Third Supervisor:	Prof. Dr. Silke Anger
Date of examination:	December 17, 2021

URN: urn:nbn:de:bvb:473-irb-533508
DOI: https://doi.org/10.20378/irb-53350

Acknowledgment

It has been an exciting, interesting and challenging journey. I would like to thank the following people, without whom my doctoral research would never have been possible.

I am particularly grateful to Professor Guido Heineck, my thesis supervisor, for his guidance, encouragement, professional and personal support throughout this project and these years. I am also grateful to Professor Bernd Süßmuth, my second supervisor, for his valuable suggestions and discussions. In addition, I am indebted to Professor Silke Anger for kindly agreeing to be my third supervisor.

I would also like to thank the Bamberg Graduate School of Social Science (BAGSS) for believing in this project and for providing the academic and financial means to develop it. My research fellowship at BAGSS was generously funded by the German Research Foundation (DFG) under the German Excellence Initiative (GSC1024), and by the *Katholischer Akademischer Ausländer-Dienst* (KAAD). I am deeply grateful to these institutions. Thanks are also due to the Inter-American Development Bank (IDB), Ecuador's Ministry of Education and the Leibniz Institute for Educational Trajectories (LIfBi) for providing the data for this research. I also gratefully acknowledge the valuable advice and support provided by my colleagues from the Chair of Empirical Microeconomics and BAGSS at the University of Bamberg.

My utmost thanks to my family for their encouragement, patience and love. I would like to dedicate this work to my daughter Isabel Valentina, who is the light of my life. Thanks to my husband Pablo for his unconditional care and love. Together we crossed the Atlantic, went back to school and started this challenging adventure five years ago. I am deeply grateful to my sister Ximena, who is also my best friend, for always sharing the perfect words and laughs. Thanks to my brother Alberto, who constantly broadens my horizons with his insight and perspectives. Moreover, my very profound gratitude to my parents, Ximena and Iván, who have always been a source of guidance, inspiration and love.

Above all, I want to thank God for providing more than we could ask for during these years, for the lives of the wonderful people who were part of them, and for the exciting journey still ahead of us.

Contents

List of Tables

List of Figures

List of Abbreviations

AM	*Acuerdo Ministerial* (Ministerial Resolution)
ATE	Average Treatment Effect
ATT	Average Treatment Effect on the Treated
BDH	*Bono de Desarrollo Humano* (Human Development Bond)
CASMIN	Comparative Analysis of Social Mobility in Industrial Nations
CIA	Conditional Independence Assumption
CLASS	Classroom Assessment Scoring System
COACTIV	Cognitive Activating Instruction and Development of Students' Mathematics Literacy Study
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
EPF	Education Production Function
FE	Fixed Effects
GPA	Grade Point Average
HDR	Human Development Reports
IDB	Inter-American Development Bank.
ISCED	International Standard Classification of Education
ISEI	International Socio-Economic Index of Occupational Status
KMK	*Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland* (Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany)
LOEI	*Ley Orgánica de Educación Intercultural* (Ecuador's Intercultural Education Law)

MPI	Global Multidimensional Poverty Index
NEPS	German National Educational Panel Study
OECD	Organisation for Economic Co-Operation and Development
OPHI	Oxford Poverty & Human Development Initiative
PIAAC	Programme for the International Assessment of Adult Competencies
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PPVT	Peabody Picture Vocabulary Test
PSM	Propensity Score Matching
PUCE	*Pontificia Universidad Católica del Ecuador* (Pontifical Catholic University of Ecuador)
RE	Random Effects
SC2	Starting Cohort 2
SC3	Starting Cohort 3
SD	Standard Deviation
SIME	*Sistema de Información del Ministerio de Educación* (Ecuador's Ministry of Education Information System)
TALIS	Teaching and Learning International Survey
TVIP	*Test de Vocabularion en Imagenes* (Peabody Picture Vocabulary Test)
UNDP	United Nations Development Programme
US	United States
VERA	*VERgleichsArbeiten*
WAIS-III	Wechsler Adult Intelligence Scale III

Chapter 1 *Introduction*

The human capital accumulation theory, formalized in the pioneering work of Becker (1962, 1964), postulates that education contributes to individuals' development of skills, which increase their productivity and enable them to achieve higher earnings. Subsequently, building on the human capital earnings function developed by Mincer (1970, 1974), a vast number of empirical studies in economics have not only shown that the association between education and earnings exists, but also that there is a causal effect of schooling on earnings (Card, 1999; Harmon, Oosterbeek and Walker, 2003; Heckman, Lochner and Todd, 2006; McMahon, 2017; Heckman, Humphries and Veramendi, 2018; Psacharopoulos and Patrinos, 2018; Gunderson and Oreopolous, 2020; Patrinos and Psacharopoulos, 2020).[1] The average increase in earnings induced by an additional year of schooling, also referred to as the returns to education, has typically ranged from six to fifteen percent in developed countries[2] (Gunderson and Oreopolous, 2020). In developing countries[3], where educational attainment is lower, returns to education tend to be higher (Patrinos and Psacharopoulos, 2020).

The linkage between education and productivity is not limited only to the quantity of schooling an individual attains, however. More recent research lead by Hanushek (Hanushek and Kimko, 2000; Hanushek and Woessmann, 2008) has incorporated learned skills as measures of human capital into the Mincer function, with

[1] A causal estimation of schooling on earnings is challenging because schooling is associated with unobserved cognitive and non-cognitive ability, which is a potential source of bias. Different strategies have been implemented in the causal analysis: proxy measures of ability, twin studies, natural experiments based on education system or policy features, instrumental variables and regression discontinuity designs (Gunderson and Oreopolous, 2020).

[2] Developed countries includes high-income economies in the World Bank's country classification by income level (World Bank, 2021).

[3] Developing countries includes middle- and low-income economies in the World Bank's country classification by income level (World Bank, 2021).

the underlying postulate that human capital is about skills developed in the school and family context, as opposed to merely time spent at school. In this framework, learned skills are typically measured by numeracy or literacy standardized test scores. Applied studies have consistently found a strong and direct impact of learned skills on individual earnings, even once school attainment is taken into account (Hanushek and Zhang, 2009; Hanushek *et al.*, 2015; Hampf, Wiederhold and Woessmann, 2017).[4]

The focus on learned skills as measures of human capital in economics, consequently, has shown that quality differences in education (and not just quantity of schooling) affect skill development, individual productivity, and future earnings (Hanushek and Kimko, 2000; Hanushek and Zhang, 2009; Hanushek and Woessmann, 2011, 2012a; Hoekstra, 2020). This framework has also allowed an alignment of the human capital theory with the long-standing education production function literature (Ben-Porath, 1967; Hanushek, 1979; Todd and Wolpin, 2003), which studies the productivity relationship between school inputs and cognitive achievement outcomes for school-age individuals, once family inputs and innate endowments have been taken into account. Accordingly, it has provided the rationale for estimating and interpreting the long run economic effects of school inputs (Hanushek, 2020).

The aim of this dissertation is to broaden our current knowledge of the effects of teachers and their qualifications, as key school inputs, on the human capital accumulation of kindergarten and primary school children in Ecuador, a developing economy, and Germany, a developed one. Even though these two countries are vastly different in their economic, scientific, social and cultural characteristics, their contexts are similar in that little research has been conducted there on teachers in the framework of the education production function, and particularly of research with causal interpretation.

[4] For instance, using data of the Programme for the International Assessment of Adult Competencies (PIAAC) survey, Hanushek et al. (2015) estimated that on average one standard deviation increase in numeracy skills is associated with an 18 percent earning increase among prime-age workers in a pooled sample of 23 OECD countries. When adding years of schooling, the point estimate on numeracy skills remains significant but goes down to 10.2 percent. Years of schooling is significant and on average associated with 5.9 percent higher earnings, conditional on numeracy skills.

Starting with the early works of Hanushek (1971) and Murnane (1975), the last few decades have seen growing attention paid to the analysis of teachers as the key school input affecting students' human capital accumulation in the education production function framework (Hanushek and Rivkin, 2006; Hanushek, 2011; Hanushek and Rivkin, 2012; Jackson, Rockoff and Staiger, 2014; Koedel, Mihaly and Rockoff, 2015; Strøm and Falch, 2020). The point of departure of these studies is a production function (Todd and Wolpin, 2003):

$$y_{it} = f(X_i(t), S_i(t), T_i(t), \alpha_{io}, \varepsilon_{it}) \tag{1.1}$$

where y_{it} is cognitive achievement for student *i* at time *t,* $X_i(t)$ represents the history of student and family inputs,[5] $S_i(t)$ corresponds to the history of school and classroom inputs,[6] $T_i(t)$ represents the history of teacher inputs,[7] α_{io} denotes student's endowed mental capacity or ability, and ε_{it} is an idiosyncratic error. A fundamental problem in the estimation of the causal effects of individual teachers or their characteristics on student achievement in this framework is the potential matching of students to teachers based on unobserved ability, which is likely to produce bias in the teacher coefficient estimates (Clotfelter, Ladd and Vigdor, 2006). Thus, random assignment of students to classrooms and teachers across time would lead to unbiased Ordinary Least Squares (OLS) estimates of teacher inputs in the education production function framework (Todd and Wolpin, 2003; Guarino, Reckase and Wooldridge, 2015).[8] Nonetheless, even though these experimental settings exist in the literature (Krueger, 1999; Kane and Staiger, 2008; Chetty *et al.*, 2011; Araujo *et al.*, 2016;

[5] Student and family inputs typically included in education production function studies are student race, gender, special-education status, socioeconomic status and parental education (Todd and Wolpin, 2003; Koedel, Mihaly and Rockoff, 2015).

[6] School and classroom characteristics usually accounted for are school size, class size and aggregates of student-level variables (Todd and Wolpin, 2003; Koedel, Mihaly and Rockoff, 2015).

[7] Teacher characteristics commonly studied include gender, experience, education degree, cognitive skills and certification test scores (Hanushek and Rivkin, 2006; Strøm and Falch, 2020).

[8] Most estimations in the literature address the potential non-random assignment of students to schools by relying on within school variation, in other words, by controlling for school fixed effects.

Bacher-Hicks *et al.*, 2019), they are rare, and most researchers have had to rely on administrative data (Strøm and Falch, 2020).

In non-experimental settings, a widely used approach for obtaining unbiased estimates of the effects of individual teachers and their characteristics is the value-added specification of the education production function. This specification differs from the regular education production function in the inclusion of a lagged or baseline achievement score, which is taken to be a sufficient statistic for unobserved input histories, as well as the unobserved endowment of mental capacity or ability (Todd and Wolpin, 2003). The basic linear value-added specification can be represented with the following equation (Todd and Wolpin, 2003; Koedel, Mihaly and Rockoff, 2015):

$$Y_{isjt} = \alpha_o + Y_{isjt-1}\beta_1 + X_{isjt}\beta_2 + S_{isjt}\beta_3 + T_{isjt}\beta_4 + \varepsilon_{isjt} \tag{1.2}$$

where Y_{isjt} is cognitive achievement for student *i* at school *s* with teacher *j* in time *t,* Y_{isjt-1} is a vector of lagged achievement, X_{isjt}, S_{isjt}, T_{isjt} are vectors of contemporary student and family, school and classroom, and teacher characteristics respectively, and ε_{isjt} the idiosyncratic error term. A novel approach within this framework, known as the teacher value-added model, redefines T_{isjt} as a vector of teacher indicator variables or individual teacher effects. The teacher value-added model aims to identify the overall contribution of individual teachers to student achievement in a specific time period, using teacher fixed or random effects (Hanushek and Rivkin, 2012; Jackson, Rockoff and Staiger, 2014; Koedel, Mihaly and Rockoff, 2015). This approach requires longitudinal data linking individual students to individual teachers within schools and across grades.

The literature has also adopted a more restrictive value-added specification of the education production function, also referred to as the gain-score value-added specification, which sets the parameter on the lagged achievement score to one (Todd and Wolpin, 2003; Koedel, Mihaly and Rockoff, 2015):

$$Y_{isjt} - Y_{isjt-1} = \alpha_o + X_{isjt}\beta_2 + S_{isjt}\beta_3 + T_{isjt}\beta_4 + \varepsilon_{isjt} \quad (1.3)$$

Strong assumptions should be met for a consistent OLS estimation of the value-added specification of the education production function in (1.2) and (1.3), as outlined by Todd and Wolpin (2003). First, in the value-added specification (1.2), the effects of all inputs must decay at the same rate – this is often referred to as the common factor restriction. In the gain-score specification (1.3), the effect of each input must be independent of the time at which it was applied, and the effect of ability endowment must likewise be independent of the achievement time. Second, in the value-added specification (1.2), the common factor restriction must also apply to the errors, so they are serially uncorrelated (Todd and Wolpin, 2003; Guarino, Reckase and Wooldridge, 2015). Finally, lagged achievement scores must serve as a good proxy for unobservable individual characteristics (Guarino, Reckase and Wooldridge, 2015). Whether and under which circumstances these assumptions hold for the consistent and unbiased estimation of teacher characteristics or individual teacher effects in the context of the education production function, are questions that are being addressed by the empirical research.[9]

Several studies have compared individual teacher effects obtained in quasi-experimental or experimental settings, where students were randomly assigned to their teachers, with those of non-experimental settings obtained through the application of teacher value-added models. They have consistently found that teacher value-added measures are unbiased predictors of teachers' impacts on student achievement, and that the scope for bias is quite small and statistically insignificant (Kane and Staiger, 2008; Kane *et al.*, 2013; Bacher-Hicks, Kane and Staiger, 2014; Chetty, Friedman and Rockoff, 2014a; Bacher-Hicks *et al.*, 2019). The inclusion of student baseline

[9] Estimation methods commonly applied in the teacher value-added literature include pooled OLS, hierarchical lineal or linear mixed models and the average residual approach (Guarino, Reckase and Wooldridge, 2015; Koedel, Mihaly and Rockoff, 2015). Less commonly, the gain-score value-added specification can be estimated with pooled OLS on the gain score, random effects on the gain score, and fixed effects on the gain score (Guarino, Reckase and Wooldridge, 2015). Finally, the Arellano and Bond Approach, a combination of first differencing and instrumental variables, could be implemented for the estimation (Guarino, Reckase and Wooldridge, 2015).

achievement measures seems to be key to the unbiased estimation of the teacher value-added model (Kane and Staiger, 2008; Chetty, Friedman and Rockoff, 2014a). In addition, empirical evidence has shown that the value-added specification (1.2) is the most robust, and that it even performs better than the gain-score specification (1.3) in the estimation of individual teacher effects (Guarino, Reckase and Wooldridge, 2015; Koedel, Mihaly and Rockoff, 2015). Likewise, the explicit role of lagged achievement as a good proxy for unobservable student characteristics on the right-hand side of the equation seems to be key to a consistent estimation of the teacher effects.

Empirical studies on teacher effectiveness in the education production function framework and the subsequent value-added and teacher value-added to student achievement models have produced some consistent findings. First, there is substantial individual teacher contribution to student achievement and significant variation within this contribution, and this is found in developed and developing countries (Nye, Konstantopoulos and Hedges, 2004; Rockoff, 2004; Rivkin, Hanushek and Kain, 2005; Aaronson, Barrow and Sander, 2007; Kane and Staiger, 2008; Kane, Rockoff and Staiger, 2008; Hanushek and Rivkin, 2010, 2012; Chetty, Friedman and Rockoff, 2014a, 2014b; Jackson, Rockoff and Staiger, 2014; Koedel, Mihaly and Rockoff, 2015; Araujo *et al.*, 2016; Bau and Das, 2020). Second, the evidence regarding the specific teacher characteristics that directly affect student cognitive achievement, or are associated with the individual teacher value-added to student achievement is mostly mixed (Hanushek and Rivkin, 2006, 2012; Glewwe *et al.*, 2014), and seems to be context dependent (Strøm and Falch, 2020).

In developed countries, particularly the United States (US), the teacher characteristics studied include experience, education degree, cognitive skill and certification test scores, and more recently non-commonly observed characteristics such as classroom practices, personality and attitudes. From these characteristics, experience has consistently shown a positive and mostly significant effect on student academic performance; however, this effect is not linear and seems to be more important in the first teaching years (Hanushek, 2003; Hanushek and Rivkin, 2006; Jackson, Rockoff and Staiger, 2014; Papay and Kraft, 2015; Strøm and Falch, 2020). Early research found that teachers' advanced university degrees, such as master's

degrees, were uncorrelated to student achievement (Hanushek, 2003; Hanushek and Rivkin, 2006; Clotfelter, Ladd and Vigdor, 2007a; Chingos and Peterson, 2011); however, new evidence suggests that graduate degrees in the taught subject have a positive effect (Bastian, 2019). At present, evidence on the effects of teacher cognitive skill and certification test scores on student achievement remains mixed (Wayne and Youngs, 2003; Goldhaber and Anthony, 2007; Clotfelter, Ladd and Vigdor, 2007b; Goldhaber, 2007; Angrist and Guryan, 2008; Boyd *et al.*, 2008; Kane, Rockoff and Staiger, 2008; Harris and Sass, 2009, 2011; Rockoff *et al.*, 2011; Goldhaber, Gratz and Theobald, 2017; Hanushek, Piopiunik and Wiederhold, 2019; Strøm and Falch, 2020). By contrast, new experimental research suggests that classroom observation scores are unbiased predictors of teacher effectiveness in terms of their contribution to student achievement (Kane and Staiger, 2012; Bacher-Hicks *et al.*, 2019); nonetheless, this effect depends on the observation instrument used. Finally, it has been found weak and mostly insignificant associations between teachers' effectiveness and individual personality traits evaluated using instruments such as the Big Five Personality test (Rockoff *et al.*, 2011; Jacob *et al.*, 2018), but there is still much to be learned in this area.

In developing countries, research suggests a positive but mostly insignificant effect of teacher education degree or experience on student achievement (Glewwe *et al.*, 2014), and a positive and largely significant effect of teacher skill and content knowledge test scores (Metzler and Woessmann, 2012; Glewwe *et al.*, 2014; Bietenbeck, Piopiunik and Wiederhold, 2018; Bau and Das, 2020). In addition, a pioneer experimental study by Araujo *et al*. (2016) conducted in Ecuador identified a significantly positive association between student cognitive achievement and teacher classroom observation scores, but no significant association with specific personality traits.

Interestingly, research in the education production function framework has also shown that teacher characteristics combined into indices are better predictors of future teacher impact on student achievement than specific individual characteristics (Boyd *et al.*, 2008; Rockoff *et al.*, 2011). Furthermore, more recent evidence from the US suggests that teacher hiring processes that combine teacher background characteristics

and screening measures, such as content knowledge tests, interviews and demonstration classes, can strongly predict future teacher effectiveness (Goldhaber, Grout and Huntington-Klein, 2017; Jacob *et al.*, 2018; Bruno and Strunk, 2019).

The study of teachers as the key school input affecting students' human capital accumulation has proven to be highly relevant for educational policy in both developing and developed countries. On the one hand, of all the education production function inputs directly controlled by policymakers within schools, teachers exhibit the largest and more persistent effects on student achievement gains (Hanushek, 2011). On the other hand, research has shown that more effective teachers, who show persistent positive effects on student achievement, also positively affect student later-life outcomes such as earnings (Chetty *et al.*, 2011; Hanushek, 2011; Hanushek and Rivkin, 2012; Chetty, Friedman and Rockoff, 2014b). In this context, research on teachers, their credential and individual effects has the potential to contribute to maximizing students' current educational- and later economic outcomes by providing evidence on policies that aim to improve teacher workforce effectiveness. Indeed, teacher research in economics has already been used to deliver policy recommendations to improve processes of teacher selection at the hiring stage (Goldhaber, Gratz and Theobald, 2017; Jacob *et al.*, 2018), as well as teacher evaluation, promotion, payment and retention (Hanushek, 2011; Hanushek and Rivkin, 2012; Jackson, Rockoff and Staiger, 2014; Koedel, Mihaly and Rockoff, 2015; Steinberg and Donaldson, 2016).

The literature on teacher effects, nonetheless, requires context and country specific evidence for public policy applications. As noted earlier, found effects of specific teacher characteristics are mixed and largely context dependent. In this regard, my dissertation provides specific and unique evidence of the effects of teachers and their qualifications on students' human capital accumulation in Ecuador and Germany; countries where there is scarce research on teachers in the framework of the education production function and the subsequent value-added and teacher value-added models. Moreover, the purpose of my thesis is to provide scientific evidence that contributes to the permanent improvement of the current teacher recruitment, promotion and retention policies in both countries.

This dissertation consists of three independent scientific articles organized into Chapters (see also table 1.1). In Chapter 2, I evaluate whether Ecuador's new teacher recruitment process, which requires teacher candidates to pass national entrance tests before they are allowed to participate in merit-based competitions for tenure at public schools, has served as an effective screening device and ultimately helped to improve student achievement in the first grades of primary school. To answer these questions, I analyze data from a unique Ecuadorian survey of schools in the academic year 2011-2012, which I match to individual teacher recruitment records from Ecuador's Ministry of Education. I first estimate a value-added to student achievement model using OLS. I then use propensity score matching (PSM) to simulate a random assignment of students to teachers and estimate causal treatment effects. The evidence suggests that teachers awarded tenure through the new policy (*test-screened tenured* teachers) were no more effective overall in improving students' achievement in reading or math over three months of instruction. Nonetheless, in contrast to previous evidence from Ecuador (Cruz-Aguayo, Ibarrarán and Schady, 2017), the OLS and PSM results also suggest that these teachers had a significant positive effect on reading achievement for students from poor households. The PSM estimations show an average treatment effect of about a 9 percent of a standard deviation in reading achievement for students living in poverty. This effect holds even after other teacher credentials such as academic degree and experience are taken into account, which suggests that the standardized exams and classroom practice assessments used in national competitions to award teacher tenured positions helped to differentiate between candidates. Overall, this finding suggests that Ecuador's teacher recruitment policy had a positive impact on the nation's most vulnerable students.

In Chapter 3, Guido Heineck, Yyannú Cruz-Aguayo and I implement an experimental evaluation of the impact of Ecuador's new teacher recruitment process for kindergarten. We first link administrative teacher information from Ecuador's Ministry of Education to data from a unique experimental study where almost 15,000 kindergarten children were randomly assigned to their teachers in the 2012-2013 school year. Then, we use our data to confirm that kindergarten children were indeed assigned randomly to teachers tenured through the new recruitment policy, which allows us to estimate their causal effects on student learning outcomes and prevent potential bias

caused by the matching of teachers to students. Our results show positive and significant effects of *test-screened tenured* teachers of at least an 11 percent of a standard deviation for language, and a 9 percent of a standard deviation for math, over a year of instruction. These effects persist even after controlling for teacher academic degree, experience, cognitive ability, personality traits and classroom practices. In addition, our estimations show that tenured teachers significantly outperform contract teachers in Ecuador, which contrasts with recent evidence on the positive effects of fixed-term contract teachers in other developing countries (Muralidharan and Sundararaman, 2013; Duflo, Dupas and Kremer, 2015). Furthermore, we confirm that the effects of *test-screened tenured* teachers on language learning are stronger for vulnerable children who started the school year with lower baseline test scores or came from socioeconomically disadvantaged families.

The scientific articles presented in Chapters 2 and 3 first contribute to the growing economic literature on teacher effectiveness in the developing world (Glewwe *et al.*, 2020). In addition, they add to the findings of recent research on personnel policies in economics of education, which has shown that combining teacher background characteristics and screening measures in teacher hiring processes can strongly predict future effectiveness (Goldhaber, Grout and Huntington-Klein, 2017; Jacob *et al.*, 2018; Bruno and Strunk, 2019). Moreover, Chapter 3 provides the first experimental evidence of the positive effects of competitive teacher recruitment for a Latin American country. In addition, this research confirms the potential effectiveness of highly qualified teachers in closing learning gaps between socioeconomically advantaged and disadvantaged children in the developing world (Boyd *et al.*, 2008).

Johanna Sophie Quis and I also study the effect of teachers on skill formation in Chapter 4, but in the context of a developed country, Germany. We use primary school data from the German National Educational Panel Study (NEPS) to estimate classroom effects on language and math competence development, which are driven by teachers. We estimate a value-added model with individual classroom fixed- as well as random effects. Both model specifications apply empirical Bayes shrinkage to adjust the classroom effects' estimates by their level of precision. Our results show substantial classroom effects and classroom quality differences in the first grades of German

primary school. One standard deviation increase in classroom effectiveness is associated with at least 14 percent of a standard deviation increase in language competence scores, and at least a 12 percent of a standard deviation increase in student math competence scores. In addition, we find that almost none of the teacher characteristics analyzed, including gender, years of teaching experience, migration background, self-reported *Abitur* GPA, self-reported First State Examination grade, whether the teacher has passed the Second State Examination, constructivist beliefs, and self-reported levels of professional exhaustion, are significantly associated with classroom effectiveness. Remarkably, parental assessment of teacher quality is the only indicator that significantly explains the classroom effects on language competence.

Chapter 4 contributes to the teacher value-added literature in three ways. First, we present the first empirical estimations of classroom effects on language and math competence development in primary school in Germany. Second, our results show that these classroom effects are not associated with characteristics used in teacher recruitment and tenure processes, which is in line with previous findings in the US (Hanushek and Rivkin, 2012; Jackson, Rockoff and Staiger, 2014; Koedel, Mihaly and Rockoff, 2015). However, we find that parents seem to identify more effective teachers and their classrooms for language competence development, adding to the new and growing evidence of the association between parental and student evaluation and teacher quality (Araujo *et al.*, 2016; Bacher-Hicks *et al.*, 2019). Finally, our estimations add to the evidence showing the robustness of teacher and classroom value-added estimates to different settings in developed economies (Koedel, Mihaly and Rockoff, 2015).

Table 1.1: Overview of dissertation

	Chapter 2	Chapter 3	Chapter 4
Title	Measuring the Effect of Competitive Teacher Recruitment on Student Achievement: Evidence from Ecuador	Does Test-Based Teacher Recruitment Work in the Developing World? Experimental Evidence from Ecuador	Parents Can Tell! Evidence on Classroom Quality Differences in German Primary Schools
Data	IDB's survey data and Ecuador's Ministry of Education teacher records	IDB's Closing Gaps data and Ecuador's Ministry of Education teacher records	NEPS SC2
Methods	EPF Value-added Model Propensity Score Matching	EPF Value-added Model Randomization	EPF Teacher Value-added Model Fixed and Random Effects
Co-author(s)		Guido Heineck and Yyannú Cruz-Aguayo	Johanna Sophie Quis
Own contribution	**100%**	**80%**	**80%**

Notes: IDB stands for Inter-American Development Bank. NEPS SC2 refers to the Starting Cohort 2 of the German National Educational Panel Study. EPF stands for Education Production Function.

Chapter 2 *Measuring the Effect of Competitive Teacher Recruitment on Student Achievement: Evidence from Ecuador*

2.1 Introduction

Latin American countries have significantly increased primary and secondary enrolment in the last few decades, but learning outcomes are still substantially lower than in high-income and other middle-income economies (Organisation for Economic Co-Operation and Development [OECD]), 2014, 2016). Research also suggests that the low level of educational achievement in the region has accounted for its slow economic growth relative to the other regions (Hanushek and Woessmann, 2012b). In this context, several Latin American governments have implemented systems for recruiting and promoting teachers based on tests of candidates' knowledge and competences in order to raise teacher and school quality (Bruns and Luque, 2015; Elacqua *et al.*, 2017). Among them is the Ecuadorian government, which has required teacher candidates to pass mandatory tests before they can opt for long-term careers at public schools since 2007. Whether these policies have been effective remains an open question.

There has thus far been no conclusive evidence regarding the teacher characteristics that governments need to identify, select, or enhance to improve teacher and student outcomes. Among the characteristics widely used as a signal of teacher quality for recruitment purposes in high-income economies are teacher cognitive skill and content knowledge of the subject taught, generally measured by certification tests. However, the evidence regarding the effectiveness of teacher test scores as predictors of future quality is still mixed. On the one hand, research from the education production literature suggests that student achievement is positively and significantly associated with teacher certification, knowledge or competency tests (Hill, Rowan and Loewenberg Ball, 2005; Clotfelter, Ladd and Vigdor, 2007b; Goldhaber, 2007;

Goldhaber and Anthony, 2007; Boyd *et al.*, 2008; Rockoff *et al.*, 2011; Goldhaber, Gratz and Theobald, 2017; Jacob *et al.*, 2018; Hanushek, Piopiunik and Wiederhold, 2019). On the other hand, several studies from this literature have pointed out that there is little evidence that the variance in teacher quality is substantially explained by these characteristics (Aaronson, Barrow and Sander, 2007; Angrist and Guryan, 2008; Harris and Sass, 2009). In the developing world, the existing research seems to point out to a positive effect of teacher skill and knowledge of the subject taught on student achievement (Metzler and Woessmann, 2012; Glewwe *et al.*, 2014; Bietenbeck, Piopiunik and Wiederhold, 2018; Bau and Das, 2020).

In this article, I evaluate whether teachers who passed national skill and knowledge tests and were tenured by Ecuador's new competitive recruitment policy (henceforth, *test-screened tenured* teachers) have positive effects on student learning outcomes in the first grades of primary school. In addition, I particularly explore their effects on children who live in poor households given the persistence of significant and substantial differences in learning outcomes between children of high and low socioeconomic status throughout primary school in Ecuador and other Latin-American countries (Schady *et al.*, 2015). These specific effects are also important because the evidence shows that *test-screened tenured* teachers are more likely to work in disadvantaged schools located in poorer areas.

I draw upon a unique survey data of a representative sample of Ecuadorian schools in the academic year 2011-2012, where at least two teaching vacancies were opened and test-screened candidates were selected and subsequently tenured between 2007 and 2011. To identify the effects of *test-screened tenured* teachers, I first estimate a value-added to student achievement model using OLS. Then, I implement a propensity score matching (PSM) approach to simulate a random assignment of students to teachers and estimate causal treatment effects. The evidence suggests that teachers screened and tenured through the new competitive recruitment system were not necessarily more effective in generally raising student achievement in reading or math than their peers at school. Nonetheless, the OLS and PSM results do also suggest that these teachers had a significant positive effect on reading achievement for students from poor households. According to the preferred PSM estimation, the average

treatment effect of being assigned to a *test-screened tenured* teacher is at least a 0.094 standard deviation gain in reading achievement for a student living in poverty, over three months of instruction.

I also explore the effect of teachers who passed national skill and knowledge tests and became eligible candidates allowed to apply to a vacancy in a public school, regardless of whether they have successfully achieved a tenured position. Some of the PSM algorithms implemented suggest that these teachers were also more effective in reading for students living in poverty. However, these results are not as robust as the ones found for *test-screened tenured* teachers. This suggests that differences in skill, subject knowledge and classroom practice scores among *test-screened* teachers might be relevant indicators of quality.

The value-added to student achievement model applied has the potential of providing unbiased estimates of *test-screened tenured* teachers because it takes into account student lagged achievement, which potentially eliminates unobserved family, classroom and school factors from the past, and minimizes student individual specific differences (Chetty, Friedman and Rockoff, 2014a). In addition, the OLS estimation includes school fixed effects that eliminate the bias caused by nonrandom matching to schools. Nonetheless, in order to fully address the potential bias associated with the non-random matching of students to teachers in the sample (Clotfelter, Ladd and Vigdor, 2006; Rothstein, 2009, 2010), I apply different PSM algorithms that balance the probability for a student to be assigned to a *test-screened tenured* teacher based on relevant individual, family, classroom, school and even teacher observed pre-treatment characteristics (Rosenbaum and Rubin, 1983).

This study contributes to the recent work on personnel policies within the economics of education (Goldhaber, Grout and Huntington-Klein, 2017; Jacob *et al.*, 2018; Bruno and Strunk, 2019), as well as the growing literature on teacher quality in the developing world (Glewwe *et al.*, 2020). Specifically, this research offers crucial evidence of positive and significant effects of teachers screened and tenured through Ecuador's new competitive recruitment policy on learning outcomes of students living in poverty, in contrast to earlier findings (Cruz-Aguayo, Ibarrarán and Schady, 2017). These effects are not explained by credentials such as academic degree and experience,

which suggests that national entrance exams helped to screen the best teacher candidates.

In the context of Latin America, this study helps to inform the current debate about the effectiveness of competitive teacher recruitment based on candidates' skill, content-knowledge and classroom practice assessments (Bruns and Luque, 2015; Elacqua *et al.*, 2017). The results suggest that a competitive recruitment policy could lead to improvement in academic outcomes, particularly for socio-economically disadvantaged students.

The remainder of this paper is organized as follows. Section 2.2 presents an overview of teacher recruitment policy reforms in Latin American and impact evaluation attempts, as well as a detail description of the teacher recruitment policy in Ecuador. Section 2.3 describes data sources and reports summary statistics. Section 2.4 presents the OLS estimation strategy and results. The PSM estimation strategy and results are presented in Section 2.5. Some conclusions are drawn in Section 2.6.

2.2 Background and Evidence

2.2.1 Teacher Recruitment Policy Reforms in Latin American

Even though there is no conclusive evidence regarding the impact of teacher certification and screening policies on education quality, several Latin American countries have implemented teacher recruitment systems based on testing candidates' knowledge and competencies in the last two decades (Elacqua *et al.*, 2017). The purpose of these reforms has been to recruit highly qualified teacher candidates for public schools in order to raise quality and equity. For instance, Colombia (since 2002), Ecuador (2007) and Mexico (2008) and Peru (2012) all require candidates to pass national mandatory tests before they can opt for long-term careers at public schools (Bruns and Luque, 2015; Elacqua *et al.*, 2017).

Very little research has been done on the effects of these policies on teacher quality and student outcomes in Latin America. For the Colombian case, Brutti and Sánchez (2017) estimate the impact of the new teacher entry competition on student performance in the national high school exit examination between 2008 and 2013. Starting in 2002, the new teacher entry regulation required candidates to pass national

standardized tests of teaching aptitude and subject knowledge, before they could compete for teaching vacancies. For their identification strategy, Brutti and Sánchez take advantage of the fact that this regulation applied only to newly hired teachers, which created a mix of new-regulation and old-regulation teachers at schools. Their results show that when the share of the teachers who had passed the entrance examinations goes from zero to one, the high school exit exam score increases by about a 0.06 subject standard deviation.

Estrada (2019) evaluates Mexico's teacher hiring reform that required teacher candidates to pass a national standardized test designed to measure cognitive skills, knowledge of the teaching subject, mastery of teaching methods and ethics. The reform was introduced in 2008 to fill some teaching positions and extended in 2013 to all teaching vacancies in public education institutions. Given that the new and traditional recruitment processes coexisted between 2008 and 2013, Estrada is able to compare student outcomes from a group of secondary schools that received only test-hired teachers or only traditionally hired teachers in 2010. Using a difference-in-difference analysis as identification strategy, his results show that moving from traditionally based hires in a school to only test-based hires significantly increases the school's math test score by 0.53 standard deviations and the language score by 0.32 standard deviations. Estrada concludes that test-based hiring might have selected and attracted better applicants, and also might have changed the informal incentives faced by teacher candidates in the traditional recruitment process influenced by union connections.

In Ecuador, Cruz-Aguayo, Ibarrarán and Schady (2017) use data on a sample of children in the first school grades to analyze whether children taught by teachers with higher test scores in the new competitive recruitment, as established in 2007, had higher achievement in language and math in the 2011-2012 school year. They report no indication that teachers with higher (or lower) test scores were assigned to children with different observable characteristics, which allows them to estimate level and value-added to achievement specifications with OLS regressions. Their analysis shows no evidence that the test scores or the aggregate scores on the teacher entry competition predicted child achievement in language or math. Therefore, Cruz-Aguayo et al. come to the conclusion that the evaluation used to make tenure decisions in Ecuador did not

predict how effective teachers were at increasing children's test scores. There are serious limitations in this study, nonetheless. First, it only compares student outcomes among successful teacher candidates who passed entry tests, who therefore belonged to less than 10 percent of all tested teachers. Since test score data from teachers who did not pass entry exams is not available in the study, its findings do not seem to fully support its conclusion. Second, the study does not attempt to compare the outcomes of students whose teachers passed the new selective recruitment process to those whose teachers did not. Consequently, there are still open questions regarding the effectiveness of the screening device of Ecuador's policy. Finally, the study does not take into account possible heterogeneous effects among students, and the changes implemented in the selection process between 2007 and 2011.

This paper addresses the unanswered questions of Cruz-Aguayo, Ibarrarán and Schady (2017). First, I assess the effectiveness of Ecuador's new teacher recruitment process as a quality screening device by incorporating information on teachers who did not pass national entry examinations and were not recruited by the new selective competitions, but were working at the same schools and teaching the same grades during the 2011-2012 school year. Second, I include unique information about the rules applied to each recruitment process to estimate teacher effects by type of competition. Third, I incorporate data on student poverty status in order to examine teacher heterogeneous effects. Finally, I implement a PSM approach to simulate a random assignment of students to teachers and estimate causal treatment effects.

2.2.2 The Teacher Recruitment Policy Reform in Ecuador

Between 2006 and 2017, the Ecuadorian government implemented major education reforms to improve teacher quality (Schneider, Cevallos Estarellas and Bruns, 2019), which had drastically decreased since the 1970s due to the lack of academic standards for pre-service programs and severe decreases in teaching wages relative to other professions (Elacqua *et al.*, 2017). One of the policy reforms introduced was teacher recruitment based on cognitive skill, content-knowledge and classroom practice assessments.

The Ecuadorian government has required teacher candidates to pass national standardized tests before they could participate in merit-based selection competitions

for tenure[10] at public schools since 2007 (Ministerio de Educación del Ecuador, 2007). Ecuador's Education Law of 1990 (*Ley de Carrera Docente y Escalafón del Magisterio Nacional*) established that teachers should be selected through merit-based competitions; however, a national standardized examination for teacher candidates was not implemented until the release of the Executive Order No. 708 in November 2007. [11] The new regulation was applied to teachers seeking tenure in public schools starting in December of the same year. Teachers who were granted tenure before the 2007 regulation were exempted from taking national entrance exams.[12] Moreover, local education authorities were only allowed to hire teachers who had not gone through the new recruitment process temporarily with fixed-term contracts. Teacher vacancies were to be filled permanently by tenured teachers through the merit-based selection competitions.

In March 2011, the National Assembly of Ecuador passed the new Intercultural Education Law (*Ley Orgánica de Educación Intercultural*, LOEI) that ratified national entrance exams as mandatory for teacher candidates (Asamblea Nacional del Ecuador, 2011). The new recruitment rules not only modified the teacher career path that had been in place since 1990, but also removed the discretionary influence of the national teachers' union over teacher selection in Ecuador (Bruns and Luque, 2015; Schneider, Cevallos Estarellas and Bruns, 2019).

Ecuador's Ministry of Education has regulated the new teacher recruitment process through several Ministerial Resolutions (*Acuerdos Ministeriales*, AM). From 2007 to 2009, the recruitment process was regulated by Ministerial Resolution AM No. 438-07 of December 2007, which divided the process into two stages. At the first stage, teacher candidates were required to take a logical-verbal reasoning test, a pedagogical knowledge test and a subject-specific knowledge test that added up to 45 points of the total merit-based competition score, equivalent to 45 percent of the total score. Also at

[10] In Ecuador, tenured teachers hold permanent job positions in public educational institutions. These teachers are civil servants.

[11] Prior to 2007, teacher selection processes were locally organized by Provincial Directorates of Education with no national standards other than academic degree requirements.

[12] Tenured teachers who wanted to be transferred to another school were also required to take these exams and compete for an available position (Ministerio de Educación del Ecuador 2008)

this stage, teachers had to present a demonstration class in front of a school board, which was granted 20 points of the total merit-based competition score, (equivalent to 20 percent of the final evaluation).[13] Teacher candidates had to achieve at least 39 out of the 65 possible points at the first stage in order to become eligible candidates (60 percent). The second stage of the recruitment process was the evaluation of the eligible candidates' credentials. At this stage, scores were assigned to academic degrees, teaching experience, additional training courses and academic publications. The credentials added up to 35 points equivalent to 35 percent of the total merit-based competition score (Ministerio de Educación del Ecuador, 2007). The total score was used to rank teacher candidates who applied to vacancies at public schools. Tenure was awarded to the teacher with the highest score among candidates.

In January 2010, the new teacher recruitment process was adjusted slightly and reorganized into three stages under the Ministerial Resolution AM No. 018-10 (Ministerio de Educación del Ecuador, 2010). At the first stage, teacher candidates were required to take the previously mentioned tests, but not carry out the teaching demonstration. Tests' results added up to 45 points of the total merit-based competition score, equivalent to 45 percent. Accordingly, teacher candidates had to achieve at least 27 points out of 45 possible in order to become eligible candidates for the next stages (60 percent). The second stage again represented the evaluation of the eligible candidates' credentials, but the score granted to teacher experience increased. Credentials therefore now added up to 40 points, or 40 percent of the total merit-based competition score. Finally, the demonstration class was conducted in the third stage and represented 15 points equivalent to 15 percent of the total merit-based competition score.

The LOEI, approved in November 2011, made it harder for teacher candidates to become eligible. The weighting of test results in the competition increased, as well as the minimum scores required to pass the tests (Ministerio de Educación del Ecuador,

[13] The school board comprised the school principal or deputy, a peer teacher and two parents elected by the school's general assembly. For positions in lower and upper secondary education, a student was also included.

2011).[14] The merit-based competition regulations applied between 2007 and 2011 are summarized in table 2.1.

Table 2.1: Regulations and Components of Ecuador's Competitive Teacher Recruitment

	Ministerial Resolutions					
	AM No. 438-07 December 2007		**AM No. 018-10 January 2010**		**AM No. 379-11 November 2011**	
Competition Components	Weight	Use of score	Weight	Use of score	Weight	Use of score
Tests	**45%**		**45%**		**55%**	
– Logical-verbal reasoning	15%	eligibility& ranking	15%	eligibility& ranking	15%	eligibility& ranking
– Pedagogical knowledge	15%	eligibility& ranking	15%	eligibility& ranking	15%	eligibility& ranking
– Subject-specific knowledge	15%	eligibility& ranking	15%	eligibility& ranking	25%	eligibility& ranking
Demonstration Class	**20%**	eligibility& ranking	**15%**	ranking	**10%**	ranking
Teacher Credentials	**35%**		**40%**		**35%**	
– Academic degree	20%	ranking	20%	ranking	20%	ranking
– Training and publications	10%	ranking	10%	ranking	5%	ranking
– Teaching experience	5%	ranking	10%	ranking	10%	ranking
Minimum eligibility threshold	– 60% of eligibility instruments		– 60% of eligibility instruments		– 60% of logical-verbal and pedagogical knowledge tests – 70% of subject-specific knowledge test	
Issued	Dec-07		Jan-10		Nov-11	
Abolished	Jan-10		Nov-11		May-13	

Source: Prepared by the author based on the Ministerial Resolutions AM No 438-07 of December 2007, AM No. 018-10 of January 2010 and AM No. 379-11 of November 2011 issued by Ecuador's Ministry of Education.

Along with the new recruitment regulation, the Ecuadorian government introduced strong economic incentives to attract highly competitive teacher candidates into the public educational system. The LOEI homogenized the teacher payment scale to the public service payment scale, which meant that the real salary of a teacher who

[14] In addition, from July 2013 onwards, teacher tests have been designed by the National Institute of Educational Evaluation (*Instituto Nacional de Evaluación Educativa*, INEVAL) created by the LOEI

started her career in the public sector rose 160 percent between 2006 and 2014 (Elacqua *et al.*, 2017). The economic incentives were granted only to tenured teachers. Therefore, there was great incentive for teachers working at public schools with temporary contracts, teachers working at private schools and recently graduated teachers to engage with the process. Moreover, the LOEI officially opened the teaching career to university graduates who did not hold teaching degrees, but were specialists in subjects taught at public schools.

The Ecuadorian new teacher recruitment process became highly competitive. In December 2012, the Ministry of Education reported that, since 2007, 320.000 teacher candidates had registered for eligibility tests. Meanwhile, 34.250 teaching positions were made available through public selection competitions, 21.200 eligible candidates passed entry tests, and 18.820 successful candidates were granted a permanent teaching position (Ministerio de Educación del Ecuador, 2012).

Ecuador's teacher recruitment process has changed since its inauguration in late 2007, but mandatory teacher tests have consistently acted as a screening device for the process throughout the years. In this paper, I evaluate the effectiveness of the new recruitment process for teachers who were granted tenure between 2007 and 2011 (under the 2007 and 2010 Ministerial Resolutions), and who were working during the 2011-2012 school year.

2.3 Data and Descriptive Statistics

Ecuador is a middle-income country with compulsory schooling from 5 to 15 years of age. The education system is structured in three levels: initial education, general basic education and high school (Asamblea Nacional del Ecuador, 2011). The initial education or early education serves children under 5 years of age (equivalent to level 0 of the International Standard Classification of Education [ISCED]). The general basic education starts at 5 years of age and consists of one year of kindergarten, six years of primary education (ISCED level 1) and three years of lower secondary education (ISCED level 2). High school corresponds to three years of upper secondary education (ISCED level 3).

In 2011, Ecuador's Ministry of Education and the Inter-American Development Bank (IDB) conducted a school survey to analyze the attitudes and pedagogical practices of teachers recruited by the new competitive recruitment process and their impact on educational outcomes. A random sample of 239 primary public schools from the coastal region of Ecuador was chosen. [15] Originally, the sample was drawn from schools that had at least two teachers from the 1st to 3th grades[16] of primary who had received tenure through the new competitive recruitment process. Nonetheless, when schools were actually visited, information from teachers who had not gone through this process was also collected[17] (Pontificia Universidad Católica del Ecuador [PUCE], 2012). Schools were visited twice in the 2011-2012 school year, at end of the second and third quarter.

In total, 476 teachers were interviewed at the 239 primary public schools. Using information from the surveys, variables of teacher characteristics such as gender, years of teaching experience and level of education are generated. With respect to the educational level, a binary variable of whether the teacher had a university degree is also built, because Ecuadorian teachers obtain their teaching degrees from technical institutes[18] as well as from universities.

In addition, information on teachers' recruitment processes was obtained from administrative records of Ecuador's Ministry of Education. The Ministry matched teachers' recruitment process data to the original survey and granted me access to this information. I am able to identify all teachers who received tenure from the start of the merit-based selection competitions to the beginning of the 2011-2012 school year in

[15] Ecuador has four natural regions: Coastal (*Costa*), Andean (*Sierra*), Amazon (*Amazonía*) and Insular (*Islas Galápagos)*. Because of particular weather conditions of each Region, the school year starts at different months. The 2011-2012 school year in the Coastal and Insular Regions stared in April 2011 and ended in January 2012. The same school year in the Andean and Amazon Regions started in September 2011 and ended in June 2012.

[16] The 1st grade of primary school in the ISCED classification is equivalent to the Ecuadorian 2nd grade of Basic Education. Children start this grade when they are around 6 years old.

[17] Researchers found at field that some preselected schools did not have two merit-based recruited teachers at the specific grades, but one. In order to complete the sample, primary teachers in 1st to 3rd grades were randomly chosen at the moment of the survey's application. Thus, teachers tenured throughout the new competitive recruitment process are oversampled in the survey.

[18] Non-university tertiary education.

the coastal region, as well as the specific competition they faced. Approximately a third of teachers surveyed were not awarded tenure under the new competitive selection process at the time of the survey.[19] This additional administrative information obtained from the Ministry of Education makes this data set unique. Although it is similar to that used by Cruz-Aguayo et al. (2017), it incorporates information from teachers who had not been recruited through the new competitive selection process, and differentiates newly tenured teachers by the type of competition they faced. Consequently, my analyses can potentially enable causal effects of the reform to be inferred.

From the interviewed teachers' classes, around 10 students were randomly selected and tested in reading with an adaptation of the Early Grade Reading Assessment (EGRA)[20] (RTI International, 2009b), and in math with an adaptation of the Early Grade Mathematics Assessment (EGMA)[21] (RTI International, 2009a). EGRA is a tool used to measure student progress toward learning to read, and it is administered orally to individual students. EGRA has been adapted for use in more than 65 countries and in over a 100 languages, and can be used for program evaluation purposes (Dubeck and Gove, 2015). Likewise, EGMA is an international assessment of early mathematics learning, with emphasis on numbers and operations. It is also an oral assessment individually administered to students. Internationally, EGMA has been used to measure learning change over time, usually during the course of a program intervention (Platas *et al.*, 2014). In total, 4,520 students completed both assessments of EGRA and EGMA in the second and third quarter of the 2011-2012 school year. Total EGRA and EGMA scores are calculated and standardized by grade.[22] From the

[19] From these teachers, about 51 percent were contract teachers who had not passed entry tests, 36 percent were contract teachers who had passed the tests (eligible teachers) but who had not yet won a merit-based competition, and 13 percent were teachers tenured before the application of the 2007 regulation

[20] The EGRA version applied in Ecuador contained eight tasks or subtests: (1) letter name knowledge; (2) phonemic awareness; (3) letter sound knowledge; (4) familiar word reading; (5) unfamiliar word reading; (6) oral reading fluency with comprehension; (7) listening comprehension; and (8) dictation (PUCE, 2012).

[21] The EGMA adaptation applied in Ecuador had six components: (1) number identification; (2) quantity discrimination; (3) recognition of number patterns (missing number); (4) addition and subtraction; (5) word problems; and (6) geometry (PUCE, 2012).

[22] The EGMA and EGRA Toolkits' guidelines were followed to calculate the scores for each of their subtasks, as well as the global EGMA and EGRA scores for the first and second assessment (RTI International, 2009a, 2009b, 2014, 2016).

student questionnaires, information about student gender, age in years, and class size is also obtained.

The Ministry's survey included a questionnaire applied to students' parents or representatives[23] on their socio-economic context. In total, parents and representatives of 3,937 students were surveyed. From the family questionnaires, I generate a variable of parents' years of education, a binary variable of whether the student attended an early education program, and a binary variable of whether the family receives a monthly cash transfer from the government called a Human Development Bond (*Bono de Desarrollo Humano,* BDH), which is a strong indicator of poverty in Ecuador. [24]

Additional information on school characteristics was provided by principals. The survey has information about: school area (urban or rural), whether the school is multi-grade[25] or complete, enrolment and repetition rates, among others. Multi-grade schools and schools located in rural areas tend to serve impoverished communities in Ecuador. Therefore, they are strong indicators of school socio-economic condition and quality.

The final data set results from matching student, teacher, family and school surveys. I am able to match 4,435 students to their teachers and schools. Of these, 3,661 students (82.55 percent) provide complete family context information.

Table 2.2 reports sample statistics of selected student, teacher and school characteristics for the full sample, as well as whether the student is assigned to a *test-screened tenured* teacher, and whether she comes from a household living in poverty (BDH receiver). On average, reading and math test scores significantly increase

[23] This category included: stepmother, stepfather, grandmother, grandfather, brother, sister, uncle, aunt, other relative, other non-relative.

[24] The Human Development Bond (BDH) is the largest cash transfer program in Latin America for households living in poverty. A poverty score indicator is used to determine households' eligibility for BDH transfers in Ecuador. Information about household composition, education levels, work, dwelling characteristics and access to services is aggregated into the poverty score indicator by principal components. Poverty censuses to gather this information and update the indicator and households' eligibility were conducted in 2000/02, 2007/08 and 2013/14 (Araujo, Bosch and Schady, 2019)

[25] Multi-grade schools are primary schools where teachers have to teach two or more student grades in the same class. Even though multi-grade schools are not prevalent in Ecuador, they can be found in rural and distant communities.

between the second and third quarter of the school year. About half of the student sample are girls, the average student age is 8 years old and about 78 percent attended to an early education program. Children's parents completed around 9 years of education and about 75 percent of students come from BDH households and can be considered poor. With respect to teacher characteristics, almost 90 percent of students have a female teacher, 70 percent are exposed to teachers with university degrees and their teachers' experience is on average 11 years. Approximately 68 percent of the students in the sample are taught by a test-screened tenured teacher. Additionally, about 81 percent of them attend a complete school and 63 percent a rural school.

Looking at students assigned to *test-screened tenured* teachers, table 2.2 shows that they, on average, have lower reading and math test scores in the first and second assessments. However, the reading and math test score gaps between these students and those who were not assigned to *test-screened tenured teachers* are reduced across the two assessment periods, particularly for reading. In addition, students assigned to *test-screened tenured* teachers have parents with fewer years of education, and come from poorer families. The proportion of teachers with a university degree is significantly higher and class size is significantly smaller for this group of students. Nonetheless, they are significantly more likely to attend multi-grade, rural or higher repetition rate schools.

It can also be observed from table 2.2 that students living in poverty have significantly lower reading and math test scores, are older and come from parents with fewer years of education. The proportion of students from poor families assigned to *test-screened tenured* teachers is significantly higher, their class sizes are smaller, and they are also more likely to attend multi-grade, rural and higher repetition rate schools

The descriptive statistics confirm, on the one hand, that *test-screened tenured* teachers tend to serve disadvantaged students whose parents have less year of education and are more likely to live in poverty. On the other hand, they suggest that test-*screened tenured* teachers are also more likely to work in disadvantaged schools located in poorer areas.

Table 2.2: Sample Characteristics

	Full Sample	Test-screened Tenured Teacher			Household is living in poverty (BDH)		
		YES	NO	Difference	YES	NO	Difference
Student:							
Reading t_{-1} (EGRA t_{-1})	142.492	138.189	151.584	-13.395***	137.759	156.410	-18.651***
	(1.481)	(1.785)	(2.633)	(3.181)	(1.702)	(2.953)	(3.409)
Reading t (EGRA t)	171.341	167.716	178.999	-11.283***	166.132	186.659	-20.527***
	(1.446)	(1.759)	(2.526)	(3.078)	(1.677)	(2.792)	(3.257)
Math t_{-1} (EGMA t_{-1})	40.258	39.479	41.903	-2.424***	39.134	43.564	-4.430***
	(0.373)	(0.448)	(0.669)	(0.806)	(0.428)	(0.746)	(0.860)
Math t (EGMA t)	47.056	46.367	48.511	-2.144***	45.956	50.291	-4.335***
	(0.381)	(0.459)	(0.681)	(0.821)	(0.441)	(0.745)	(0.866)
Female	0.492	0.487	0.503	-0.016	0.495	0.484	0.010
	(0.008)	(0.010)	(0.015)	(0.018)	(0.010)	(0.016)	(0.019)
Age (years)	7.942	7.937	7.951	-0.014	7.999	7.773	0.226***
	(0.021)	(0.026)	(0.036)	(0.045)	(0.025)	(0.038)	(0.046)
Attended early education	0.784	0.785	0.782	0.002	0.779	0.798	-0.018
	(0.007)	(0.008)	(0.012)	(0.015)	(0.008)	(0.013)	(0.015)
Family:							
Parents' years of education	8.812	8.530	9.406	-0.876***	8.214	10.568	-2.354***
	(0.061)	(0.073)	(0.111)	(0.133)	(0.065)	(0.132)	(0.147)
Household is poor (BDH)	0.746	0.782	0.670	0.112***			
	(0.007)	(0.008)	(0.014)	(0.016)			
Teacher:							
Female	0.882	0.879	0.889	-0.011	0.870	0.917	-0.047***
	(0.005)	(0.007)	(0.009)	(0.011)	(0.006)	(0.009)	(0.011)
University degree	0.700	0.720	0.656	0.064***	0.693	0.719	-0.026
	(0.008)	(0.009)	(0.014)	(0.017)	(0.009)	(0.015)	(0.017)
Years of experience	10.790	10.707	10.964	-0.257	10.873	10.545	0.328
	(0.104)	(0.118)	(0.207)	(0.238)	(0.120)	(0.208)	(0.240)
Test-screened tenured	0.679				0.712	0.582	0.129***
	(0.008)				(0.009)	(0.016)	(0.018)
Classroom:							
Class size	29.629	27.977	33.119	-5.142***	28.706	32.342	-3.637***
	(0.158)	(0.185)	(0.274)	(0.330)	(0.183)	(0.299)	(0.351)
School:							
Complete	0.812	0.803	0.830	-0.027**	0.791	0.872	-0.081***
	(0.006)	(0.008)	(0.011)	(0.014)	(0.008)	(0.011)	(0.013)
Rural	0.631	0.774	0.329	0.445***	0.693	0.450	0.243***
	(0.008)	(0.008)	(0.014)	(0.016)	(0.009)	(0.016)	(0.019)
Repetition rate (%)	2.057	2.155	1.851	0.304***	2.154	1.771	0.383***
	(0.047)	(0.057)	(0.083)	(0.101)	(0.057)	(0.083)	(0.100)
N Students	3661	2485	1176		2732	929	
N Teachers	461	313	148				
N Schools	239						

Note: This table reports means and standard deviations of student, family, teacher and school characteristics for the full sample, as well as whether the teacher passed national entrance tests and won a competition for tenure (test-screened tenured teacher) and whether the student household is living in poverty (BDH receiver). * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level

In the 2011-2012 academic year, only 22 percent of students at public schools in the coastal region of Ecuador attended a rural school (Ministerio de Educación del Ecuador, 2020), which shows that our sample was disproportionally drawn from rural schools. Accordingly, it also suggests that teacher merit-based competitions for tenured positions were disproportionately opened at more vulnerable schools.

2.4 OLS Estimation Strategy and Results

2.4.1 OLS Estimates of Test-Screened Tenured Teacher Effects

To assess the effectiveness of the new teacher recruitment screening device in Ecuador, I apply a restricted value-added to student achievement specification of the education production function formalized by Todd and Wolpin (2003), but rooted in the longstanding empirical education production literature (Ben-Porath, 1967; Hanushek, 1971, 1979). The value-added model specification differs from the regular education production function only in the inclusion of a lagged (baseline) achievement measure, which is taken to be a sufficient statistic for unobserved input histories (family, classroom and school) as well as the unobserved endowment of mental capacity (Todd and Wolpin, 2003). Accordingly, the model has the potential of providing unbiased education production function estimates, particularly of the effects of teachers and their characteristics on student achievement gains (Chetty, Friedman and Rockoff, 2014a). The restricted value-added model[26] parameters are estimated with the following OLS regression:

[26] A restricted value-added model with a specification that places baseline student achievement (Y_{it-1}) at the left-hand side of the equation (Todd and Wolpin 2003) was chosen because of two methodological reasons. On the one hand, baseline reading and math tests were not administered at the beginning of the school year, before students were exposed to their teachers. In this context, the left-hand side specification does allow minimizing potential problems of endogeneity bias, while still controlling for previous student achievement. On the other hand, there is a very short period of time between the first and second student assessments (three months). Consequently, it is accurate to assume that the student achievement function is non-age-varying, at least over the ages used in implementing the value-added model, as discussed by Todd and Wolpin (2003). OLS estimations of the value-added to student achievement model with a specification that places previous student achievement (Y_{it-1}) at the right-hand side of the equation as a control variable were also conducted for comparison reasons. All results were confirmatory and are available upon request.

$$Y_{isct} - Y_{isct-1} = \alpha_0 + test\_tenured_{cst}\beta_1 + X_{icst}\beta_2 + \bar{X}_{icst}\beta_3 + C_{cst}\beta_4 + T_{cst}\beta_5 + u_{it} \quad (2.1)$$

where the subscripts denote students (*i*), classrooms (*c*), schools (*s*) and time (*t*). The relevant variables or vector of variables are defined as follows: Y_{isct} is achievement of student *i* in year *t* as measured by normalized test scores in reading or math by grade; Y_{isct-1} is previous achievement of the student *i* as measured by normalized test scores in reading or math by grade; α_s is a school fixed component; $test\_tenured_{cst}$ is an indicator of whether the student's current teacher was tenured by the new merit-based competition process; X_{icst} is a vector of measurable student and family characteristics such as gender, age, early education attendance, parents' education and family poverty status; $\bar{X}_{icst}$ is a vector of classroom averages of student and parent characteristics; C_{jmt} is an indicator of class size; and T_{cst} is a vector of additional teacher characteristics such as gender, academic degree and teacher experience.

The main focus of interest is the estimation of β_1 which is identified by the comparison of teachers who passed mandatory standardized tests and were tenured by the new merit-based recruitment process to those who did not. Therefore, it is a measure of the average differential in student achievement between these two types of teachers, holding observed student, family and classroom factors constant within schools. As mention, the potential bias problem in the OLS estimation is addressed first by taking into account student lagged achievement, which eliminates unobserved family, classroom and school factors from the past and minimizes student individual specific differences. In addition, the OLS estimation includes school fixed effects that control for all school specific characteristics, and therefore eliminates the bias caused by nonrandom matching to schools. Standard errors are clustered at the school level, which takes into account cross-classroom correlations in errors within schools (Chetty *et al.*, 2011).

Another potential source of bias in carrying out the analysis described is the non-random matching of students to teachers. Existing research suggests that teachers with stronger credentials tend to be matched to more advantaged students, which is likely to produce upward bias in teacher coefficient estimates (Clotfelter, Ladd and

Vigdor, 2006). The potential bias that comes from non-random matching can be, nonetheless, minimized by estimations of value-added models that control for student's own lagged achievement (Chetty, Friedman, and Rockoff 2014). In addition, the sample descriptive statistics presented in this study suggest that *test-screened tenured* teachers are matched to disadvantaged students and schools in Ecuador. If there are indeed selection processes in the assignment of students to these teachers, we would expect the OLS teacher estimates to be at some extent downward biased.

The estimated effects of *test-screened tenured* teachers on student achievement gains are reported in table 2.3. Columns (1) and (3) present the result of the OLS model estimation for reading and math respectively, with student, family, classroom controls and school fixed effects. In this first estimation I do not take into account other teacher characteristics such as academic degree and experience, because one could argue that these factors already featured in the competitive recruitment process. Columns (2) and (4) present the OLS estimates for reading and math after including controls for teacher characteristics. These additional controls do not substantially change the estimates. All the model specifications show that *test-screened tenured* teachers do not have a statistically significant impact on student learning gains in reading or math, when compared to their peer teachers. Moreover, none of the teacher characteristic have a significant impact on student achievement gains.

The effectiveness of the new competitive teacher recruitment process in Ecuador may vary by student background. Several studies have found that differences in socioeconomic status are strongly associated with variations in children's cognitive and language outcomes. In Ecuador, Schady et al. (2015) found that the differences in language development between children of high and low socioeconomic status are statistically significant, substantial and constant throughout elementary school.[27] Thus, there is an ongoing debate on how much highly qualified teachers can do to close the gap between socioeconomically advantaged and disadvantaged students (Borman and Kimball, 2005; Boyd et al., 2008; Phillips, 2010; Hanushek et al., 2020; James and Wyckoff, 2020).

[27] For example, the difference in language development between children in the richest and poorest quartiles is 1.21 standard deviations in rural Ecuador.

Table 2.3: Estimates of Test-screened Tenured Teacher Effects on Reading and Math Achievement Gains

	Reading		Math	
	(1)	(2)	(3)	(4)
Teacher:				
Test-screened Tenured	0.098	0.106	-0.065	-0.063
	(0.099)	(0.099)	(0.071)	(0.071)
Female		0.005		-0.076
		(0.089)		(0.087)
University degree		0.084		0.072
		(0.070)		(0.068)
Years of experience		0.011		0.008
		(0.016)		(0.014)
Years of experience squared		-0.000		-0.000
		(0.000)		(0.000)
Student:				
Female	0.044	0.045	-0.019	-0.019
	(0.034)	(0.034)	(0.036)	(0.036)
Age (months)	0.034	0.034	-0.047**	-0.047**
	(0.022)	(0.022)	(0.019)	(0.019)
Attended early education	0.112**	0.112**	0.040	0.039
	(0.048)	(0.048)	(0.047)	(0.047)
Family:				
Parents' years of education	0.007	0.007	0.005	0.005
	(0.006)	(0.006)	(0.005)	(0.005)
Household is poor (BDH)	-0.013	-0.013	-0.012	-0.012
	(0.042)	(0.042)	(0.049)	(0.049)
Classroom:				
Class size	0.002	0.001	-0.000	-0.001
	(0.007)	(0.006)	(0.005)	(0.005)
Average age	-0.002	0.002	0.063**	0.063**
	(0.038)	(0.039)	(0.031)	(0.031)
Proportion females	-0.248	-0.196	-0.125	-0.048
	(0.222)	(0.229)	(0.171)	(0.183)
Proportion attended early education	-0.117	-0.090	-0.144	-0.112
	(0.211)	(0.210)	(0.159)	(0.162)
Average parents' education	0.026	0.025	-0.005	-0.005
	(0.026)	(0.026)	(0.024)	(0.024)
Proportion poor household	-0.100	-0.096	0.106	0.099
	(0.273)	(0.275)	(0.230)	(0.232)
Constant	-0.648	-0.866	-0.078	-0.153
	(0.499)	(0.550)	(0.351)	(0.403)
Observations	3661	3661	3661	3661
R^2	0.143	0.143	0.125	0.126

Notes: Each column reports coefficients from OLS regressions estimated with school fixed effects and clustered standard errors (in parentheses) at the school level. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 2.4 presents the effect estimates of *test-screened tenured* teachers on student achievement gains in reading, sorted by student households' poverty condition. Columns (1) and (3) present the result of the OLS estimation with student, family,

classroom controls and school fixed effects, but without controls for other teacher characteristics, for poor and non-poor student households respectively. Columns (2) and (4) report OLS estimates with a full set of controls. We observe a marginally significantly positive effect of *test-screened tenured* teachers on reading achievement gains for students who live in poverty in column (1), which holds even after controlling for other teacher observable characteristics in column (2). For students living in poverty, having a *test-screened tenured* teacher is associated with between a 0.183 and 0.192 standard deviation gain in reading achievement over three months of instruction. By contrasts, no effect is found for students from non-poor households.

Table 2.4: Estimates of Test-screened Tenured Teacher Effects on Reading and Math Achievement Gains by Student Poverty Condition

	Reading				Math			
	Poor Household		Non-poor Household		Poor Household		Non-poor Household	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher:								
Test-screened tenured	0.183*	0.192*	-0.116	-0.135	-0.087	-0.064	-0.143	-0.223
	(0.109)	(0.107)	(0.207)	(0.208)	(0.068)	(0.070)	(0.202)	(0.207)
Female		0.000		-0.068		-0.045		-0.176
		(0.100)		(0.171)		(0.091)		(0.210)
University degree		0.103		-0.067		0.152*		-0.174
		(0.085)		(0.146)		(0.080)		(0.142)
Years of experience		0.006		0.005		0.013		-0.001
		(0.020)		(0.027)		(0.016)		(0.037)
Years of experience squared		-0.000		-0.000		-0.000		-0.001
		(0.001)		(0.001)		(0.000)		(0.001)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	YES	YES	YES	YES	YES	YES	YES	YES
Parent controls	YES	YES	YES	YES	YES	YES	YES	YES
Classroom controls	YES	YES	YES	YES	YES	YES	YES	YES
Observations	2732	2732	929	929	2732	2732	929	929
R^2	0.164	0.165	0.329	0.330	0.144	0.146	0.305	0.312

Notes: Each column reports coefficients from OLS regressions estimated with school fixed effects and clustered standard errors (in parentheses) at the school level. Columns (1)-(8) control for the following student characteristics: age, gender, attendance to early education; parent characteristics: years of education; classroom characteristics: class size, classroom averages of student and parent characteristics. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 2.4 also reports the results for student achievement gains in math sorted by student household poverty condition. There appears to be no significant effect of

test-screened tenured teachers on math learning gains for students who come from poor or non-poor families. The only teacher characteristic that has a positive significant effect on the achievement of students from poor households is whether the teacher has a university degree. None of the teacher characteristics have any significant effect on students from non-poor households.

2.4.2 OLS Estimates of Test-Screened Tenured Teacher Effects, accounting for Ministerial Resolution of Competitions

In our sample around 96.5 percent of the *test-screened tenured* teachers participated in competitions regulated by the original Ministerial Resolution of December 2007 (AM No. 438-07). The remaining 3.5 percent participated in competitions organized under the Ministerial Resolution of January 2010 (AM No. 018-10), which mainly differs from the 2007 Regulation by the timing and weighting of the candidate demonstration class. The impact of each competition might be different because of the changes in the competitions' component weighting or due to test quality differences among competitions[28].

Accordingly, I estimate the effect of *test-screened tenured* teachers on learning outcomes accounting for the Ministerial Resolution that regulated each selection competition. Since there are very few observations from teachers who were granted tenure under the 2010 Regulation, the purpose of these estimations is to isolate the effect of the *test-screened* teachers granted tenure under the 2007 Regulation.[29] I use an extended specification of equation (2.1):

$$Y_{isct} - Y_{isct-1} = \alpha_0 + test\_tenured_{cst}^{AM\ 438-07}\beta_1 + test\_tenured_{cst}^{AM\ 018-10}\beta_2 + X_{icst}\beta_3 + \bar{X}_{icst}\beta_4 + C_{cst}\beta_5 + T_{cst}\beta_6 + u_{it} \quad (2.2)$$

[28] Even though the same type of teacher skill and subject knowledge tests were applied between 2007 and 2012, there is no evidence that they were psychometrically comparable.

[29] Estimates of test-screened tenured teachers that participated in competitions organized under the 2010 Rule cannot be considered robust because of the lack of a sufficient sample and should be considered with caution.

Table 2.5: Estimates of Test-screened Tenured Teacher Effects on Reading and Math Achievement Gains, Accounting for Ministerial Resolution of Competitions

	Reading		Math	
	(1)	(2)	(3)	(4)
Teacher:				
Test-screened tenured	0.130	0.137	-0.055	-0.052
AM 438-07	(0.109)	(0.107)	(0.077)	(0.078)
Test-screened tenured	-0.125	-0.109	-0.137	-0.139
AM 018-10	(0.234)	(0.238)	(0.189)	(0.179)
Female		-0.000		-0.078
		(0.090)		(0.087)
University degree		0.083		0.072
		(0.070)		(0.068)
Years of experience		0.011		0.008
		(0.016)		(0.014)
Years of experience Squared		-0.000		-0.000
		(0.000)		(0.000)
Student:				
Female	0.044	0.045	-0.019	-0.019
	(0.034)	(0.034)	(0.036)	(0.036)
Age (months)	0.034	0.034	-0.047**	-0.047**
	(0.022)	(0.022)	(0.019)	(0.019)
Attended early education	0.112**	0.112**	0.040	0.039
	(0.048)	(0.048)	(0.047)	(0.047)
Family:				
Parents' years of education	0.007	0.007	0.005	0.005
	(0.006)	(0.006)	(0.005)	(0.005)
Household is poor (BDH)	-0.013	-0.013	-0.012	-0.012
	(0.042)	(0.042)	(0.049)	(0.049)
Classroom:				
Class size	0.003	0.002	0.000	-0.001
	(0.007)	(0.006)	(0.005)	(0.005)
Average age	0.000	0.004	0.064**	0.064**
	(0.038)	(0.039)	(0.031)	(0.031)
Proportion females	-0.227	-0.175	-0.118	-0.040
	(0.224)	(0.231)	(0.173)	(0.186)
Proportion attended early education	-0.117	-0.089	-0.144	-0.112
	(0.211)	(0.210)	(0.159)	(0.162)
Average parents' education	0.024	0.023	-0.006	-0.006
	(0.026)	(0.026)	(0.024)	(0.024)
Average poor household	-0.093	-0.091	0.109	0.101
	(0.274)	(0.276)	(0.230)	(0.232)
Constant	-0.711	-0.916	-0.099	-0.171
	(0.515)	(0.560)	(0.358)	(0.409)
Observations	3661	3661	3661	3661
R^2	0.143	0.143	0.125	0.126

Notes: Each column reports coefficients from OLS regressions estimated with school fixed effects and clustered standard errors (in parentheses) at the school level. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 2.5 reports the estimated effects of *test-screened tenured* teachers on reading and math student achievement gains, differentiating by the Ministerial Resolution Competition. All the model specifications show that *test-screened tenured* teachers who participated in competitions organized under the 2007 Regulation do not have a statistically significant effect on student learning gains in reading or math, when compared to other teachers.

Similar to the model estimations outlined before, the results in table 2.6 differentiate according to household poverty status. We do observe in Column 2 that *test-screened tenured* teachers of the 2007 Regulation have a positive effect on reading for students who come from poor households, at the 10 percent significance level, when additional teacher characteristics are taken into account as controls in the OLS estimations. For students living in poverty, having a *test-screened* teacher *tenured* under the 2007 Regulation is associated with a 0.195 standard deviation gain in reading achievement over three months of instruction. By contrast, none of the teacher characteristics have any significant effect on reading for students from non-poor households.

This finding, however, is only observed for reading, not for math. From table 2.6 we can see that there is no significant effect of having a *test-screened* teacher tenured under the 2007 Regulation. There is also some evidence of a positive significant effect of *test-screened* teachers tenured under the 2010 Regulation; however, given that the estimations are based on very small sample, results should thus be treated with considerable caution. In addition, the teacher characteristic that has a positive significant association with math achievement gains is whether the teacher has a university degree for students from poor households. Then again, no effect is found for students from non-poor households.

Table 2.6: Estimates of Test-screened Tenured Teacher Effects on Reading and Math Achievement Gains by Student Poverty Condition, Accounting for Ministerial Resolution of Competitions

	Reading				Math			
	Poor Household		Non-poor Household		Poor Household		Non-poor Household	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher:								
Test-screened tenured AM 438-07	0.189	0.195*	-0.074	-0.099	-0.113	-0.093	-0.032	-0.138
	(0.116)	(0.114)	(0.279)	(0.283)	(0.071)	(0.072)	(0.239)	(0.249)
Test-screened tenured AM 018-10	0.076	0.129	-0.197	-0.202	0.343**	0.439***	-0.357	-0.384
	(0.086)	(0.091)	(0.284)	(0.283)	(0.145)	(0.144)	(0.373)	(0.366)
Female		0.000		-0.072		-0.044		-0.185
		(0.100)		(0.173)		(0.091)		(0.216)
University degree		0.103		-0.062		0.159*		-0.163
		(0.086)		(0.148)		(0.081)		(0.144)
Years of experience		0.006		0.005		0.014		-0.001
		(0.020)		(0.027)		(0.016)		(0.037)
Years of experience squared		-0.000		-0.000		-0.000		-0.001
		(0.001)		(0.001)		(0.000)		(0.001)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	YES	YES	YES	YES	YES	YES	YES	YES
Parent controls	YES	YES	YES	YES	YES	YES	YES	YES
Classroom controls	YES	YES	YES	YES	YES	YES	YES	YES
Observations	2732	2732	929	929	2732	2732	929	929
R^2	0.164	0.165	0.329	0.330	0.144	0.146	0.306	0.313

Notes: Each column reports coefficients from OLS regressions estimated with school fixed effects and clustered standard errors (in parentheses) at the school level. Columns (1)-(8) control for the following student characteristics: age, gender, attendance to early education; parent characteristics: years of education; classroom characteristics: class size, classroom averages of student and parent characteristics. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

2.4.3 OLS Estimates of Test-Screened Teacher Effects

As previously discussed, teacher candidates in Ecuador are required to pass national entrance tests before they can participate in merit-based selection competitions for tenure at public schools. Teacher candidates who pass entrance examinations become eligible to compete for teacher vacancies. In our school sample, approximately 68 percent of teachers are *test-screened tenured* teachers who had won merit-based competitions. Previously, we compared this group with all of their peer teachers. However, in our comparison group there are contract teachers who had passed national entrance exams but had not yet won a competition for tenure (around 12 percent of the sample). These teachers are *test-screened* candidates in the process of applying to merit-

based competitions for vacancies at public schools. In this section I analyze whether *test-screened* teachers, regardless of their tenure status, produced significantly different results in reading or math compared to their peers who had not taken or passed the national entrance examinations.

The effect of *test-screened* teachers on learning outcomes is estimated using a specification analogous to equation (2.1):

$$Y_{isct} - Y_{isct-1} = \alpha_0 + test\_screened_{cst}\beta_1 + X_{icst}\beta_2 + \bar{X}_{icst}\beta_3 + C_{cst}\beta_4 + T_{cst}\beta_5 + u_{it} \quad (2.3)$$

Table 2.7 reports *test-screened* teacher effects on reading and math student achievement gains. The estimations do not identify any significant effect.

The effects of *test-screened* teachers on the reading and math achievement gains of students from poor and non-poor households are also estimated and reported in table 2.8. Here, we do not observe any effect from *test-screened* teachers on reading or math learning gains for students who come from poor and non-poor households. [30]

The differences between the effects of *test-screened* and *test-screened tenured* teachers on reading achievement of students from poor households have two potential explanations. First, differences in skill, subject knowledge and classroom practice scores among *test-screened* teachers might be relevant indicators of quality that helped to select and tenure the best candidates. Second, there could be a positive association between tenure status and performance at public schools in Ecuador.

[30] I run several specifications for this model in order to have a clearer understanding of the differences in the effects of *test-screened* and *test-screened tenured* teachers. First, I control for tenure status in the original model specification; however, results do not change. Given that the correlation between tenure status and test-screened tenured teacher is above 0.9, it is not ideal to introduce this variable in the model. In addition, I divide the variable of interest into *test-screened tenured* and *test-screened contract* teachers, and indeed find that while *test-screened tenured* teachers outperform the control group for students living in poverty, *test-screened contract* teachers do not.

Table 2.7: Estimates of Test-Screened Teacher Effects on Reading and Math Achievement Gains

	Reading		Math	
	(1)	(2)	(3)	(4)
Teacher:				
Test-Screened	0.125	0.124	-0.039	-0.044
	(0.081)	(0.083)	(0.058)	(0.062)
Female		0.001		-0.075
		(0.091)		(0.087)
University degree		0.073		0.076
		(0.071)		(0.069)
Years of experience		0.010		0.009
		(0.016)		(0.014)
Years of experience squared		-0.000		-0.000
		(0.000)		(0.000)
Student:				
Female	0.044	0.045	-0.019	-0.019
	(0.034)	(0.034)	(0.036)	(0.036)
Age (months)	0.034	0.034	-0.047**	-0.047**
	(0.022)	(0.022)	(0.019)	(0.019)
Attended early education	0.112**	0.112**	0.040	0.039
	(0.048)	(0.048)	(0.047)	(0.047)
Family:				
Parents' years of education	0.007	0.007	0.005	0.005
	(0.006)	(0.006)	(0.005)	(0.005)
Household is poor (BDH)	-0.013	-0.013	-0.012	-0.012
	(0.042)	(0.042)	(0.049)	(0.049)
School:				
Class size	0.003	0.002	-0.000	-0.001
	(0.007)	(0.007)	(0.005)	(0.005)
Average age	-0.000	0.003	0.063**	0.064**
	(0.038)	(0.039)	(0.031)	(0.031)
Proportion females	-0.229	-0.185	-0.123	-0.046
	(0.224)	(0.231)	(0.173)	(0.184)
Proportion attended early education	-0.105	-0.079	-0.151	-0.118
	(0.211)	(0.210)	(0.158)	(0.162)
Average parents' education	0.025	0.024	-0.004	-0.004
	(0.025)	(0.025)	(0.024)	(0.024)
Average poor household	-0.090	-0.091	0.108	0.100
	(0.271)	(0.273)	(0.229)	(0.231)
Constant	-0.716	-0.894*	-0.108	-0.182
	(0.493)	(0.532)	(0.352)	(0.398)
Observations	3661	3661	3661	3661
R^2	0.143	0.144	0.125	0.126

Notes: Each column reports coefficients from OLS regressions estimated with school fixed effects and clustered standard errors (in parentheses) at the school level. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 2.8: Estimates of Test-Screened Teacher Effects on Reading and Math Achievement Gains by Student Poverty Condition

	Reading				Math			
	Poor Household		Non-poor Household		Poor Household		Non-poor Household	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher:								
Test-screened	0.152	0.153	0.050	0.048	-0.035	-0.025	-0.142	-0.185
	(0.095)	(0.097)	(0.184)	(0.186)	(0.056)	(0.062)	(0.173)	(0.173)
Female		-0.003		-0.050		-0.046		-0.166
		(0.102)		(0.166)		(0.092)		(0.202)
University degree		0.093		-0.051		0.154*		-0.136
		(0.087)		(0.154)		(0.081)		(0.138)
Years of experience		0.005		0.009		0.014		0.005
		(0.020)		(0.027)		(0.016)		(0.036)
Years of experience squared		-0.000		-0.000		-0.000		-0.001
		(0.001)		(0.001)		(0.000)		(0.001)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	YES	YES	YES	YES	YES	YES	YES	YES
Parent controls	YES	YES	YES	YES	YES	YES	YES	YES
Classroom controls	YES	YES	YES	YES	YES	YES	YES	YES
Observations	2732	2732	929	929	2732	2732	929	929
R^2	0.164	0.165	0.329	0.329	0.144	0.146	0.306	0.312

Notes: Each column reports coefficients from OLS regressions estimated with school fixed effects and clustered standard errors (in parentheses) at the school level. Columns (1)-(8) control for the following student characteristics: age, gender, attendance to early education; parent characteristics: years of education; classroom characteristics: class size, classroom averages of student and parent characteristics. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

2.5 Propensity Score Matching Estimation Strategy and Results

2.5.1 PSM Estimates of Test-Screened Tenured Teacher Effects

As previously discussed, *test-screened tenured* teachers were not randomly allocated to students or schools. In order to fully address this potential source of bias in the OLS regressions and to estimate causal treatment effects, I apply a PSM technique following Rosenbaum and Rubin's (1983) approach. For the PSM estimation, I consider *test-screened tenured* teachers as the "treatment." The PSM technique balances the probability for a student to be assigned to a *test-screened tenured* teacher based on her observed pre-treatment characteristics. Moreover, it postulates that conditional on a set of relevant exogenous covariates, systematic differences in outcomes between treated and comparisons students with the same propensity score, or probability of being assigned to a *test-screened tenured* teacher in this case, are attributable to treatment

(conditional independence assumption, CIA). In addition, the PSM technique addresses the problematic over-representation of *test-screened tenured* teachers in the sample.

As mentioned, for the PSM model specification, assignment to a *test-screened tenured* teacher is the "treatment." In a first model specification, I further consider that treatment selection is on the student, family and classroom observable characteristics featured in the original value-added model, and additional school level observable characteristics (S_{st}).[31] that are absorbed by the school fixed effect component in the OLS estimation. Under this model specification, *test-screened tenured* teachers are the "treatment" regardless of their individual teacher characteristics such as gender, educational level and experience. The first model aims to reproduce a randomized experiment where *test-screened tenured* teachers are a pure treatment. This model will be henceforth referred to as the unconditional to teacher characteristics' model or unconditional model.

Assuming that the CIA holds and imposing an overlap between both treated and comparison students, the PSM estimators for the average treatment effect (ATE) and average treatment effect on the treated (ATT) of the unconditional model can be written as:

$$ATE = E_{P(X,\bar{X},C,S)}\{E[Y_{isct} - Y_{isct-1}(1)|test\_tenured = 1, P(X_{icst}, \bar{X}_{icst}, C_{cst}, S_{st})] - E[Y_{isct} - Y_{isct-1}(0)|test\_tenured = 0, P(X_{icst}, \bar{X}_{icst}, C_{cst}, S_{st})]\} \quad (2.4)$$

$$ATT = E_{P(X,\bar{X},C,S)|test\_tenured=1}\{E[Y_{isct} - Y_{isct-1}(1)|test\_tenured = 1, P(X_{icst}, \bar{X}_{icst}, C_{cst}, S_{st})] - E[Y_{isct} - Y_{isct-1}(0)|test\_tenured = 0, P(X_{icst}, \bar{X}_{icst}, C_{cst}, S_{st})]\} \quad (2.5)$$

[31] School area (urban or rural), type (complete or multi-grade) and repetition rate.

Applied to our case, the ATE is defined as the expected average causal achievement gain in reading or math of being assigned to a *test-screened tenured* teacher for a randomly chosen student from the population. Alternatively, the ATT measures the expected average causal achievement gain in reading or math for those students who have actually being assigned to a *test-screened tenured* teacher.

In a second model specification, I consider that treatment selection is not only on the student, family, classroom and school observable characteristics, but also on other teacher observable characteristics (T_{cst}). Under the second model specification, I intend to reproduce a randomized experiment where the treatment (*test-screened tenured* teacher) was blocked by teacher observed characteristics. This model will be henceforth referred to as the full model. The PSM estimators for the ATE and ATT of the full model can be written as:

$$ATE = E_{P(T,X,\bar{X},C,S)}\{E[Y_{isct} - Y_{isct-1}(1)|test\_tenured = 1, P(T_{cst}, X_{icst}, \bar{X}_{icst}, C_{cst}, S_{st})] - E[Y_{isct} - Y_{isct-1}(0)|test\_tenured = 0, P(T_{cst}, X_{icst}, \bar{X}_{icst}, C_{cst}, S_{st})]\} \quad (2.6)$$

$$ATT = E_{P(T,X,\bar{X},C,S)|Test\_tenured=1}\{E[Y_{isct} - Y_{isct-1}(1)|test\_tenured = 1, P(T_{cst}, X_{icst}, \bar{X}_{icst}, C_{cst}, S_{st})] - E[Y_{isct} - Y_{isct-1}(0)|test\_tenured = 0, P(T_{cst}, X_{icst}, \bar{X}_{icst}, C_{cst}, S_{st})]\} \quad (2.7)$$

In order to implement the PSM for both model specifications, first I estimate the probability of participation in the treatment with logit regressions, presented in table A2.1 of the Appendix. According to both propensity score estimations, students in classrooms with a higher proportion of poor households or attending rural schools were significantly more likely to be assigned to *test-screened tenured* teachers. Once again, the evidence supports the idea that more vulnerable students and schools are matched to *test-screened tenured* teachers.

In a second step of the PSM estimation for both model specifications, I apply three matching algorithms: i) nearest neighbor matching with replacement using one neighbor, ii) nearest neighbor matching with replacement using five neighbors and imposing a caliper, and iii) Gaussian Kernel matching. The algorithms group the most

comparable students into treatment and control groups in order to estimate the ATE and ATT of *test-screened tenured* teachers on student achievement gains in reading and math. The PSM algorithms also reduce imbalance in the pre-treatment covariates between the treated and control groups, thereby reducing the degree of model dependence and potential for bias. In order to assess whether the matching procedures have been successful in balancing the distribution of the observable relevant pre-treatment covariates in both the control and treatment groups, I estimate the following quality indicators: mean standardized bias[32], and pseudo-R^2 and joint significance of probably of participation before and after matching (Caliendo and Kopeinig, 2008).

In addition, for all PSM estimations, I impose a common support that drops treatment observations whose propensity score is higher than the maximum or less than the minimum propensity score of the controls.

Table A 2.2 of the Appendix reports the PSM quality indicators for reading and math achievement gains for both model specifications. Colum 1 reports the unmatched sample indicators. Subsequently, column 2 reports quality indicators for the one nearest neighbor matching, column 3 for the five nearest neighbors matching, and column 4 for the Gaussian Kernel matching. All matching algorithms reduce the mean standardized bias below 5 percent after matching, which is a sufficient balance measure, with the exception of the one nearest neighbor matching for the full model. The pseudo-R^2, which indicates how well the observable characteristics explain the participation probability in the logit regression, is smaller and close to zero after matching for all procedures. Finally, the likelihood ratio test on the joint significance of all regressors in the logit model is rejected in all matching approaches after matching for the unconditional model, with the exception of the one nearest neighbor matching. The ratio test on the joint significance is just rejected by the Gaussian Kernel matching for the full model. Overall, the Gaussian Kernel matching outperforms the other approaches for reading and for math in the unconditional and full models. This can also

[32] Mean bias corresponds to the mean of the standardized bias of observable pre-treatment characteristics in the PSM model. The standardized bias for each covariate is defined as the difference of sample means in the treated and matched control subsamples divided by the square root of the average of sample variances in both groups, then multiply by 100 (Rosenbaum and Rubin, 1985).

be observed in Figure A 2.1 of the Appendix, which shows the mean standardized bias of the PSM pre-treatment covariates before and after matching for the three algorithms and the two model specifications.

Table 2.9 reports ATT and ATE of *test-screened tenured* teachers on reading and math achievement gains estimated with the three matching algorithms, for both model specifications. I do not find evidence of a causal effect of *test-screened tenured* teachers on reading or math achievement gains from the PSM estimations, either for the unconditional to teacher characteristics model or for the full model.

Table 2.9: PSM Estimates of Test-screened Tenured Teacher Effects on Reading and Math Achievement Gains

	Reading			Math		
	PSM Nearest Neighbor 1 (1)	PSM Nearest Neighbor 5 (2)	PSM Kernel (3)	PSM Nearest Neighbor 1 (4)	PSM Nearest Neighbor 5 (5)	PSM Kernel (6)
Unconditional Model						
ATT	0.024	0.044	0.044	-0.053	-0.024	-0.033
	(0.059)	(0.050)	(0.041)	(0.062)	(0.054)	(0.047)
ATE	0.043	0.052	0.052	-0.018	0.002	-0.004
	(0.049)	(0.045)	(0.036)	(0.052)	(0.046)	(0.047)
Full Model						
ATT	0.018	0.055	0.057	-0.060	-0.057	-0.027
	(0.056)	(0.053)	(0.049)	(0.064)	(0.056)	(0.050)
ATE	0.027	0.049	0.047	-0.032	-0.032	-0.009
	(0.048)	(0.049)	(0.039)	(0.056)	(0.053)	(0.045)
N	3661	3661	3661	3661	3661	3661

Notes: Standard errors (in parentheses) are bootstrapped each with 100 repetition for ATT and ATE. Propensity score of unconditional model estimated with the following observable student characteristics: age, gender, attendance to early education; parent characteristics: years of education, poverty status (BDH); classroom characteristics: class size, classroom averages of student and parent characteristics; school characteristics: school area, type and repetition rate. Propensity score of full model estimated in addition with observable teacher characteristics: gender, university degree, years of experience, years or experience squared. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

In order to identify causal effects of *test-screened tenured* teachers on the reading and math achievement gains of students from poor and non-poor households, I next construct a stratified PSM estimation with the same three matching algorithms for both model specifications. PSM quality indicators are presented in table A 2.3 of the

Appendix. The Gaussian Kernel matching outperforms the other approaches in both samples for reading and math in the unconditional model, and is the only procedure that achieves good balance indicators for the full model. The one and five nearest neighbors matching do not meet the PSM quality indicators for balance in the full model. Therefore, their results should be considered with caution.

Table 2.10: PSM Estimates of Test-screened Tenured Teacher Effects on Reading and Math Achievement Gains, Stratified by Student Poverty Condition

	Poor Household			Non-Poor Household		
	PSM Nearest Neighbor 1	PSM Nearest Neighbor 5	PSM Kernel	PSM Nearest Neighbor 1	PSM Nearest Neighbor 5	PSM Kernel
	(1)	(2)	(3)	(4)	(5)	(6)
			Reading			
Unconditional Model						
ATT	0.193***	0.151***	0.105**	-0.077	-0.122	-0.131*
	(0.074)	(0.045)	(0.048)	(0.128)	(0.095)	(0.076)
ATE	0.149***	0.137***	0.103**	-0.039	-0.092	-0.080
	(0.057)	(0.045)	(0.047)	(0.090)	(0.089)	(0.073)
Full Model						
ATT	0.092	0.066	0.107*	-0.163	-0.126	-0.105
	(0.067)	(0.066)	(0.055)	(0.129)	(0.111)	(0.108)
ATE	0.106*	0.060	0.094*	-0.101	-0.072	-0.066
	(0.058)	(0.050)	(0.051)	(0.105)	(0.098)	(0.087)
			Math			
Unconditional Model						
ATT	0.005	0.021	-0.017	-0.138	-0.099	-0.099
	(0.067)	(0.056)	(0.053)	(0.128)	(0.112)	(0.102)
ATE	-0.000	0.023	0.001	-0.104	-0.073	-0.044
	(0.066)	(0.055)	(0.052)	(0.100)	(0.093)	(0.088)
Full Model						
ATT	0.018	0.001	0.003	0.015	-0.083	-0.066
	(0.078)	(0.072)	(0.051)	(0.158)	(0.113)	(0.107)
ATE	0.044	0.005	0.013	0.023	-0.012	-0.021
	(0.052)	(0.051)	(0.050)	(0.107)	(0.099)	(0.097)
N	2732	2732	2732	929	929	929

Notes: Standard errors (in parentheses) are bootstrapped each with 100 repetition for ATT and ATE. Propensity score of unconditional model estimated with the following observable student characteristics: age, gender, attendance to Early Education; parent characteristics: years of education; classroom characteristics: class size, classroom averages of student and parent characteristics; school characteristics: school area, type and repetition rate. Propensity score of full model estimated in addition with observable teacher characteristics: gender, university degree, years of experience, years or experience squared. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 2.10 presents the ATT and the ATE of *test-screened tenured* teachers on reading and math stratified by student poverty condition. Interestingly, according to the matching algorithms that balance the pre-treatment covariates, the ATT of *test-screened*

tenured teachers on reading achievement is positive and significant for disadvantaged students in both the unconditional and full models. This ATT ranges between a 0.105 and a 0.107 standard deviation gain, under the Gaussian Kernel matching. A significant ATE for reading is also consistently found in both models. Thus, the ATE of being assigned to a *test-screened tenured* teacher ranges between a 0.094 and a 0.103 standard deviation gain in reading achievement for a student living in poverty, under the Gaussian Kernel matching.

Besides this, no significant effect is found on students from non-poor households in reading, and no significant effect is found in math either for students from poor households or for students from non-poor households.

2.5.2 PSM Estimates of Test-Screened Tenured Teacher Effects, Accounting for Ministerial Resolution of Competitions

The causal effects of *test-screened tenured* teachers who participated in competitions ruled by the 2007 Regulation are also estimated for the unconditional and full models presented in the previous section. In this case, students of *test-screened* teachers *tenured* under the 2007 Regulation are the treatment group and the control group corresponds to students assigned to other teachers that had not been tenured under any of the new merit-based competitions.[33] Table A 2.4 of the Appendix reports the PSM quality indicators for reading and math achievement gains. Overall, the Gaussian Kernel matching outperforms the other approaches for the unconditional model and is the only one that achieves good balance indicators for the full model. This is also apparent from Figure A 2.2 of the Appendix, which shows the mean standardized bias of the PSM pre-treatment covariates before and after matching.

Table 2.11 presents the ATT and ATE of *test-screened* teachers *tenured* under the 2007 Regulation on reading and math achievement gains. A PSM estimation of the unconditional model show some evidence of a positive and significant ATT of these teachers on reading; however, this effect is not consistently found and does not hold in

[33] Test-screened teachers tenured under the Ministerial Regulation of January 2010 and their students are not taken into account. They constitute a different treatment.

the full model estimations. In contrast, we do not find any evidence of a causal effect of these *test-screened tenured* teachers on math achievement gains.

Table 2.11: PSM Estimates of Test-screened Tenured Teacher Effects (2007 Regulation) on Reading and Math Achievement Gains

	Reading			Math		
	PSM Nearest Neighbor 1 (1)	PSM Nearest Neighbor 5 (2)	PSM Kernel (3)	PSM Nearest Neighbor 1 (4)	PSM Nearest Neighbor 5 (5)	PSM Kernel (6)
Unconditional Model						
ATT	0.061	0.081*	0.056	-0.062	-0.053	-0.036
	(0.061)	(0.045)	(0.039)	(0.061)	(0.057)	(0.048)
ATE	0.026	0.068	0.061	-0.026	-0.019	-0.007
	(0.047)	(0.043)	(0.039)	(0.054)	(0.043)	(0.042)
Full Model						
ATT	0.025	0.091	0.080	0.068	0.031	-0.019
	(0.060)	(0.062)	(0.053)	(0.076)	(0.059)	(0.056)
ATE	-0.014	0.059	0.060	0.046	0.022	-0.004
	(0.051)	(0.047)	(0.041)	(0.055)	(0.054)	(0.044)
N	3581	3581	3581	3581	3581	3581

Notes: Standard errors (in parentheses) are bootstrapped each with 100 repetition for ATT and ATE. Propensity score of unconditional model estimated with the following observable student characteristics: age, gender, attendance to early education; parent characteristics: years of education, poverty status (BDH); classroom characteristics: class size, classroom averages of student and parent characteristics; school characteristics: school area, type and repetition rate. Propensity score of full model estimated in addition with observable teacher characteristics: gender, university degree, years of experience, years or experience squared. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

A stratified PSM estimation by poverty condition is also implemented to identify the causal effects of *test-screened* teachers *tenured* by the 2007 Regulation. Table A 2.5 of the Appendix reports the PSM quality indicators for reading and math achievement gains. Overall, the Gaussian Kernel matching outperforms the other approaches for reading and for math for both samples of students from poor and non-poor households, and both model specifications. Estimations of the one nearest neighbor matching for the full model should be viewed with caution, since it does not achieve quality balance standards.

The ATT and ATE effects on reading and math achievement gains stratified by poverty condition are presented in table 2.12. It is noteworthy that all algorithms find

positive and significant ATT and ATE effects on reading among students living in poor households in both model specifications. Under the Gaussian Kernel matching, the ATT of having a *test-screened* teacher *tenured* under the 2007 Regulation ranges between a 0.109 and a 0.117 standard deviation gain in reading achievement among poor students. The ATE ranges between a 0.104 and a 0.109 standard deviation gain. In addition, no significant effect is found on students from non-poor households in reading achievement and no significant effect is found in math.

Table 2.12: PSM Estimates of Test-screened Tenured Teacher Effects (2007 Regulation) on Reading and Math Achievement Gains, Stratified by Student Poverty Condition

	Poor Household			Non-Poor Household		
	PSM Nearest Neighbor 1	PSM Nearest Neighbor 5	PSM Kernel	PSM Nearest Neighbor 1	PSM Nearest Neighbor 5	PSM Kernel
	(1)	(2)	(3)	(4)	(5)	(6)
			Reading			
Unconditional Model						
ATT	0.125*	0.108*	0.109**	-0.099	-0.092	-0.129
	(0.066)	(0.057)	(0.048)	(0.115)	(0.094)	(0.089)
ATE	0.110*	0.112**	0.109**	-0.077	-0.062	-0.068
	(0.063)	(0.052)	(0.046)	(0.105)	(0.091)	(0.086)
Full Model						
ATT	0.125*	0.136*	0.117**	0.180	-0.003	-0.033
	(0.065)	(0.071)	(0.058)	(0.152)	(0.117)	(0.086)
ATE	0.114*	0.125**	0.104**	0.073	0.005	-0.031
	(0.065)	(0.057)	(0.051)	(0.107)	(0.091)	(0.095)
			Math			
Unconditional Model						
ATT	0.039	-0.017	-0.019	-0.137	-0.064	-0.106
	(0.067)	(0.061)	(0.052)	(0.121)	(0.113)	(0.104)
ATE	0.020	0.007	0.001	-0.041	-0.008	-0.030
	(0.057)	(0.058)	(0.045)	(0.116)	(0.089)	(0.088)
Full Model						
ATT	0.014	0.017	0.000	-0.235	-0.091	-0.091
	(0.082)	(0.064)	(0.053)	(0.154)	(0.140)	(0.134)
ATE	0.017	0.005	0.007	-0.045	-0.029	-0.044
	(0.061)	(0.053)	(0.050)	(0.132)	(0.104)	(0.092)
N	2698	2698	2698	883	883	883

Notes: Standard errors (in parentheses) are bootstrapped each with 100 repetition for ATT and ATE. Propensity score of unconditional model estimated with the following observable student characteristics: age, gender, attendance to early education; parent characteristics: years of education; classroom characteristics: class size, classroom averages of student and parent characteristics; school characteristics: school area, type and repetition rate. Propensity score of full model estimated in addition with observable teacher characteristics: gender, university degree, years of experience, years or experience squared. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

2.5.3 PSM Estimates of Test-Screened Teacher Effect

Finally, I use the PSM approach to estimate the causal effect of *test-screened* teachers, regardless of their tenure status, on reading and math achievement gains for the unconditional and full models. Under this scenario, *test-screened* teachers are the "treatment." Table A 2.6 of the Appendix reports the PSM quality indicator for reading and math. All matching algorithms achieve sufficient balance measures, with the exception of the one nearest neighbor matching. Figure A 2.3 of the Appendix details the mean standardized bias of the PSM pre-treatment covariates before and after matching.

Table 2.13 presents the ATT and ATE of *test-screened* teachers for reading and math achievement gains, estimated with the three matching algorithms. We do not find evidence of a significantly positive causal effect on reading or math achievement gains.

Table 2.13: PSM Estimates of Test-Screened Teacher Effects on Reading and Math Achievement Gains

	Reading			Math		
	PSM Nearest Neighbor 1	PSM Nearest Neighbor 5	PSM Kernel	PSM Nearest Neighbor 1	PSM Nearest Neighbor 5	PSM Kernel
	(1)	(2)	(3)	(4)	(5)	(6)
Unconditional Model						
ATT	0.020	0.076	0.056	0.027	-0.081	-0.018
	(0.055)	(0.052)	(0.044)	(0.071)	(0.067)	(0.051)
ATE	0.040	0.074	0.057	0.014	-0.070	-0.017
	(0.055)	(0.049)	(0.037)	(0.056)	(0.053)	(0.043)
Full Model						
ATT	0.035	0.051	0.057	0.092	0.005	0.005
	(0.063)	(0.060)	(0.049)	(0.070)	(0.075)	(0.048)
ATE	0.029	0.046	0.056	0.067	0.004	0.003
	(0.056)	(0.050)	(0.046)	(0.062)	(0.050)	(0.049)
N	3661	3661	3661	3661	3661	3661

Notes: Standard errors (in parentheses) are bootstrapped each with 100 repetition for ATT and ATE. Propensity score of unconditional model estimated with the following observable student characteristics: age, gender, attendance to early education; parent characteristics: years of education, poverty status (BDH); classroom characteristics: class size, classroom averages of student and parent characteristics; school characteristics: school area, type and repetition rate. Propensity score of full model estimated in addition with observable teacher characteristics: gender, university degree, years of experience, years or experience squared. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

I also implement a stratified PSM estimation by poverty condition to identify the causal effects of *test-screened* teachers for both the unconditional and full model. Table A2.7 of the Appendix reports the PSM quality indicators for reading and math achievement gains. The Gaussian Kernel matching outperforms the other approaches for the unconditional model, for the sample of students from poor households. The five nearest neighbors matching outperforms the other approaches in the full model for samples of students from poor households; however, it does not achieve sufficient balance indicators for the unconditional model. In addition, the one nearest neighbor matching is not able to balance the sample of students from poor households in both models. Only the Gaussian Kernel matching algorithm is close to balance the non-poor household sample.

Table 2.14 reports the ATT and ATE effects of *test-screened* teachers on reading and math achievement gains stratified by poverty condition. Under the matching procedures that show good balance indicators in the unconditional and full models, some evidence of significantly positive ATT and ATE is found for reading among poor households. For a student living in poverty, the ATT of being assigned to a *test-screened* teacher ranges between a 0.108 (Gaussian Kernel matching) and a 0.118 (five nearest neighbors matching) standard deviation in reading achievement gains. The ATE ranges between a 0.103 (Gaussian Kernel matching) and a 0.111 (five nearest neighbors matching) standard deviation in reading achievement gains. No significant effect is found on students from non-poor households. Likewise, no significant effect is found on math achievement, either for students of poor households or for students of non-poor households.

Table 2.14: PSM Estimates of Test-Screened Teacher Effects on Reading and Math Achievement Gains, Stratified by Student Poverty Condition

	Poor Household			Non-Poor Household		
	PSM Nearest Neighbor 1 (1)	PSM Nearest Neighbor 5 (2)	PSM Kernel (3)	PSM Nearest Neighbor 1 (4)	PSM Nearest Neighbor 5 (5)	PSM Kernel (6)
			Reading			
Unconditional Model						
ATT	0.018	0.082	0.108**	0.054	-0.093	-0.113
	(0.073)	(0.057)	(0.055)	(0.120)	(0.116)	(0.090)
ATE	0.051	0.091*	0.103**	0.036	-0.070	-0.074
	(0.059)	(0.053)	(0.048)	(0.097)	(0.084)	(0.082)
Full Model						
ATT	0.094	0.118*	0.086	-0.157	-0.060	-0.069
	(0.085)	(0.071)	(0.057)	(0.134)	(0.108)	(0.099)
ATE	0.089	0.111*	0.083	-0.121	-0.035	-0.045
	(0.074)	(0.061)	(0.056)	(0.106)	(0.084)	(0.093)
			Math			
Unconditional Model						
ATT	-0.035	-0.040	-0.003	-0.106	-0.133	-0.116
	(0.077)	(0.066)	(0.061)	(0.143)	(0.144)	(0.099)
ATE	-0.043	-0.050	-0.004	-0.091	-0.099	-0.092
	(0.075)	(0.063)	(0.054)	(0.136)	(0.103)	(0.112)
Full Model						
ATT	0.043	0.058	0.029	-0.155	-0.099	-0.096
	(0.091)	(0.079)	(0.061)	(0.137)	(0.121)	(0.099)
ATE	0.044	0.051	0.026	-0.133	-0.081	-0.078
	(0.075)	(0.067)	(0.056)	(0.145)	(0.124)	(0.089)
N	2732	2732	2732	929	929	929

Notes: Standard errors (in parentheses) are bootstrapped each with 100 repetition for ATT and ATE. Propensity score of unconditional model estimated with the following observable student characteristics: age, gender, attendance to early education; parent characteristics: years of education; classroom characteristics: class size, classroom averages of student and parent characteristics; school characteristics: school area, type and repetition rate. Propensity score of full model estimated in addition with observable teacher characteristics: gender, university degree, years of experience, years or experience squared. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

2.6 Conclusions

In order to improve teacher quality, the Ecuadorian government has required teacher candidates to pass national standardized tests since 2007. Passing these exams is necessary to participate in Ecuador's merit-based selection competitions for tenure at public schools. The question tackled in this article is whether teachers who passed national skill and knowledge tests and were tenured by Ecuador's new competitive recruitment have positive effects on student learning outcomes in the first grades of primary school.

Different estimations of a value-added to student achievement model show that, overall, *test-screened tenured* teachers were *not* more effective in raising student achievement in reading or math in the 2011-2012 academic year. However, merit-based and competitive tenure substantially improved reading skills for students living in poor households (BDH receivers). This finding is consistent in OLS and PSM estimations. The preferred PSM estimations show that students from poor families experienced at least a 0.094 standard deviation higher achievement gain in reading when taught by a *test-screened tenured teacher*, and at least a 0.109 standard deviation higher achievement gain when taught by a teacher tenured under the 2007 Regulation. These effects are not explained by credentials such as academic degree and experience, which suggest that the competitions' national standardized exams and the demonstration class assessments helped to differentiate the candidates.

The effects on reading and math of *test-screened* teachers, regardless of their tenure status, are also explored. The preferred PSM algorithm implemented suggests an average test-screened teacher effect of at least a 0.103 standard deviation achievement gain in reading for students living in poor households. This finding is not as robust as the one found for *test-screened tenured* teachers; nonetheless, it suggests once again that differences among teacher candidates' test scores are relevant.

In contrast to earlier findings (Cruz-Aguayo, Ibarrarán and Schady, 2017), this study offers crucial evidence of positive and significant effects of teachers screened and tenured through Ecuador's new competitive recruitment policy on the outcomes of students living in poverty. The sizes of these effects are substantial when compared to other studies conducted in the US and Ecuador. Estimations of the effects of certified teachers in the US have typically ranged between 0.01 and 0.07 standard deviations for reading and math test scores (Clotfelter, Ladd and Vigdor, 2007b; Goldhaber, 2007; Goldhaber and Anthony, 2007; Harris and Sass, 2009; Goldhaber, Gratz and Theobald, 2017). In their experimental study on teacher quality in Kindergarten in Ecuador, Araujo et al. (2016) found that a classroom practice score[34] was the only teacher

[34] The Classroom Assessment Scoring System (CLASS) was applied.

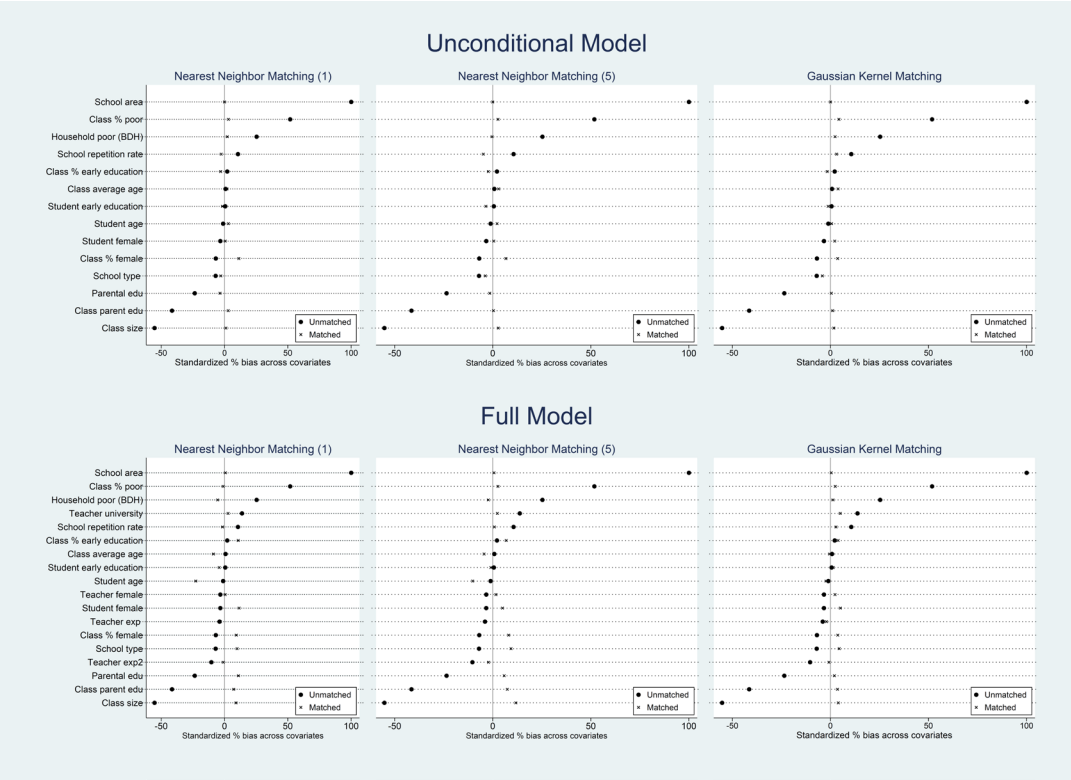
characteristic positively and significantly associated with learning outcomes[35], with effect sizes of a 0.06 and 0.08 standard deviation higher end of year tests scores for reading and math respectively.

Some policy implications can be drawn from the research results. On the one hand, the results show that the Ecuadorian reform partially succeeded in raising the quality and equity of the educational system between 2007 and 2011. The policy had a positive significant and substantial effect on reading achievement gains for students living in poverty. This particular population should not be ignored because it is concentrated among public schools. On the other hand, the results suggest that the policy was unable to recruit candidates that outperformed their peers in raising students' math achievement or in producing better results for non-poor students in primary schools. Thus, the quality of the tests and instruments used in the merit-based selection process should be carefully evaluated in order to improve the process and its effectiveness. Further research is needed to evaluate the policy effects beyond its first four years of implementation.

[35] In addition, children with inexperienced (less than 3 years of experience) teachers had test scores that were 0.17 standard deviation lower. None of the other teacher characteristics analyzed (tenure status, IQ, the Big Five dimensions of personality, inhibitory control and attention, and early circumstances) were associated with student learning.

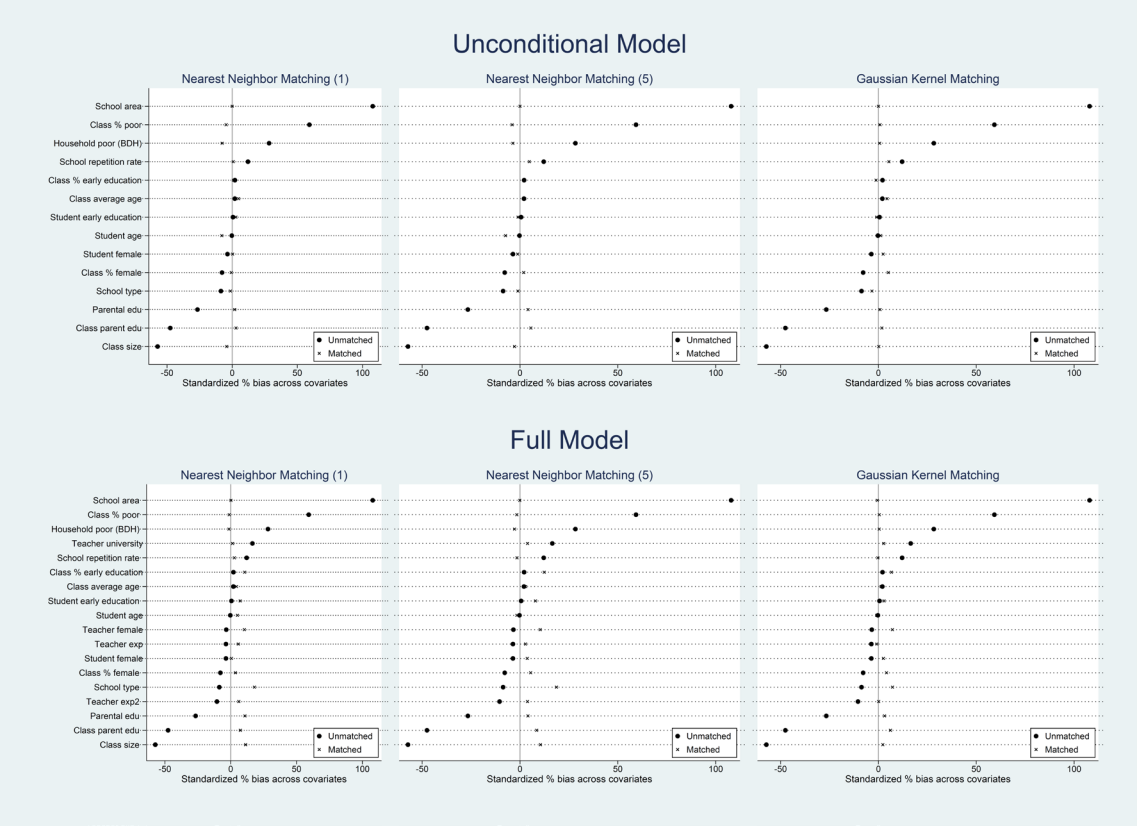
2.7 Appendix

Figure A 2.1: Mean Standardized Bias of Pre-treatment Covariates Before and After Matching for Reading and Math Achievement Gains, when Treatment is Test-Screened Tenured Teacher



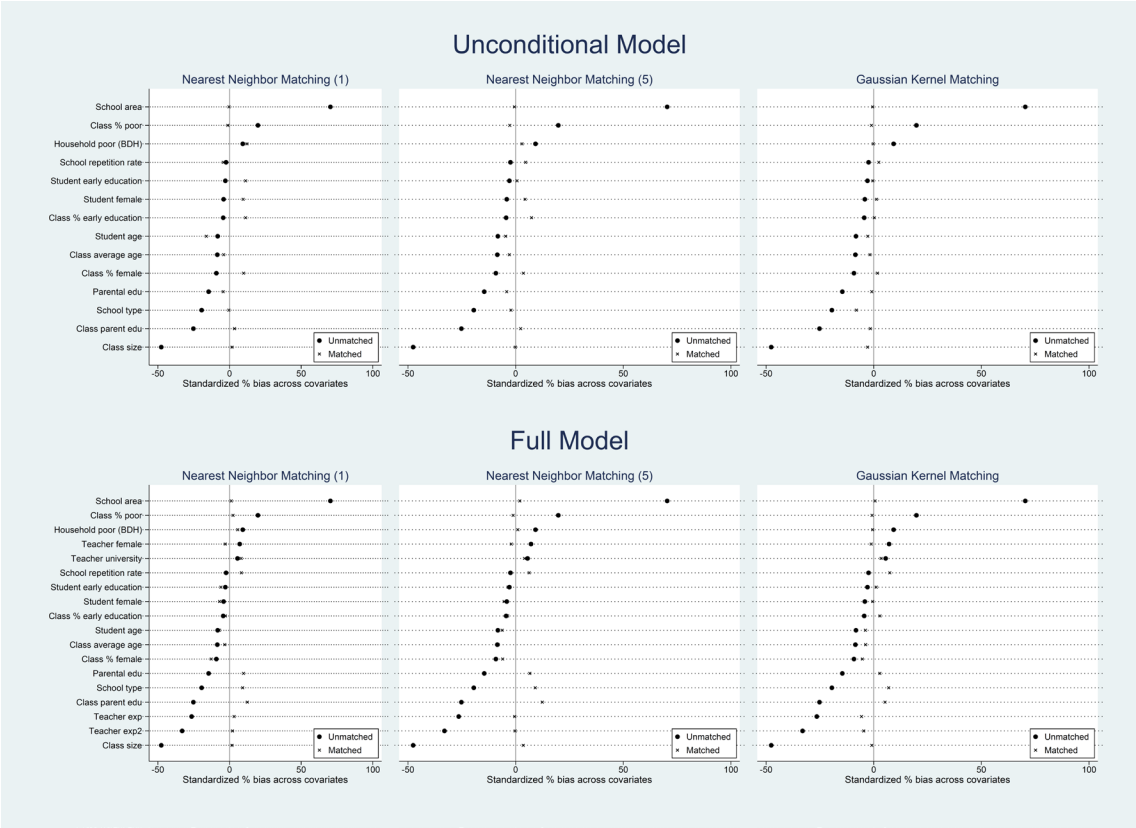
Notes: This table reports mean standardized bias of pre-treatment covariates before and after matching in the PSM model, when treatment is *test-screened tenured* teacher. The standardized bias for each covariate is defined as the difference of sample means in the treated and matched control subsamples divided by the square root of the average of sample variances in both groups, then multiply by 100 (Rosenbaum and Rubin, 1985).

Figure A 2.2: Mean Standardized Bias of Pre-treatment Covariates Before and After Matching for Reading and Math Achievement Gains, when Treatment is Test-Screened Tenured Teacher (2007 Regulation)



Notes: This table reports mean standardized bias of pre-treatment covariates before and after matching in the PSM model, when treatment is *test-Screened tenured* teacher (2007 Regulation). The standardized bias for each covariate is defined as the difference of sample means in the treated and matched control subsamples divided by the square root of the average of sample variances in both groups, then multiply by 100 (Rosenbaum and Rubin, 1985).

Figure A 2.3: Mean Standardized Bias of Pre-treatment Covariates Before and After Matching for Reading and Math Achievement Gains, when Treatment is Test-Screened Teacher



Notes: This table reports mean standardized bias of pre-treatment covariates before and after matching in the PSM model, when treatment is test-screened teacher. The standardized bias for each covariate is defined as the difference of sample means in the treated and matched control subsamples divided by the square root of the average of sample variances in both groups, then multiply by 100 (Rosenbaum and Rubin, 1985).

Table A 2.1: Propensity Score Estimation of Assignment to Test-Screened Tenured Teacher, Unconditional and Full Model

	Test-screened Tenured Teacher	
	Unconditional Model	Full Model
	(1)	(2)
Teacher:		
Female		0.286**
		(0.128)
University degree		0.419***
		(0.088)
Years of experience		0.152***
		(0.022)
Years of experience equared		-0.006***
		(0.001)
Student:		
Female	-0.019	-0.027
	(0.083)	(0.084)
Age (months)	-0.026	-0.034
	(0.049)	(0.049)
Attended early education	-0.030	-0.030
	(0.111)	(0.112)
Family:		
Parents' years of education	-0.002	-0.002
	(0.013)	(0.014)
Household is poor (BDH)	-0.011	-0.007
	(0.103)	(0.104)
Classroom:		
Class size	-0.010*	-0.015***
	(0.005)	(0.005)
Average age	0.053	0.096
	(0.065)	(0.066)
Proportion females	-0.264	-0.371
	(0.276)	(0.281)
Proportion attended early education	0.275	0.387*
	(0.227)	(0.231)
Average parents' education	0.018	0.040
	(0.028)	(0.028)
Proportion poor household	0.738***	0.770***
	(0.242)	(0.248)
School:		
Complete	-0.011	0.073
	(0.109)	(0.112)
Rural	1.730***	1.843***
	(0.098)	(0.101)
Repetition rate	0.010	0.018
	(0.014)	(0.015)
Constant	-0.872	-2.691***
	(0.561)	(0.618)
N	3661	3661
chi2	697.538***	801.847***

Notes: Propensity Score estimated with logit. Standard errors in parentheses. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table A 2.2: PSM Quality Indicators Before and After Matching for Reading and Math Achievement Gains, when Treatment is Test-screened Tenured Teacher

Sample	Unmatched (1)	PSM Nearest Neighbor 1 Matched (2)	PSM Nearest Neighbor 5 Matched (3)	PSM Kernel Matched (4)
Unconditional Model				
Mean Bias (%)	23.6	2.9	2.5	2.2
Pseudo-R^2	0.152	0.005	0.003	0.002
LR chi^2	696.59	35.82	18.99	14.68
p >chi^2	0.000	0.001	0.165	0.400
Full Model				
Mean Bias (%)	20.1	6.7	4.8	2.7
Pseudo-R^2	0.174	0.021	0.012	0.004
LR chi^2	801.39	145.14	81.24	24.83
p >chi^2	0.000	0.000	0.000	0.130

Notes: This table reports PSM quality indicators before and after matching. Mean bias corresponds to the mean of the standardized bias of observable pre-treatment characteristics in the PSM model. The standardized bias for each covariate is defined as the difference of sample means in the treated and matched control subsamples divided by the square root of the average of sample variances in both groups, then multiply by 100 (Rosenbaum and Rubin, 1985). The Pseudo-R^2 indicates how well the observable pre-treatment characteristics explain participation probability in the propensity score estimation (logit). LR chi^2 corresponds to the likelihood ratio test on the joint significance of all observable pre-treatment characteristics in the propensity score estimation (logit), and p>chi^2 corresponds to its p-value. Column (1) reports the unmatched sample indicators. Column (2) reports quality indicators of one nearest neighbor matching, column (3) of five nearest neighbors matching, and column (4) of Gaussian Kernel matching.

Table A 2.3: PSM Quality Indicators Before and after Matching for Reading and Math Achievement Gains by Student Poverty Condition, when Treatment is Test-screened Tenured Teacher

	Poor Household				Non-Poor Household			
Sample	Unmatched (1)	PSM Nearest Neighbor 1 Matched (2)	PSM Nearest Neighbor 5 Matched (3)	PSM Kernel Matched (4)	Unmatched (5)	PSM Nearest Neighbor 1 Matched (6)	PSM Nearest Neighbor 5 Matched (7)	PSM Kernel Matched (8)
Unconditional Model								
Mean Bias (%)	21.4	2.9	1.5	2.1	23.2	5.7	5.1	4.3
Pseudo-R^2	0.142	0.005	0.002	0.002	0.154	0.021	0.01	0.009
LR chi^2	466.99	24.98	8.97	10.48	194.56	31.49	14.46	12.74
p >chi^2	0.000	0.023	0.775	0.654	0.000	0.003	0.342	0.468
Full Model								
Mean Bias (%)	19.2	5.9	5.4	2.5	22.0	6.9	6.3	5.8
Pseudo-R^2	0.173	0.014	0.011	0.004	0.177	0.023	0.02	0.018
LR chi^2	568.41	74.61	58.77	22.37	223.58	34.18	29.3	27.39
p >chi^2	0.000	0.000	0.000	0.171	0.000	0.008	0.032	0.053

Notes: This table reports PSM quality indicators before and after matching. Mean bias corresponds to the mean of the standardized bias of observable pre-treatment characteristics in the PSM model. The standardized bias for each covariate is defined as the difference of sample means in the treated and matched control subsamples divided by the square root of the average of sample variances in both groups, then multiply by 100 (Rosenbaum and Rubin, 1985).The Pseudo-R^2 indicates how well the observable pre-treatment characteristics explain participation probability in the propensity score estimation (logit). LR chi^2 corresponds to the likelihood ratio test on the joint significance of all observable pre-treatment characteristics in the propensity score estimation (logit), and p>chi^2 corresponds to its p-value. Columns (1) and (5) report unmatched sample indicators. Column (2) and (6) report quality indicators of one nearest neighbor matching, columns (3) and (7) of five nearest neighbors matching, and columns (4) and (8) of Gaussian Kernel matching.

Table A 2.4: PSM Quality Indicators Before and After Matching for Reading and Math Achievement Gains, when Treatment is Test-screened Tenured Teacher (2007 Regulation)

Sample	Unmatched (1)	PSM Nearest Neighbor 1 Matched (2)	PSM Nearest Neighbor 5 Matched (3)	PSM Kernel Matched (4)
Unconditional Model				
Mean Bias (%)	26.0	2.9	3.0	2.0
Pseudo-R^2	0.177	0.007	0.006	0.002
LR chi^2	803.13	47.77	39.01	14.30
p >chi^2	0.000	0.000	0.000	0.427
Full Model				
Mean Bias (%)	22.1	6.0	5.7	2.8
Pseudo-R^2	0.207	0.015	0.014	0.004
LR chi^2	939.60	96.82	92.41	29.40
p >chi^2	0.000	0.000	0.000	0.044

Notes: This table reports PSM quality indicators before and after matching. Mean bias corresponds to the mean of the standardized bias of observable pre-treatment characteristics in the PSM model. The standardized bias for each covariate is defined as the difference of sample means in the treated and matched control subsamples divided by the square root of the average of sample variances in both groups, then multiply by 100 (Rosenbaum and Rubin, 1985). The Pseudo-R^2 indicates how well the observable pre-treatment characteristics explain participation probability in the propensity score estimation (logit). LR chi^2 corresponds to the likelihood ratio test on the joint significance of all observable pre-treatment characteristics in the propensity score estimation (logit), and p>chi^2 corresponds to its p-value. Column (1) reports the unmatched sample indicators. Column (2) reports quality indicators of one nearest neighbor matching, column (3) of five nearest neighbors matching, and column (4) of Gaussian Kernel matching.

Table A 2.5: PSM Quality Indicators Before and After Matching for Reading and Math by Student Poverty Condition, when Treatment is Test-screened Tenured Teacher (2007 Regulation)

	Poor Household				Poor Household			
Sample	Unmatched (1)	PSM Nearest Neighbor 1 Matched (2)	PSM Nearest Neighbor 5 Matched (3)	PSM Kernel Matched (4)	Unmatched (5)	PSM Nearest Neighbor 1 Matched (6)	PSM Nearest Neighbor 5 Matched (7)	PSM Kernel Matched (8)
Unconditional Model								
Mean Bias (%)	22.5	3.3	2.1	2.0	28.2	3.1	3.4	4.6
Pseudo-R^2	0.155	0.004	0.002	0.002	0.203	0.007	0.004	0.007
LR chi^2	506.77	19.88	12.79	11.53	245.86	9.10	5.08	9.22
p >chi^2	0.000	0.098	0.464	0.567	0.000	0.765	0.974	0.756
Full Model								
Mean Bias (%)	20.0	3.3	4.1	3.2	25.9	6.1	4.9	4.8
Pseudo-R^2	0.19	0.007	0.007	0.006	0.239	0.021	0.014	0.014
LR chi^2	619.22	36.45	37.27	29.48	289.04	27.72	18.98	18.73
p >chi^2	0.000	0.004	0.003	0.030	0.000	0.048	0.330	0.344

Notes: This table reports PSM quality indicators before and after matching. Mean bias corresponds to the mean of the standardized bias of observable pre-treatment characteristics in the PSM model. The standardized bias for each covariate is defined as the difference of sample means in the treated and matched control subsamples divided by the square root of the average of sample variances in both groups, then multiply by 100 (Rosenbaum and Rubin, 1985).The Pseudo-R^2 indicates how well the observable pre-treatment characteristics explain participation probability in the propensity score estimation (logit). LR chi^2 corresponds to the likelihood ratio test on the joint significance of all observable pre-treatment characteristics in the propensity score estimation (logit), and p>chi^2 corresponds to its p-value. Columns (1) and (5) report unmatched sample indicators. Column (2) and (6) report quality indicators of one nearest neighbor matching, columns (3) and (7) of five nearest neighbors matching, and columns (4) and (8) of Gaussian Kernel matching.

Table A 2.6: PSM Quality Indicators Before and After Matching for Reading and Math, when Treatment is Test-Screened Teacher

Sample	Unmatched (1)	PSM Nearest Neighbor 1 Matched (2)	PSM Nearest Neighbor 5 Matched (3)	PSM Kernel Matched (4)
Unconditional Model				
Mean Bias (%)	17.6	6.5	3.1	1.8
Pseudo-R^2	0.085	0.018	0.004	0.002
LR chi^2	318.44	140.84	33.70	13.38
p >chi^2	0.000	0.000	0.002	0.497
Full Model				
Mean Bias (%)	17.7	5.9	4.6	3.2
Pseudo-R^2	0.126	0.015	0.009	0.005
LR chi^2	470.12	114.16	72.13	40.91
p >chi^2	0.000	0.000	0.000	0.002

Notes: This table reports PSM quality indicators before and after matching. Mean bias corresponds to the mean of the standardized bias of observable pre-treatment characteristics in the PSM model The standardized bias for each covariate is defined as the difference of sample means in the treated and matched control subsamples divided by the square root of the average of sample variances in both groups, then multiply by 100 (Rosenbaum and Rubin, 1985). The Pseudo-R^2 indicates how well the observable pre-treatment characteristics explain participation probability in the propensity score estimation (logit). LR chi^2 corresponds to the likelihood ratio test on the joint significance of all observable pre-treatment characteristics in the propensity score estimation (logit), and p>chi^2 corresponds to its p-value. Column (1) reports the unmatched sample indicators. Column (2) reports quality indicators of one nearest neighbor matching, column (3) of five nearest neighbors matching, and column (4) of Gaussian Kernel matching.

Table A 2.7: PSM Quality Indicators Before and After Matching for Reading and Math by Student Poverty Condition, when Treatment is Test-Screened Teachers

	Poor Household				Non-Poor Household			
Sample	Unmatched (1)	PSM Nearest Neighbor 1 Matched (2)	PSM Nearest Neighbor 5 Matched (3)	PSM Kernel Matched (4)	Unmatched (5)	PSM Nearest Neighbor 1 Matched (6)	PSM Nearest Neighbor 5 Matched (7)	PSM Kernel Matched (8)
Unconditional Model								
Mean Bias	17.9	10.0	7.5	1.9	19.2	12.6	7.5	6.3
Pseudo R2	0.088	0.026	0.017	0.002	0.099	0.039	0.015	0.010
LR chi2	239.97	153.45	102.36	10.58	99.45	75.18	29.29	19.33
p >chi2	0.000	0.000	0.000	0.646	0.000	0.000	0.006	0.113
Full Model								
Mean Bias	18.8	4.0	3.1	3.8	16.7	4.7	5.6	5.8
Pseudo R2	0.141	0.010	0.008	0.009	0.118	0.020	0.013	0.012
LR chi2	383.45	57.46	48.23	52.23	119.01	38.07	25.09	22.31
p >chi2	0.000	0.000	0.000	0.000	0.000	0.002	0.093	0.173

Notes: This table reports PSM quality indicators before and after matching. Mean bias corresponds to the mean of the standardized bias of observable pre-treatment characteristics in the PSM model. The standardized bias for each covariate is defined as the difference of sample means in the treated and matched control subsamples divided by the square root of the average of sample variances in both groups, then multiply by 100 (Rosenbaum and Rubin, 1985). The Pseudo-R^2 indicates how well the observable pre-treatment characteristics explain participation probability in the propensity score estimation (logit). LR chi^2 corresponds to the likelihood ratio test on the joint significance of all observable pre-treatment characteristics in the propensity score estimation (logit), and p>chi^2 corresponds to its p-value. Columns (1) and (5) report unmatched sample indicators. Column (2) and (6) report quality indicators of one nearest neighbor matching, columns (3) and (7) of five nearest neighbors matching, and columns (4) and (8) of Gaussian Kernel matching.

Chapter 3 *Does Test-Based Teacher Recruitment Work in the Developing World? Experimental Evidence from Ecuador*

María Daniela Araujo P, Guido Heineck and Yyannú Cruz-Aguayo

3.1 Introduction

Identifying high-quality teachers who substantially contribute to student learning has been one of the main challenges faced by policy-makers and researchers in education in recent decades. Much current research has shown that teachers are a key factor in student learning, although what makes a good teacher remains a puzzle. Value-added to student achievement models derived from the education production function literature have not only found substantial teacher effects or individual teacher contribution to student achievement, but also substantial variation in this contribution (Hanushek and Rivkin, 2006, 2010, 2012; Chetty, Friedman and Rockoff, 2014b; Jackson, Rockoff and Staiger, 2014; Koedel, Mihaly and Rockoff, 2015). Interestingly, it has also been observed that easily quantifiable teacher characteristics such as academic degree, experience beyond the first years, training or test scores explain little of the individual teacher contribution to learning.

Teacher cognitive skill[36] and content knowledge of the subject taught measured by certification tests are among the widely-used observable characteristics as a signal of teacher quality for recruitment purposes in high-income economies. In the US, written tests have been used to certify teachers since the early-20th century, as a means to ensure high and uniform academic standards for teacher pre-service programs and

[36] For example, approximated by teachers' verbal and mathematical skills.

safeguard the public from faulty teacher candidates (D'Agostino and Powers, 2009). Nonetheless, the evidence regarding the effectiveness of teacher test scores as predictors of future quality remains mixed. On the one hand, early education production function studies typically found positive effects of teacher skill and certification test scores on student achievement (Wayne and Youngs, 2003). On the other hand, more recent studies using large longitudinal data and applying parametric and non-parametric value-added to student achievement models have not been able to consistently find positive or significant effects (Clotfelter, Ladd and Vigdor, 2007b; Goldhaber, 2007; Goldhaber and Anthony, 2007; Angrist and Guryan, 2008; Boyd *et al.*, 2008; Kane, Rockoff and Staiger, 2008; Harris and Sass, 2009, 2011; Rockoff *et al.*, 2011; Goldhaber, Gratz and Theobald, 2017). On top of the mixed evidence, all such studies suffer from potential estimation bias caused by the non-random matching of teachers to students (Clotfelter, Ladd and Vigdor, 2006; Rothstein, 2009, 2010; Koedel and Betts, 2011).

In contrast to the evidence on specific teacher test scores, new research on personnel policies from the US has consistently shown that combining teacher background characteristics and screening measures in teacher recruitment processes, such as content knowledge tests and interviews, can strongly predict future teacher effectiveness (Goldhaber, Grout and Huntington-Klein, 2017; Jacob *et al.*, 2018; Bruno and Strunk, 2019). These studies, nonetheless, also rely on non-experimental data where matching of teachers to students might be a source of concern.

Latin American countries have not traditionally had certification processes for the teaching profession to ensure that quality standards are met. However, in the last fifteen years several countries have implemented teacher recruitment policies based on skill and subject knowledge tests as screening measures to improve teacher and school quality. At present, Colombia (since 2002), Ecuador (2007), Mexico (2008) and Peru (2012) all require teacher candidates to pass national mandatory tests before they can opt for long-term careers at public schools (Bruns and Luque, 2015; Elacqua *et al.*, 2017). Nevertheless, much remains to be learned about the effects of these teacher recruitment policies, since the vast majority have not been subject to evaluation.

In this article, we evaluate whether teachers who passed national entry tests and were tenured by Ecuador's new recruitment policy (henceforth, *test-screened tenured* teachers) have positive effects on student learning outcomes by linking unique administrative teacher records to the rich experimental data produced by the "Closing Gaps" project. In the project, a representative cohort of Ecuadorian kindergarten children were assigned to their teachers in the 2012-2013 school year, using a rule that is as good as random (Araujo *et al.*, 2016). We confirm that successful random assignment of these kindergarten children to *test-screened tenured* teachers is also given in our data, which allows us to estimate teacher causal effects and prevent potential bias caused by the matching of teachers to students. Our results show that kindergarten students benefit from randomly-assigned teachers who passed mandatory entry tests and won merit-based competitions for tenure in Ecuador, having significantly higher end-of-year test scores of at least a 0.105 standard deviation in language and a 0.085 standard deviation in math.

In addition to the experimental nature of the data, the "Closing Gaps" project collected rich information on students, families and teachers, which allows us to check the robustness of our experimental estimations. Moreover, we not only have access to common observable teacher characteristics as such as gender, education and experience, but also measurements of teacher cognitive ability (Wechsler Adult Intelligence Scale, WAIS-III), personality traits (the Big Five personality test) and classroom practices (the Classroom Assessment Scoring System, CLASS). The positive and significant effects of teachers tenured by the new recruitment policy persist even after controlling for the full set of teacher characteristics.

Our estimations are highly robust to several specifications. We test a different control group specification and confirm that *test-screened tenured* teachers significantly outperform teachers tenured before the policy implementation who were not required to pass national entry tests, contract teachers who had passed national entry exams but had not yet won a completion for tenure, and contract teachers who had not gone through the screening process. In addition, we run our analysis for the subsample of schools that had at least one teacher tenured by the new recruitment policy. The size and significance of our estimates are consistent for all specifications.

This paper makes three key contributions to teacher quality research in the context of the education production function literature. First, we provide the first experimental estimations of the effects of teachers screened by skill and subject knowledge tests and tenured by merit-based competitions in a Latin American country. Accordingly, our results add to the findings of the recent research on personnel policies that has shown that combining teacher background characteristics and screening measures in recruitment processes can strongly predict future teacher effectiveness (Goldhaber, Grout and Huntington-Klein, 2017; Jacob *et al.*, 2018; Bruno and Strunk, 2019). Furthermore, we provide an insight into the effectiveness of the new teacher recruitment policies implemented in the region in recent years.

Second, we contribute to the current debate around the teacher characteristics associated with student learning in the developing world. We show that teachers screened by entry tests who competed to earn a permanent job position have positive and significant effects on student learning in Ecuador. This outcome is to some extent aligned with previous findings on the positive learning effects associated with teacher skill and subject knowledge in developing countries (Metzler and Woessmann, 2012; Glewwe *et al.*, 2014; Bietenbeck, Piopiunik and Wiederhold, 2018; Bau and Das, 2020). However, our results also provide new evidence of the connection between teacher qualification, job status and performance. Contrary to recent evidence on the positive effects of fixed-term contract teachers in developing countries (Muralidharan and Sundararaman, 2013; Duflo, Dupas and Kremer, 2015), our estimations show that tenured teachers significantly outperform contract teachers in Ecuador.

Third, our study also confirms the potential effectiveness of highly-qualified teachers in closing learning gaps between socioeconomically advantaged and disadvantaged children in the developing world. We find that the effects of *test-screened tenured* teachers on language learning are stronger for children who started the school year with lower baseline test scores or came from socioeconomically disadvantaged households.

The remainder of the paper is structured as follow. In Section 3.2, we provide background on Ecuador's new teacher recruitment policy and previous evidence connected to our research question. Section 3.3 reviews the "Closing Gaps" data

experimental design and assesses the validity of the experiment. Section 3.4 presents our model, estimation strategy and results. Section 3.5 details our heterogeneous effects analysis. Section 3.6 presents our robustness check, and finally Section 3.7 concludes.

3.2 Background and Evidence

3.2.1 New Teacher Recruitment Policy in Ecuador

The Ecuadorian government implemented major education reforms between 2006 and 2017 to increase enrollment and improve learning outcomes (Araujo P. and Bramwell, 2015; Schneider, Cevallos Estarellas and Bruns, 2019). There is evidence of the success achieved in education expansion: while kindergarten and primary school net enrollment remained above 90 percent, secondary school net enrollment increased from 62 percent in 2005 to 85 percent in 2017 (World Bank, 2019). Nonetheless, questions about the reforms' impact on the system quality remain open. These reforms particularly targeted the teaching profession, whose prestige and quality had progressively declined since the 1970s due to the lack of academic standards for teacher pre-service programs and the drastic decrease in teaching wages (Elacqua *et al.*, 2017).

In November 2007, the Ecuadorian government issued the Executive Order No. 708[37], which required teacher candidates to pass mandatory skill and content knowledge tests before they were allowed to participate in merit-based selection competitions for tenure at public schools. The new regulation was exclusively applied to teachers seeking tenure[38] from December 2007 onwards. Already-tenured teachers were exempted from the regulation[39]. The policy reform was institutionalized in 2011, when Ecuador's new Intercultural Education Law (*Ley Orgánica de Educación*

[37] The Executive Order No. 708 reformed the Regulation to Ecuador's Education Law of 1990 (*Reglamento de Ley de Carrera Docente y Escalafón del Magisterio Nacional*). Even though Ecuador's Education Law of 1990 already established that teachers should be tested and recruited through merit-based competitions, national entrance examinations for teacher candidates were not implemented until the release of the Executive Order No. 708 in 2007.

[38] In Ecuador, tenured teachers hold permanent job positions in public educational institutions.

[39] Tenured teachers who wanted to be transferred to another school were also required to take the exams and compete for an available position.

Intercultural, LOEI) ratified national entrance exams as a mandatory requirement for teacher candidates (Asamblea Nacional del Ecuador, 2011).

Prior to December 2007, teacher selection processes were locally organized by Provincial Directorates of Education without national standards other than academic degree requirements. After December 2007, teacher recruitment processes were centrally organized by Ecuador's Ministry of Education. Permanent teacher vacancies had to be filled by teachers who had passed national entrance exams and won merit-based selection competitions. Local education authorities were only allowed to hire teachers who had not undergone the new recruitment process with fixed-term contracts until the positions were permanently filled by tenured teachers.

Ecuador's Ministry of Education regulated the new competitive teacher recruitment processes through Ministerial Resolutions (*Acuerdos Ministeriales*, *AM*) and organized them into stages or components (Ministerio de Educación del Ecuador, 2007, 2008, 2010, 2011). In a first stage, teacher candidates were required to pass a logical-verbal reasoning test, a pedagogical knowledge test and a subject-specific knowledge test. Teacher candidates who achieved a required minimum score became eligible candidates to compete for a permanent position. Test scores were part of a total competition score for each candidate. A second stage of the recruitment process comprised evaluating the eligible candidates' credentials: academic degrees, teaching experience, additional training courses and academic publications. Credentials were graded and added up to the total competition score. Finally, teachers were also required to present a demonstration class in front of a school board,[40] which was also evaluated and added to the total competition score. The demonstration class was part of the first stage in the beginning of the policy implementation, but afterwards it became a third stage of the process itself. In each merit-based competition, the eligible candidate who achieved the highest total competition score was entitled to a tenured position.

In addition, it is worth mentioning that the stages of the recruitment process were highly automated and transparent. Teacher candidates were required to open an

[40] The school board comprised the school principal or deputy, a peer teacher and two parents elected by the school's general assembly. For positions in lower and upper secondary education, a student was also included.

account on the Ministry's online recruitment platform in order to register for eligibility tests. In the first stage, standardized exams were scanned and graded using a specialized software. Afterwards, test scores were automatically uploaded and published on the Ministry's online platform. In the second stage, eligible candidates were allowed to apply for a permanent position from among all open vacancies at public schools through the Ministry's online platform, and were also required to upload their credentials and certificates. Even though personnel from the Provincial Directorates of Education were in charge of checking that information and corresponding certificates were accurate, a software incorporated into the Ministry's online platform was used to score candidates' credentials. In the third stage, school boards were required to use a standardized instrument provided by the Ministry to assess the demonstration class. Assessment sheets were scanned, automatically graded and uploaded to the Ministry's online platform. Teacher candidates were able to follow the recruitment process at each stage on the Ministry's online platform and had access to the scores of all candidates participating in their specific competitions. Moreover, teacher candidates had the right to appeal the results at each stage of the process.

The recruitment process stage weighting was slightly changed between 2007 and 2013. Nonetheless, test scores permanently represented the highest weighting, going from 45 to 55 percent of the total competition score. Table 2.1 of Chapter 2 describes the Ministerial Resolutions applied from the beginning of the process to the first semester of 2013.[41] Our analysis focuses on this period because it covers all of the possible recruitment processes for teachers tenured after December 2007 and employed in the 2012-2013 school year.

Along with the new recruitment regulation, the Ecuadorian government opened around 34,000 new permanent teacher positions between 2007 and 2012 (Ministerio de Educación del Ecuador, 2012). Strong economic incentives to attract highly competitive teacher candidates into the public educational system were also introduced.

[41] The competitive teacher recruitment process is still in place in Ecuador. In July 2013, it was relaunched as the "I Want to be Teacher" (*Quiero Ser Maestro*) competition. Since then, entry tests have been independently designed by the National Institute of Education Evaluation (*Instituto National de Evaluación Educativa, INEVAL*).

The nominal monthly entry wage for a new tenured teacher steadily increased from US$291 in 2006 to US$396 in 2010 and quite strongly to US$775 in 2011, when Ecuador's new Intercultural Education Law homogenized the teacher payment scale in line with the public service payment scale (Ministerio de Educación del Ecuador, 2012; Schneider, Cevallos Estarellas and Bruns, 2019). [42] The incentives were very attractive to contract teachers[43] working at public schools, teachers working at private schools, recently-graduated teachers, and university graduates who did not hold teaching degrees but were specialists in subjects taught at schools[44].

The first years of the new Ecuadorian competitive teacher recruitment process were highly competitive. Between 2007 and 2012, 320,000 teacher candidates registered for eligibility tests, 21,200 eligible candidates passed entry tests and 18,820 successful candidates were granted a permanent teaching position (Ministerio de Educación del Ecuador, 2012).

3.2.2 Previous Policy Evaluation Studies in Ecuador

There is scarce and mixed evidence on the effects of mandatory certification tests and competitive teacher recruitment in Ecuador. Cruz-Aguayo, Ibarrarán and Schady (2017) use data on a representative sample of children in first primary school grades to analyze whether children taught by teachers with higher test scores in Ecuador's new competitive recruitment had higher achievements in language and math in the 2011-2012 school year. They report no indication that teachers with higher (or lower) test scores were assigned to children with different observable characteristics, which allows them to estimate level and value-added to achievement models with OLS regressions. Their results do not suggest that test scores on the teacher entry competition were associated with child achievement in language or math. Subsequently, Cruz-Aguayo et al. (2017) conclude that the instrument used to decide which teachers receive

[42] Teaching wage increases were not the product of high inflation rates. Ecuador's average inflation rate between 2006 and 2011 was 4.37 percent (INEC, 2020).

[43] Contract teachers are fixed-term employed teachers. The regular contract period is one year with the possibility of renewal.

[44] Ecuador's new Intercultural Education Law of 2011 officially opened the teaching career to university graduates who did not hold teaching degrees.

tenure in Ecuador does not predict how effective a teacher is at raising math and language achievement. Nonetheless, a serious limitation with this study is that it only compares student outcomes among successful teacher candidates who passed entry tests, who therefore belong to less than 10 percent of all tested teachers. Since test score data from teachers who did not pass entry exams is not available in the study, its findings do not seem to fully support its conclusion.

Araujo P. (2019) assesses the effectiveness of Ecuador's teacher recruitment process as a quality screening device using the same school sample, but taking into account information on teachers who were not recruited through the selective entry competitions and were working at the same schools and grades in the 2011-2012 academic year. Her analysis shows that students assigned to teachers tenured by the new recruitment policy had on average parents with fewer years of education and were more likely to live in poor households, which suggests a matching of more vulnerable students to these teachers. Using propensity score matching to estimate a value-added to student achievement model, her results suggest that teachers who passed national entrance examinations and won tenured positions were more effective in raising language achievement among students living in poverty. The average treatment effect of a test-screened teacher who won an entry competition is estimated to be at least a 0.09 standard deviation gain in language for a student living in a poor household. By contrast, no effect was found for math achievement.

Nonetheless, the aforementioned studies and much of the teacher quality literature suffer from potential bias caused by the non-random assignment of students to teachers. We address this issue by using the data gathered by the "Closing Gaps" project, where a representative cohort of Ecuadorian kindergarten children were assigned to their classes and teachers with a rule as good as random starting in the 2012-2013 school year. Araujo et al. (2016) were the first to use the "Closing Gaps" data to examine the impact of teacher quality on learning outcomes in kindergarten. Their study finds teacher effects of a 0.09 standard deviation in language and math learning. The results also suggest that children assigned to teachers with higher classroom practice scores had significantly higher achievement. A one standard deviation higher teacher score is associated with higher student end-of-year test scores of between a 0.06 and

0.08 standard deviation. By contrast, children assigned to inexperienced teachers had test scores that were 0.17 standard deviations lower. None of the other teacher characteristics analyzed – including cognitive skills and personality traits – were associated with student learning.

We combine the "Closing Gaps" data with administrative teacher recruitment information from Ecuador's Ministry of Education. This allows us to estimate the causal effect on student learning of teachers screened by national entrance examinations and tenured through the new competitive recruitment process.

3.3 Experimental Design and Data

3.3.1 Background on the "Closing Gaps" Project

The Inter-American Development Bank (IDB) in cooperation with Ecuador's Ministry of Education started in 2011 the "Closing Gaps" project, a longitudinal experimental study to evaluate different dimensions of teacher quality in public schools, starting with the first grade of general basic education or kindergarten.[45]

The "Closing Gaps" Project randomly chose a sample of 204 public schools from the coastal region of Ecuador for its implementation.[46] Even though the study sample was drawn from the coastal region, children and households in the study were generally similar to national samples (Araujo *et al.*, 2016). The sample was limited to schools that had at least two kindergarten classes. Starting in the 2012-2013 school year, 14,930 children enrolled in kindergarten in the participating schools were randomly assigned to their classrooms and teachers. In terms of assignment, children enrolled for kindergarten in a given school were ordered by their last and first names,

[45] The Ecuadorian education system is organized at three levels: initial education, general basic education and high school (Asamblea Nacional del Ecuador, 2011). The initial education or early education serves children under 5 years of age (equivalent to the ISCED level 0). The general basic education starts at 5 years of age and comprises one year of kindergarten, six years of primary education (ISCED level 1) and three years of lower secondary education (ISCED level 2). Finally, high school corresponds to three years of upper secondary education (ISCED level 3).

[46] Ecuador has four natural regions: Coastal (Costa), Andean (Sierra), Amazon (Amazonía) and Insular (Islas Galápagos). Due to the particular weather conditions of each region, the school year starts in different months. The 2012-2013 school year started in April 2012 and ended in February 2013 in the Coastal and Insular regions. The same school year in the Andean and Amazon Regions started in September 2012 and ended in June 2013.

and then assigned to kindergarten classrooms going down the list in alternating order. Compliance with the assignment rule was very high: only 1.7 percent of children were found in classrooms other than those to which they had been assigned (Araujo *et al.*, 2016). [47]

3.3.2 Data on Children and Families

At the beginning of the 2012-2013 school year, the "Closing Gaps" project collected baseline data on characteristics of children's age, gender and attendance to early education. Children were also tested with the *Test de Vocabularion en Imagenes Peabody (TVIP),* a measurement of children's past learning. [48] The project also ran a household survey that included questions on parents' education and living conditions. We use information on household assets and access to basic services from this survey to calculate a household living standard indicator, with a scale from 0 to 6.[49]

To measure student learning at the end of the school year, four tests of language and early literacy and four tests of math were applied to children individually.[50] The language and early literacy tests covered child vocabulary, oral comprehension, and sound, letter and word recognition.[51] The math tests covered number recognition,

[47] No-compliers in our analysis are assigned to the classrooms to which they were originally randomly assigned. This means that we actually estimate intention to treat. Nonetheless, our intention to treat estimators should be very close to the average treatment effects, since compliance is almost 100 percent.

[48] It is the Spanish version of the Peabody Picture Vocabulary Test (PPVT) (Dunn *et al.*, 1986). The "Closing Gaps" project provided TVIP test scores standardized on a sample of Mexican and Puerto Rican children, whose mean was set at 100 and the standard deviation at 15 at each age. In the standardizing procedure, some of the observations were imputed to the lowest or highest possible value if the raw score obtained by the child did not have a correspondence to the standardized score in the population of reference.

[49] The household living standard indicator is based on the Global Multidimensional Poverty Index (MPI) developed for the Human Development Reports (HDR) by the Human Development Report Office at the United Nations Development Programme (UNDP), in collaboration with the Oxford Poverty & Human Development Initiative (OPHI) (Alkire *et al.*, 2016). It aggregates the following households' characteristics: access to improved sanitation and safe drinking water, type of floor, roof and exterior walls material, and asset ownership.

[50] Tests of children's inhibitory control, working memory, capacity to pay attention, and cognitive flexibility were also applied. These processes are jointly known as executive function. Our study did not find effects on children's executive function. Results are available upon request.

[51] The TVIP was applied to evaluate child vocabulary. Oral comprehension, and sound, letter and word recognition tests were taken from the Spanish-speaking version of the Woodcock-Johnson battery of tests of child development and achievement (Muñoz-Sandoval *et al.*, 2005) and an adapted version of the Early Grade Reading Assessment (RTI International, 2009b).

sequencing, applied math problems and identifying basic geometric figures. [52] Test aggregates for language and math were normalized to have zero mean and unit standard deviation.

As indicated, 14,930 children were enrolled in kindergarten in the participating schools; however, 740 children did not start the academic year in these schools. We have full information for around 98 percent of the children who enrolled and actually started the school year in the participating schools, and household information on about 94 percent of them. However, our student sample size decreases to 12,632 children when additional information on teacher characteristics is taken into account in further analyses.

Table 3.1: Summary Statistics for Children and Families' Characteristics

	Full sample		Analysis sample		
	Mean (SD)	Obs.	Mean (SD)	Obs.	Difference
	(1)	(2)	(3)	(4)	(5)
Children:					
Proportion female	0.49	14930	0.49	12632	0.00
	(0.50)		(0.50)		(1.00)
Proportion who attended early education	0.56	14925	0.61	12632	0.05***
	(0.50)		(0.49)		(0.00)
Age (months)	60.34	14841	60.23	12632	-0.11*
	(5.11)		(4.92)		(0.07)
TVIP	83.24	14187	83.32	12632	0.08
	(16.89)		(16.85)		(0.70)
Family:					
Parents' years of schooling	8.69	13275	8.66	12632	-0.03
	(3.42)		(3.43)		(0.48)
Living Standard Indicator	3.33	13744	3.32	12632	-0.01
	(1.38)		(1.38)		(0.56)

Note: This table reports means and standard deviations (in parenthesis) of the characteristics of kindergarten children and their families for the original and analysis samples. Column (5) displays differences between the original and analysis samples, and the respective p-value (in parenthesis) from a t-test for equality. TVIP stands for *Test de Vocabulario en Imágenes Peabod*y, the Spanish version of the Peabody Picture Vocabulary Test (PPVT). The test was standardized to have a mean of 100 and the standard deviation of 15 at each age, based on a reference sample of Mexican and Puerto Rican children. Family living standard indicator aggregates the following households' characteristics: access to improved sanitation and safe drinking water, type of floor, roof and exterior walls material, and assets ownership. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

[52] All of the math tests were taken from the Spanish-speaking version of the Woodcock-Johnson battery (Muñoz-Sandoval *et al.*, 2005) and an adapted version of the Early Grade Math Assessment (RTI International, 2009a).

Table 3.1 reports summary statistics for children and families' characteristics of the full and analysis samples. Children in the analysis sample are approximately 5 years old at the beginning of the school year. Girls account for almost half of the sample and around 61 percent of children attended early childhood education. The average TVIP score is approximately 83, one standard deviation lower than the Mexican and Puerto Rican reference population used to norm the test. Parents had completed almost nine years of education on average. In addition, column (5) of table 3.1 displays differences between the full and analysis samples, and the respective p-value from a t-test for equality. We observe that our analysis sample is not substantially different from the original sample, with the only exception of the proportion of children who attended early childhood education, which is about five percentage points higher in the analysis sample.

3.3.3 Data on Teachers

At the beginning of the school year, the project also collected data on conventional teacher characteristics such as gender, experience and education. Additional rich non-conventional data on teachers' cognitive ability and personality was collected by the end of the school year. The Spanish version of the Wechsler Adult Intelligence Scale (WAIS-III) was applied to measure teacher cognitive skills. The NEO PI-R psychometric instrument (Costa Jr. and McCrae, 2008) was used to assess teacher personality by measuring the so-called Big Five personality traits of neuroticism, extraversion, openness, agreeableness and conscientiousness.[53] Under the NEO PI-R profile scale, each trait can be scored as very low (20-35), low (35-45), average (45-55), high (55-65) or very high (65-80). We normalized the cognitive skill

[53] The APA Dictionary of Psychology (VandenBos, 2007) defines the traits as follows: i. Neuroticism: chronic level of emotional instability and proneness to psychological distress; ii. Extraversion: an orientation of one's interest and energies toward the outer world of people and things rather than the inner world of subjective experience, characterized by positive affect and sociability; iii. Openness to Experience: the tendency to be open to new aesthetic, cultural or intellectual experiences; iv. Agreeableness: the tendency to act in a cooperative, unselfish manner; v. Conscientiousness: the tendency to be organized, responsible and hardworking. Psychology and economics' research have well established that standardized tests of cognitive ability and personality traits predict a variety of job performance outcomes across professions (Barrick and Mount, 1991; Heineck and Anger, 2010; Kuncel, Ones and Sackett, 2010; Almlund *et al.*, 2011).

and the Big Five personality trait scores to have zero mean and a unit standard deviation for our regression analyses.

Finally, to retrieve information on teacher behaviors and classroom practices, the CLASS was applied in the middle of the school year. CLASS measures teacher behaviors in three domains: emotional support, classroom organization and instructional support (Hamre, La Paro and Pianta, 2007). Each domain's score ranges from low (1-2) to medium (3-5) or high (6-7). As part of the standard application of CLASS, all teachers were filmed teaching for a full-day during the 2012-2013 school year.[54] The videos were coded by experts according to the CLASS protocol and scores for each domain were calculated. All kindergarten teachers in the school sample were also filmed in the previous 2011-2012 school year, although there are fewer observations than for the 2012-2013 school year due to changes in the allocation of teacher staff within schools and teachers leaving the participating schools.[55] We normalized the total CLASS scores to have zero mean and a unit standard deviation for our regression analyses.

We obtained additional unique information from administrative records of Ecuador's Ministry of Education on the tenure status and recruitment processes of all teachers participating in the "Closing Gaps" project at the beginning of the 2012-2013 school year.[56] In our sample, about 13 percent of teachers had passed national entry tests and won a merit-based competitions organized by the Ministry of Education between 2007 and 2013, which granted them tenure.

[54] Teachers did not have previous knowledge about the day on which they would be filmed.

[55] The full sample of teachers has CLASS scores for the 2012-2013 school year; however, only 341 of them have CLASS scores for the previous 2011-2012 school year.

[56] In response to our request, the National Directorate of Educational Research of Ecuador's Ministry of Education merged the "Closing Gaps" project data with administrative records of teacher recruitment processes that took place between 2007 and 2013 and were stored in the *Sistema de Información del Ministerio de Educación* (SIME) platform. Personnel of the National Directorate of Educational Research used the teacher's personal ID number as unique identifier in the merging process.

Table 3.2 reports summary statistics for teacher characteristics. Virtually all teachers are females.[57] Around 58 percent are tenured. On average, they have 15 years of experience and almost all of them have a university degree. The mean cognitive skill score is around 87, which is in the low average range of the WAIS-III international scale (Wechsler and Psychological Corporation, 1997).[58] With respect to personality traits, the NEO PI-R profile scale shows a high mean score for conscientiousness, a low mean score for neuroticism, and average mean scores for openness, extraversion and agreeableness. CLASS average scores are in the medium range for socioemotional support and class management, and in the low range for instructional support in both school years. The CLASS scores show that even though teachers maintain positive relations with their students and moderately well-organized classrooms, they engage in very few interactions that support learning. We have full information on 430 teachers in the 2012-2013 school year.[59]

In addition, table 3.2 displays summary statistics of the analyzed sample by whether the teacher passed national entry tests and won a competition for tenure. Some characteristics significantly differ between the group of *test-screened tenured* teachers and their colleagues. *Test-screened tenured* teachers have on average about two years' less experience, and all of them have university degrees and tenured positions. In terms of personality, on average they have significantly higher cognitive skill scores and are more open, extroverted and agreeable than their peers. Moreover, they have higher CLASS scores, which are statistically significant only for the 2011-2012 school year. *Test-screened tenured* teachers have also slightly more years of education on average, are more conscientious and less neurotic, although these differences are not statistically significant in our sample.

[57] Traditionally, teachers in the first school grades in Ecuador are women. Nonetheless, as a robustness check, we ran all of our analyses only for the female teacher sample. The results were consistent and are available upon request.

[58] The average teacher cognitive skill score is not a source of concern for this study. Vast neuropsychological literature has shown that there are cultural differences in cognitive test performance. The performance of people from different cultures in the same cognitive test may vary according to the importance of the specific cognitive ability in one's own culture (Rosselli and Ardila, 2003; Bakos *et al.*, 2010; Fasfous *et al.*, 2013).

[59] We start our teacher sample with 450 observations; however, only 430 teachers have cognitive skill and personality trait scores.

Table 3.2: Summary Statistics for Teacher Characteristics

	Full sample	Test-screened tenured teacher		
		YES	NO	Difference
	(1)	(2)	(3)	(4)
Proportion female	0.99	0.96	0.99	-0.03
	(0.10)	(0.19)	(0.07)	(0.23)
Proportion tenured	0.58	1.00	0.52	0.48***
	(0.49)	(0.00)	(0.50)	(0.00)
Years of experience	14.74	12.84	15.01	-2.17**
	(8.60)	(7.08)	(8.78)	(0.04)
Years of education	17.15	17.46	17.11	0.36
	(1.92)	(1.90)	(1.92)	(0.20)
University degree	0.99	1.00	0.99	0.01**
	(0.11)	(0.00)	(0.11)	(0.03)
Cognitive skills	86.46	89.93	85.96	3.96***
	(9.53)	(9.73)	(9.40)	(0.01)
Neuroticism	43.85	43.22	43.94	-0.73
	(6.72)	(6.80)	(6.72)	(0.47)
Extraversion	45.65	48.55	45.24	3.31***
	(6.83)	(6.46)	(6.79)	(0.00)
Openness	50.82	52.97	50.51	2.46**
	(6.75)	(6.52)	(6.73)	(0.01)
Agreeableness	48.22	50.63	47.87	2.76***
	(7.56)	(7.02)	(7.58)	(0.01)
Conscientiousness	57.55	58.64	57.39	1.25
	(8.15)	(6.99)	(8.30)	(0.24)
CLASS average	3.48	3.50	3.48	0.02
	(0.28)	(0.27)	(0.28)	(0.62)
CLASS 2011-2012	3.62	3.80	3.61	0.19**
	(0.37)	(0.35)	(0.37)	(0.03)
Socio emotional support	4.30	4.48	4.29	0.19*
	(0.40)	(0.44)	(0.39)	(0.07)
Classroom management	4.99	5.26	4.97	0.29**
	(0.63)	(0.55)	(0.63)	(0.03)
Instructional support	1.36	1.44	1.36	0.08
	(0.27)	(0.26)	(0.27)	(0.18)
CLASS 2012-2013	3.41	3.45	3.41	0.04
	(0.28)	(0.25)	(0.29)	(0.24)
Socio emotional support	4.07	4.11	4.07	0.05
	(0.33)	(0.25)	(0.34)	(0.23)
Classroom management	4.79	4.87	4.78	0.09
	(0.47)	(0.46)	(0.47)	(0.21)
Instructional support	1.15	1.15	1.15	-0.00
	(0.18)	(0.15)	(0.18)	(0.96)
N	430	54	376	430

Note: This table reports means and standard deviations (in parenthesis) of teacher characteristics for the full sample, as well as whether the teacher passed national entry tests and won a competition for tenure (*test-screened tenured* teacher). Column (4) displays differences between *test-screened tenured* and non- *test-screened tenured* teachers, and the respective p-value (in parenthesis) from a t-test for equality. Cognitive skills were measured with the Spanish version of the Wechsler Adult Intelligence Scale (WAIS-III). The test is internationally normed so that 100 is the median score for the adult population. The Big Five personality trait scores (neuroticism, extraversion, openness, agreeableness and conscientiousness) were obtained with the NEO PI-R psychometric instrument. Each personality trait can be scored as very low (20-35), low (35-45), average (45-55), high (55-65) or very high (65-80). CLASS stands for Classroom Assessment Scoring System. CLASS domains can be scored as low (scores 1-2), medium (3-5) or high (6-7). CLASS 2011-2012 statistics based on 341 observations. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

3.3.4 Validity of the Experimental Design

The successful randomization of students into classrooms and to teachers within schools is the fundamental condition behind the validity of the causal inferences that we aim to establish. To evaluate whether the random assignment to *test-screened tenured* teachers was successfully implemented, we test for balance in predetermined variables across classrooms. For this purpose, we regress teacher *test-screened tenured* status on student and family predetermined variables. We condition our estimations on school fixed effects because the random assignment of students to teachers was conducted within schools.

Table 3.3: Randomization Test

	Test-screened tenured teachers
Children:	
Age (months)	-0.000
	(0.000)
Gender	0.001
	(0.003)
TVIP	0.000
	(0.000)
Proportion who attended preschool	-0.005
	(0.005)
Family:	
Parents' years of schooling	0.001**
	(0.001)
Living standard indicator	-0.000
	(0.003)
Observations	12632
R^2	0.580
F	1.09
p	0.372

Note: OLS model estimated with clustered standard errors (in parentheses) at the school level and school fixed effects. TVIP stands for Test de Vocabulario en Imágenes Peabody, the Spanish version of the Peabody Picture Vocabulary Test (PPVT). Total number of observations restricted to student sample with full information on student, parent, classroom and teacher characteristics. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 3.3 provides the results of our randomization test. As shown, none of the student or family characteristics predict the likelihood that a child is assigned to a *test-screened tenured* teacher, with the exception of parents' years of schooling.

Nonetheless, this variable has a regression coefficient close to zero. Moreover, an F-test for the joint significance of all of the predetermined demographic variables is statistically insignificant (p=0.32). We conclude that the random assignment to *test-screened tenured* teachers was successful, and that there is no threat to our identification strategy.

Table 3.4: No-Show, Attrition and Late Enrollment Tests

	No-shows		Attritors		Late enrollments	
Teacher	(1)	(2)	(3)	(4)	(5)	(6)
Test-screened tenured	-0.005	-0.007	-0.004	-0.003	-0.004	0.000
	(0.007)	(0.007)	(0.007)	(0.007)	(0.013)	(0.013)
Female		-0.020		0.006		-0.019
		(0.019)		(0.019)		(0.023)
Years of experience		0.000		-0.000		0.001**
		(0.000)		(0.000)		(0.000)
Years of education		0.001		-0.001		-0.001
		(0.002)		(0.001)		(0.001)
Cognitive skills		-0.002		-0.000		-0.002
		(0.003)		(0.003)		(0.003)
Neuroticism		-0.002		-0.001		0.001
		(0.003)		(0.002)		(0.003)
Extraversion		0.002		0.002		-0.005
		(0.003)		(0.002)		(0.003)
Openness		0.002		-0.001		0.003
		(0.003)		(0.002)		(0.004)
Agreeableness		0.005		-0.000		0.001
		(0.003)		(0.002)		(0.004)
Conscientiousness		0.002		0.002		0.001
		(0.003)		(0.002)		(0.003)
CLASS average		-0.001		0.007***		-0.000
		(0.003)		(0.003)		(0.004)
Observations	14909	14235	14170	13543	14495	13855
R^2	0.045	0.044	0.033	0.034	0.045	0.048
F	0.55	1.32	0.42	1.46	0.11	1.11
p	0.459	0.215	0.519	0.150	0.743	0.356

Note: OLS linear probability models estimated with clustered standard errors (in parentheses) at the school level and school fixed effects. No-shows is a dummy variable that takes the value of one if an enrolled student did not show up in the beginning of the school year. Attritors is a dummy variable that takes the value of one if the student dropped out of the school. Late enrollment is a dummy variable that takes the value of one if the student enrolled after the school year started. CLASS stands for Classroom Assessment Scoring System.* Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

We also look at the sample attrition, which poses another threat to the experimental design. In our sample, we have children who were enrolled in kindergarten in the participating schools but never showed up (5 percent of the original

sample), children who dropped out during the school year (2.8 percent), and finally children who were enrolled later (5 percent). We evaluate whether being a no-show, attritor or late enrollment is correlated with assignment to our treatment by running individual linear probability regressions of these conditions on the *test-screened tenured* teacher status, controlling for school fixed effects. Our results are presented in table 3.4. Columns 1, 3 and 5, respectively, show that correlations between no-shows, attritors or late enrollments and *test-screened tenured* teacher status are statistically insignificant and virtually equal to zero. These results do not change even when we add additional teacher characteristics in Columns (2), (4) and (6). We conclude that there is no evidence that the decisions for being a no-show, attritor or late enrollment are affected by being randomly assigned to a *test-screened tenured* teacher.

3.4 Estimation Strategy and Results

3.4.1 Estimates of Test-Screened Tenured Teacher Effects

We evaluate the effect of *test-screened tenured* teachers on learning outcomes by estimating a value-added to student achievement specification of the education production function –formalized by Todd and Wolpin (2003) – with the following OLS regression:

$$Y_{ics} = \alpha_s + test\_tenured_{cs}\beta_1 + X_{ics}\beta_2 + \bar{X}_{ics}\,\beta_3 + C_{cs}\beta_4 + T_{cs}\beta_5 + P_{cs}\beta_6 + CLASS_{cs}\beta_7 + u_{ics}, \quad (3.1)$$

where Y_{ics} represents the end-of-year test score in language or math of child *i* in classroom *c* in school *s*. $Test\_tenured_{cs}$ is a dummy variable indicating whether the student was assigned to a *test-screened tenured* teacher and α_s is a school fixed effect component, which must be included because the random assignment of students to teachers was conducted within schools. Given that students were randomly assigned to *test-screened tenured* teachers, we do not need to include any additional control in the regression. However, we also estimate our model with additional covariates to examine the robustness of our results and increase the precision of the estimates (Duflo, Glennerster and Kremer, 2008). We gradually include as controls in the regression a

vector of observable student and parent characteristics (X_{ics}),[60] a vector of classroom averages of student and parent characteristics ($\bar{X}_{ics}$), an indicator of class size (C_{cs}), a vector of teacher observable characteristics (T_{cs}), a vector of teacher cognitive ability and personality (P_{cs}) and an indicator of teacher class practices ($CLASS_{cs}$).

The fact that measurements of teacher cognitive skills and personality were collected by the end of the school year could raise concern about reverse causality, in case one assumes that these characteristics could be affected by the classroom environment. However, it should be noted that we do not try to find the causal effect of teacher cognitive skills and personality on student learning, but that we add these covariates at the end of our estimations to observe whether our experimental results are affected. In addition, despite evidence on a change in personality over the life cycle, most of the evidence does not point to single environmental events as the source of the change but rather a combination of factors importantly influenced by genetics (Almlund *et al.*, 2011).

The same reverse causality concern could apply to the information on teacher practices, given that CLASS was applied in the middle of the 2012-2013 school year. We use the average CLASS scores of 2011-2012 and 2012-2013 to deal with this problem, as suggested by Araujo et al. (2016).[61] Furthermore, we add our CLASS covariate at the end of our estimations to check where our results importantly change.

Standard errors are clustered at the school level in all regressions. We take this approach because although treatment occurs more precisely at the classroom level, clustering by school is a more conservative estimate of standard errors that takes into account cross-classroom correlations in errors within schools (Chetty *et al.*, 2011).[62] In

[60] Student and parent characteristics included: child baseline TVIP score, age, gender and attendance to preschool, parents' average years of education and living standard indicator.

[61] Our findings do not change if we run the entire analysis with the lagged CLASS score of 2011-2012 instead of the average CLASS score. Results for math are even stronger in significance level when the lagged CLASS score is used for the estimations. However, the sample is restricted to teachers who taught kindergarten in the school sample for the entire 2011-2012 and 2012-2013 school years. All results are available upon request.

[62] We also estimated all the regressions with standard errors cluster at the classroom level. The results are in line with those provided here, but with higher significance level for the variables of interest. They are available upon request.

addition, our school sample comprises purely treated schools where all kindergarten classrooms are taught by a *test-screened tenured* teacher, purely control schools where no kindergarten classroom is taught by a *test-screened tenured* teacher, and schools with a combination of treated and control kindergarten classrooms.

As previously mentioned, the final student sample size of 12,632 children in our estimations is limited to the number of observations obtained after including all student, parent, classroom and teacher controls.[63]

We report regression results for language learning in table 3.5. Column (1) presents the effect of *test-screened tenured* teachers on language learning estimated without additional controls and taking into account school fixed effects. We find a significant and positive causal effect of *test-screened tenured* teachers on language learning at the 5 percent significance level. Children assigned to *test-screened tenured* teachers have a 0.105 standard deviation higher end-of-year test score in language. We present regression results that incorporate controls for student characteristics in Column (2), family characteristics in Column (3), and classroom characteristics in Column (4). Here, the results do not substantially change in terms of significance or size. In Column (5), we include controls for additional teacher observable characteristics of gender, experience and education. We find that the effect of *test-screened tenured* teachers is significant and its size increases to a 0.115 standard deviation higher end-of-year language test score. Subsequently, we add controls for teacher cognitive ability and personality in Column (6). The effect does not change in significance or size. Finally, in Column (7) we introduce our measure of teacher classroom practices (CLASS score) as a control. Interestingly, we still find a significant and positive effect of *test-screened tenured* teachers, which increases to a 0.125 standard deviation higher end-of-year language test score.

[63] We also estimated all regressions without limiting the sample size to the final number of observations obtained after a full set of controls are included. All results reconfirm our findings here and are available upon request.

Table 3.5: Estimates of Effects of Test-Screened Tenured Teachers on Language

Teacher	Language						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Test-screened tenured	0.105**	0.102***	0.096**	0.089**	0.115***	0.115***	0.125***
	(0.043)	(0.038)	(0.038)	(0.039)	(0.037)	(0.038)	(0.035)
Female teacher					0.245**	0.211*	0.162
					(0.113)	(0.108)	(0.101)
Years of experience					0.005**	0.004**	0.003*
					(0.002)	(0.002)	(0.002)
Years of education					0.001	-0.001	-0.002
					(0.007)	(0.007)	(0.007)
Cognitive skills						0.038***	0.037**
						(0.014)	(0.014)
Neuroticism						-0.003	0.001
						(0.013)	(0.012)
Extraversion						-0.001	-0.012
						(0.014)	(0.014)
Openness						0.006	0.006
						(0.015)	(0.015)
Agreeableness						-0.014	-0.009
						(0.018)	(0.017)
Conscientiousness						-0.005	-0.004
						(0.015)	(0.015)
CLASS average							0.045***
							(0.015)
School fixed effects	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	YES	YES	YES	YES	YES
Parent controls	NO	NO	YES	YES	YES	YES	YES
Classroom controls	NO	NO	NO	YES	YES	YES	YES
Observations	12632	12632	12632	12632	12632	12632	12632
R^2	0.149	0.433	0.440	0.440	0.441	0.442	0.442

Notes: Each column reports coefficients from OLS regressions estimated with clustered standard errors (in parentheses) at the school level. Columns (2)-(7) control for the following student characteristics: baseline TVIP score, age, gender, attendance to preschool; parent characteristics: years of education and living standard conditions; classroom characteristics: class size, classroom averages of student and parent characteristics. CLASS stands for Classroom Assessment Scoring System.* Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

We report our estimates for math learning in table 3.6, following the same structure of table 3.5. As shown in Column (1), we observe a positive causal effect of *test-screened tenured* teachers of a 0.085 standard deviation higher end-of-year test score in math when no additional controls are included, albeit only at the 10 percent significance level. Once additional teacher observable characteristics are taken into account as controls in Column (5), we find some evidence of a significant effect at the 5 percent level. It finally rises to a 0.099 standard deviation higher end-of-year math test score when we include a full set of teacher characteristics in Column (7).

Table 3.6: Estimates of Effects of Test-Screened Tenured Teachers on Math

Teacher	Math						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Test-screened tenured	0.085*	0.086*	0.080*	0.068	0.093**	0.086*	0.099**
	(0.048)	(0.045)	(0.045)	(0.046)	(0.046)	(0.045)	(0.044)
Female teacher					0.097	0.092	0.028
					(0.236)	(0.241)	(0.233)
Years of experience					0.005***	0.006***	0.004**
					(0.002)	(0.002)	(0.002)
Years of education					-0.004	-0.004	-0.005
					(0.007)	(0.008)	(0.007)
Cognitive skills						0.014	0.011
						(0.015)	(0.016)
Neuroticism						0.006	0.012
						(0.015)	(0.014)
Extraversion						0.022*	0.008
						(0.013)	(0.014)
Openness						0.013	0.013
						(0.018)	(0.018)
Agreeableness						-0.010	-0.003
						(0.018)	(0.017)
Conscientiousness						-0.018	-0.017
						(0.018)	(0.018)
CLASS average							0.059***
							(0.019)
School fixed effects	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	YES	YES	YES	YES	YES
Parent controls	NO	NO	YES	YES	YES	YES	YES
Classroom controls	NO	NO	NO	YES	YES	YES	YES
Observations	12632	12632	12632	12632	12632	12632	12632
R^2	0.123	0.303	0.309	0.309	0.310	0.310	0.311

Notes: Each column reports coefficients from OLS regressions estimated with clustered standard errors (in parentheses) at the school level. Columns (2)-(7) control for the following student characteristics: baseline TVIP score, age, gender, attendance to preschool; parent characteristics: years of education and living standard conditions; classroom characteristics: class size, classroom averages of student and parent characteristics. CLASS stands for Classroom Assessment Scoring System.* Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Aside from the effect of *test-screened tenured* teachers, the only teacher characteristics that seem to correlate with language and math end-of-year test scores are teacher experience and classroom practices (CLASS), which is in line with the conclusions of Araujo et al. (2016). In addition, we find a significant and positive association between teacher cognitive skills and language learning.

It is important to note that the size of the estimated effects are substantial. Our basic estimations of the effect of a *test-screened tenured* teacher range between 10.5 and 12.5 percent of a standard deviation of end-of-year test scores for language, and

from 8.5 to 9.9 percent for math. By contrast, existing estimations of the effects of certified or test-screened teachers in the US typically range between 1 and 7 percent of a standard deviation for reading and math test scores (Clotfelter, Ladd and Vigdor, 2007b; Goldhaber, 2007; Goldhaber and Anthony, 2007; Harris and Sass, 2009; Goldhaber, Gratz and Theobald, 2017).

3.4.2 Estimates of Test-Screened Tenured, Other-Tenured and Test-Screened Contract Teachers

As shown in table 3.2, about 13 percent of the teachers in our sample are *test-screened tenured* teachers who passed national entry exams and won a merit-based competition. Previously, we compared this group with all of their peer teachers who had not undergone the new competitive recruitment process. However, in our comparison group there are teachers tenured before 2007 by local authorities (45 percent of the full sample), contract teachers who had passed national entry exams but had not yet won a competition for tenure (12 percent of the full sample), and contract teachers who had not passed national entry exams (30 percent of the full sample).

Table A 3.1 of the Appendix presents descriptive statistics of these other-tenured teachers, test-screened contract teachers and contract teachers, and differences between them and the sample of *test-screened tenured* teachers. We observe that other-tenured teachers have significantly more years of experience and have slightly higher CLASS average scores than *test-screened tenured* teachers. By contrast, other-tenured teachers exhibit lower cognitive skills and personality trait scores, which are statistically significant for extraversion, openness and agreeableness. Test-screened contract teachers are similar to *test-screened tenured* teachers in their cognitive skills and personality traits, but they have significantly fewer years of experience and lower CLASS scores. Finally, contract teachers have the lowest averages for most qualifications and characteristics among all teachers, including years of experience, cognitive skills and CLASS scores. These values are remarkably lower in size and statistical significance than the averages of *test-screened tenured* teachers.

In addition, it is worth noting that tenured teachers and contract teachers face very different incentives in terms of wages, job security and prospective careers. For instance, while the nominal monthly entry wage for a new tenured teacher rose to

US$775 in 2011, the wage of a fixed-term contract teacher stagnated around US$300. Contract teachers also face high job uncertainty since fixed-term contracts are renewed annually.

In this context, it is important to explore possible differences within the original comparison group. We estimate the effect of *test-screened tenured* teachers, other-tenured teachers and test-screened contract teachers[64] compared with contract teachers on learning outcomes using a specification analogous to equation (3.1):

$$Y_{ics} = \alpha_s + Test\_tenured_{cs}\beta_1 + Other\_tenured_{cs}\beta_2 + Test\_contract_{cs}\beta_3 + X_{ics}\beta_4 + \bar{X}_{ics}\beta_5 + C_{cs}\beta_6 + T_{cs}\beta_7 + P_{cs}\beta_8 + CLASS_{cs}\beta_9 + u_{ics}, \quad (3.2)$$

Table 3.7 presents regression results for language learning. Column (1) shows a positive and significant effect of *test-screened tenured* teachers on child learning in language when no controls are taking into account other than school fixed effects. A kindergarten student taught by a *test-screened tenured* teacher has a 0.169 standard deviation higher end-of-year test score in language compared with a student assigned to a contract teacher. Children assigned to other-tenured teachers also have significantly higher end-of-year language test scores than those assigned to contract teachers. Nonetheless, the size of this effect is about half of the *test-screened tenured* teacher effect. The impact of *test-screened tenured* and other-tenured teachers on student language learning persists when controlling for additional child, family and classroom covariates, as observed in Columns (2) to (4). The other-tenured teacher effect decreases and becomes statistically insignificant when additional teacher characteristics are taken into account in Columns (5) to (7). By contrast, the effect of a *test-screened tenured* teacher remains positive, statistically significant and only slightly decreases to a 0.148 standard deviation. Curiously, the effect of test-screened contract teachers is not statistically different from the effect of contract teachers.

[64] As a robustness check, we run all our estimations excluding test-screened contract teachers from our analysis. The results are confirmatory and available upon request.

Table 3.7: Estimates of Effects of Test-Screened Tenured, Other-Tenured, Test-Screened Contract vs. Contract Teachers on Language

Teacher	Language						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Test-screened tenured	0.169***	0.166***	0.159***	0.155***	0.157***	0.143***	0.148***
	(0.050)	(0.044)	(0.045)	(0.045)	(0.044)	(0.045)	(0.043)
Other-tenured	0.097**	0.093***	0.091**	0.098***	0.078**	0.058*	0.047
	(0.040)	(0.035)	(0.035)	(0.035)	(0.035)	(0.034)	(0.035)
Test-screened contract	-0.012	0.040	0.044	0.032	0.030	-0.013	-0.005
	(0.061)	(0.054)	(0.054)	(0.056)	(0.055)	(0.060)	(0.061)
Female teacher					0.248**	0.218**	0.171*
					(0.103)	(0.102)	(0.097)
Years of experience					0.002	0.002	0.002
					(0.002)	(0.002)	(0.002)
Years of education					-0.000	-0.001	-0.002
					(0.007)	(0.007)	(0.007)
Cognitive skills						0.038***	0.036**
						(0.014)	(0.014)
Neuroticism						-0.005	-0.001
						(0.013)	(0.013)
Extraversion						0.001	-0.010
						(0.015)	(0.015)
Openness						0.003	0.004
						(0.015)	(0.015)
Agreeableness						-0.013	-0.008
						(0.018)	(0.017)
Conscientiousness						-0.004	-0.003
						(0.015)	(0.015)
CLASS average							0.042***
							(0.016)
School fixed effects	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	YES	YES	YES	YES	YES
Parent controls	NO	NO	YES	YES	YES	YES	YES
Classroom controls	NO	NO	NO	YES	YES	YES	YES
Observations	12632	12632	12632	12632	12632	12632	12632
R^2	0.150	0.433	0.441	0.441	0.441	0.442	0.442

Notes: Each column reports coefficients from OLS regressions estimated with clustered standard errors (in parentheses) at the school level. Columns (2)-(7) control for the following student characteristics: baseline TVIP score, age, gender, attendance to preschool; parent characteristics: years of education and living standard conditions; classroom characteristics: class size, classroom averages of student and parent characteristics. CLASS stands for Classroom Assessment Scoring System. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Regression results for math learning are reported in table 3.8. We find a strong, positive and significant effect of *test-screened tenured* teachers on child learning in math. Compared with a contract teacher, the effect of a *test-screened tenured* teacher is a 0.155 standard deviation higher end-of-year math test score when no controls are included aside from school fixed effects, as shown in Column (1). Other-tenured teachers also show a positive significant effect of a 0.105 standard deviation higher end-of-year math test score. These effects hold even when additional student, family and

classroom characteristics are taken into account in Columns (2) to (4). When a full set of teacher covariates is included in Columns (5) to (7), the effect of a *test-screened tenured* teacher is still positive and statistically significant, although its size decreases to a 0.129 standard deviation. By contrast, the effect of other-tenured teachers not only decreases in size to a 0.054 standard deviation, but also becomes statistically insignificant. Once again, the effect of test-screened contract teachers is not statistically different from that of contract teachers.

Table 3.8: Estimates of Effects of Test-Screened Tenured, Other-Tenured, Test-Screened Contract vs. Contract Teachers on Math

Teacher	Math						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Test-screened tenured	0.155***	0.154***	0.147***	0.136**	0.137**	0.123**	0.129**
	(0.057)	(0.052)	(0.052)	(0.052)	(0.053)	(0.052)	(0.051)
Other-tenured	0.105**	0.100***	0.098***	0.100***	0.080**	0.068*	0.054
	(0.040)	(0.036)	(0.036)	(0.036)	(0.038)	(0.038)	(0.038)
Test-screened contract	0.008	0.047	0.049	0.043	0.042	0.020	0.030
	(0.061)	(0.053)	(0.053)	(0.055)	(0.054)	(0.058)	(0.059)
Female teacher					0.101	0.100	0.037
					(0.222)	(0.231)	(0.225)
Years of experience					0.003	0.003*	0.002
					(0.002)	(0.002)	(0.002)
Years of education					-0.005	-0.004	-0.006
					(0.007)	(0.008)	(0.007)
Cognitive skills						0.010	0.007
						(0.015)	(0.016)
Neuroticism						0.004	0.010
						(0.015)	(0.014)
Extraversion						0.022	0.008
						(0.014)	(0.014)
Openness						0.010	0.011
						(0.018)	(0.017)
Agreeableness						-0.008	-0.002
						(0.017)	(0.016)
Conscientiousness						-0.017	-0.016
						(0.018)	(0.018)
CLASS average							0.057***
							(0.019)
School fixed effects	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	YES	YES	YES	YES	YES
Parent controls	NO	NO	YES	YES	YES	YES	YES
Classroom controls	NO	NO	NO	YES	YES	YES	YES
Observations	12632	12632	12632	12632	12632	12632	12632
R^2	0.124	0.303	0.309	0.310	0.310	0.311	0.312

Notes: Each column reports coefficients from OLS regressions estimated with clustered standard errors (in parentheses) at the school level. Columns (2)-(7) control for the following student characteristics: baseline TVIP score, age, gender, attendance to preschool; parent characteristics: years of education and living standard conditions; classroom characteristics: class size, classroom averages of student and parent characteristics. CLASS stands for Classroom Assessment Scoring System. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Our estimation findings confirm the positive and significant effect of teachers who have passed national entry exams and won selection competitions for tenure in language and math learning. Nonetheless, they also show that test-screened teachers who have not won a selection competition do not have the same effect. On the one hand, these results suggest that test score differences among teachers who passed entry examinations might be relevant. On the other hand, they indicate that the entire teacher recruitment process, which combined test scores with the assessment of teacher qualifications and classroom practices, drives the positive effects on student achievement. Finally, the results point to a positive association between permanent job status and performance, in contrast to Araujo et al.'s (2016) findings.[65]

3.4.3 Estimates of Test-Screened Tenured Teachers, Accounting for Ministerial Resolution Competitions

Among the *test-screened tenured* teachers in our sample, around 25 percent won competitions regulated by the original Ministerial Resolution of December 2007 (AM No. 438-07), only about 4 percent won competitions organized under the Ministerial Resolution of January 2010 (AM No. 018-10), and 71 percent won competitions regulated by the Ministerial Resolution of November 2011 (AM No. 379-11).

The impact of each competition might be different. On the one hand, the competition component weighting slightly changed over time. The overall test weighting increased from 45 to 55 percent, while the demonstration class weighting decreased from 20 to 10 percent. On the other hand, the test quality might have differed over time and among competitions. Even though the same type of teacher skill and subject knowledge tests were applied between 2007 and 2013, there is no evidence that they were psychometrically comparable.

Table A 3.2 of the Appendix presents descriptive statistics of *test-screened tenured* teachers organized by the Ministerial Resolution that regulated each

[65] Araujo et al. (2016) did not find a significant association between tenure status and end-of-year student outcomes. Our results differ because we discriminate between *test-screen tenured* teachers and other-tenure teachers. The results of Araujo et al. might be driven by other-tenured teachers whose effects fade out after additional controls are taken into account. In addition, our information on tenure status at the beginning of the school year was confirmed by the Ministry's administrative data, whereas Araujo et al.'s came only from the "Closing Gaps" teacher survey.

competition. Interestingly, *test-screened tenured* teachers who won competitions under the 2007 Regulation have better scores in all the analyzed characteristics compared to their colleagues, who were not screened by the new recruitment policy. *Test-screened tenured* teachers who won competitions under the 2011 Regulations also exhibit higher scores and qualifications, but significantly fewer years of experience. The sample size of teachers who won competitions organized under the 2010 Resolution is too small to obtain robust statistics; however, these teachers are not excluded from our estimations to preserve the integrity of the analysis and sample size.

Accordingly, we estimate the effect of *test-screened tenured* teachers on learning outcomes accounting for the Ministerial Resolution that regulated each selection competition. We use an extended specification of equation (3.1):

$$Y_{ics} = \alpha_s + test\_tenured_{cs}^{AM\ 438-07}\beta_1 + test\_tenured_{cs}^{AM\ 018-10}\beta_2 + test\_tenured_{cs}^{AM\ 379-11}\beta_3 + X_{ics}\beta_4 + \bar{X}_{ics}\beta_5 + C_{cs}\beta_6 + T_{cs}\beta_7 + P_{cs}\beta_8 + CLASS_{cs}\beta_9 + u_{ics},$$

(3.3)

The regression results for language learning are presented in table 3.9. All of the model specifications show that the effect of *test-screened tenured* teachers who won competitions organized under the 2007 Regulation is stronger in size and significance than those of teachers who won competitions organized under other regulations. The size of the effect ranges from a 0.160 standard deviation higher end-of-year language test score in Column (1) with no controls other than school fixed effects to a 0.185 standard deviation in Column (7) where the full set of child, family, classroom and teacher covariates are taken into account. By contrast, the effect of *test-screened tenured* teachers who won competitions organized under the 2011 Regulation ranges from a positive but statistically insignificant 0.090 standard deviation higher end-of-year test score in Column (1) to a significant 0.102 standard deviation in Column (7).

Table 3.9: Estimates of Effects of Test-screened Tenured Teachers on Language, Accounting for Ministerial Resolution Competitions

Teacher	Language						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Test-screened tenured AM 438-07	0.160***	0.193***	0.193***	0.185***	0.193***	0.184***	0.185***
	(0.061)	(0.069)	(0.069)	(0.069)	(0.067)	(0.068)	(0.064)
Test-screened tenured AM 018-10	-0.105	-0.030	-0.055	-0.019	-0.058	-0.040	-0.001
	(0.234)	(0.177)	(0.167)	(0.180)	(0.249)	(0.255)	(0.237)
Test-screened tenured AM 379-11	0.090	0.068	0.060	0.051	0.088*	0.090*	0.102**
	(0.056)	(0.044)	(0.043)	(0.046)	(0.045)	(0.046)	(0.043)
Female teacher					0.263*	0.228*	0.176
					(0.135)	(0.128)	(0.119)
Years of experience					0.005**	0.004**	0.003*
					(0.002)	(0.002)	(0.002)
Years of education					0.001	-0.001	-0.002
					(0.007)	(0.007)	(0.007)
Cognitive skills						0.037***	0.036**
						(0.014)	(0.014)
Neuroticism						-0.003	0.002
						(0.013)	(0.013)
Extraversion						-0.001	-0.012
						(0.014)	(0.014)
Openness						0.008	0.008
						(0.016)	(0.015)
Agreeableness						-0.012	-0.007
						(0.018)	(0.017)
Conscientiousness						-0.005	-0.005
						(0.015)	(0.015)
CLASS average							0.044***
							(0.015)
School fixed effects	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	YES	YES	YES	YES	YES
Parent controls	NO	NO	YES	YES	YES	YES	YES
Classroom controls	NO	NO	NO	YES	YES	YES	YES
Observations	12632	12632	12632	12632	12632	12632	12632
R^2	0.149	0.433	0.440	0.440	0.441	0.442	0.442

Notes: Each column reports coefficients from OLS regressions estimated with clustered standard errors (in parentheses) at the school level. Columns (2)-(7) control for the following student characteristics: baseline TVIP score, age, gender, attendance to preschool; parent characteristics: years of education and living standard conditions; classroom characteristics: class size, classroom averages of student and parent characteristics. CLASS stands for Classroom Assessment Scoring System. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 3.10 presents estimation results for math learning. Once again, we find a stronger positive effect of *test-screened tenured* teachers who won competitions under the 2007 Regulation. Column (1) shows that the children randomly assigned to *test-screened tenured* teachers who won competitions organized under the 2007 Regulation achieve a 0.172 standard deviation higher end-of-year math test score, which is highly

significant. When additional child, family, classroom and teacher covariates are included, this effect increases to a 0.183 standard deviation in Column (7). We also find positive effects of *test-screened tenured* teachers who won competitions organized under the 2011 Regulation, although they are not statistically significant.

Table 3.10: Estimates of Effects of Test-Screened Tenured Teachers on Math, Accounting for Ministerial Resolution Competitions

Teacher	Math						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Test-screened tenured AM 438-07	0.172***	0.198***	0.197***	0.183***	0.192***	0.182**	0.183**
	(0.059)	(0.060)	(0.059)	(0.062)	(0.074)	(0.074)	(0.074)
Test-screened tenured AM 018-10	-0.155	-0.094	-0.118	-0.093	-0.093	-0.114	-0.062
	(0.262)	(0.219)	(0.209)	(0.210)	(0.253)	(0.258)	(0.236)
Test-screened tenured AM 379-11	0.058	0.045	0.037	0.025	0.057	0.051	0.067
	(0.063)	(0.058)	(0.058)	(0.059)	(0.060)	(0.059)	(0.057)
Female teacher					0.115	0.113	0.045
					(0.249)	(0.255)	(0.245)
Years of experience					0.005**	0.005***	0.004**
					(0.002)	(0.002)	(0.002)
Years of education					-0.004	-0.004	-0.005
					(0.007)	(0.007)	(0.007)
Cognitive skills						0.012	0.010
						(0.015)	(0.016)
Neuroticism						0.007	0.013
						(0.015)	(0.014)
Extraversion						0.022	0.008
						(0.013)	(0.014)
Openness						0.015	0.015
						(0.018)	(0.018)
Agreeableness						-0.008	-0.001
						(0.018)	(0.017)
Conscientiousness						-0.019	-0.018
						(0.018)	(0.018)
CLASS average							0.058***
							(0.019)
School fixed effects	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	YES	YES	YES	YES	YES
Parent controls	NO	NO	YES	YES	YES	YES	YES
Classroom controls	NO	NO	NO	YES	YES	YES	YES
Observations	12632	12632	12632	12632	12632	12632	12632
R^2	0.123	0.303	0.309	0.309	0.310	0.310	0.312

Notes: Each column reports coefficients from OLS regressions estimated with clustered standard errors (in parentheses) at the school level. Columns (2)-(7) control for the following student characteristics: baseline TVIP score, age, gender, attendance to preschool; parent characteristics: years of education and living standard conditions; classroom characteristics: class size, classroom averages of student and parent characteristics. CLASS stands for Classroom Assessment Scoring System. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Finally, estimation results of the effects of *test-screened tenured* teachers who won competitions organized under the 2010 Regulation are not statistically different from zero for language and math, although they should be treated with considerable caution due to the small sample size of these teachers.

To sum up, our findings suggest that the teacher selection competitions regulated by the original Ministerial Resolution of December 2007 were more effective in recruiting teachers who have an impact on student learning.

3.5 Heterogeneous Effects

There is an ongoing debate on the extent to which highly-qualified teachers can close learning gaps between socioeconomically advantaged and disadvantaged children (Borman and Kimball, 2005; Boyd *et al.*, 2008; Phillips, 2010; Hanushek *et al.*, 2020; James and Wyckoff, 2020). This is particularly important for Ecuador and other Latin American countries due to the large and persistent differences found in the cognitive development of children of high and low socioeconomic status throughout the school (Schady *et al.*, 2015).

Consequently, we look at heterogeneous effects of *test-screened tenured* teachers among children who started kindergarten with different language skill development levels. First, we use the TVIP baseline score to sort the student sample into quintiles from the lowest to the highest score. Subsequently, we implement our regression model for each TVIP quintile, as formalized in equation (3.1).[66] The results for language learning are presented in table 3.11, and they suggest that the effect of *test-screened tenured* teachers on language is stronger in size and significance for children at the lowest TVIP quintile. While this effect is statistically not different from zero for children at the highest TVIP quintile as displayed in Columns (9) and (10), it ranges between a 0.240 and 0.214 standard deviation higher end-of-year test score for children at the lowest TVIP quintile as shown in Columns (1) and (2).

[66] We also estimated quantile regressions for several quantile values in reading and math. The results are in line with those provided here and are available upon request.

Table 3.11: Estimates of Effects of Test-Screened Tenured Teachers on Language by TVIP Quintiles

Teacher	TVIP Q1		TVIP Q2		TVIP Q3		TVIP Q4		TVIP Q5	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Test-screened tenured	0.240***	0.214***	0.076	0.121*	0.100	0.168**	0.109	0.184**	-0.048	-0.034
	(0.078)	(0.070)	(0.083)	(0.070)	(0.067)	(0.070)	(0.068)	(0.077)	(0.086)	(0.083)
Female		0.068		0.277**		0.144		0.417**		-0.055
		(0.100)		(0.127)		(0.226)		(0.207)		(0.087)
Years of experience		0.005		-0.001		0.004		0.004		0.003
		(0.004)		(0.003)		(0.003)		(0.003)		(0.003)
Years of education		0.001		0.010		0.006		-0.012		-0.004
		(0.015)		(0.013)		(0.012)		(0.011)		(0.011)
Cognitive skills		0.030		0.041*		0.009		0.026		0.065***
		(0.032)		(0.022)		(0.024)		(0.026)		(0.023)
Neuroticism		0.001		-0.004		0.009		0.032		-0.015
		(0.027)		(0.021)		(0.024)		(0.022)		(0.026)
Extraversion		0.014		-0.015		-0.021		-0.051**		0.027
		(0.028)		(0.024)		(0.029)		(0.021)		(0.027)
Openness		-0.020		0.002		0.046*		0.008		0.001
		(0.034)		(0.023)		(0.025)		(0.029)		(0.028)
Agreeableness		-0.043		-0.010		-0.026		0.034		-0.013
		(0.030)		(0.028)		(0.027)		(0.031)		(0.032)
Conscientiousness		0.027		-0.049**		0.012		0.002		0.007
		(0.031)		(0.025)		(0.026)		(0.024)		(0.030)
CLASS average		0.025		0.033		0.026		0.101***		0.067**
		(0.029)		(0.029)		(0.028)		(0.028)		(0.026)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Parent controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Classroom controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Observations	2544	2544	2793	2793	2318	2318	2550	2550	2427	2427
R^2	0.189	0.263	0.186	0.299	0.220	0.295	0.175	0.269	0.190	0.313

Notes: Each column reports coefficients from OLS regressions estimated with clustered standard errors (in parentheses) at the school level. Columns (2), (4), (6), (8) and (10) control for the following student characteristics: baseline TVIP score, age, gender, attendance to preschool; parent characteristics: years of education and living standard conditions; classroom characteristics: class size, classroom averages of student and parent characteristics. CLASS stands for Classroom Assessment Scoring System.* Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 3.12: Estimates of Effects of Test-Screened Tenured Teachers on Math by TVIP Quintiles

Teacher	TVIP Q1		TVIP Q2		TVIP Q3		TVIP Q4		TVIP Q5	
	1	2	3	4	5	6	7	8	9	10
Test-screened tenured	0.065	0.046	0.112	0.127*	0.179**	0.272***	0.164**	0.218***	-0.052	-0.041
	(0.056)	(0.053)	(0.082)	(0.074)	(0.086)	(0.095)	(0.080)	(0.081)	(0.103)	(0.102)
Proportion female		0.020		0.272**		0.068		-0.015		-0.156
		(0.279)		(0.129)		(0.473)		(0.334)		(0.286)
Years of experience		0.007**		0.001		0.010**		0.001		0.002
		(0.003)		(0.004)		(0.004)		(0.003)		(0.004)
Years of education		-0.009		-0.001		0.007		0.001		-0.011
		(0.011)		(0.014)		(0.014)		(0.010)		(0.016)
IQ		0.012		0.011		0.012		0.007		0.010
		(0.026)		(0.023)		(0.032)		(0.027)		(0.033)
Neuroticism		-0.017		0.027		0.056*		0.040		-0.007
		(0.025)		(0.031)		(0.029)		(0.026)		(0.031)
Extraversion		0.014		0.020		0.037		-0.053*		0.033
		(0.025)		(0.025)		(0.027)		(0.027)		(0.035)
Openness		-0.030		0.013		0.041		0.035		0.023
		(0.034)		(0.029)		(0.036)		(0.028)		(0.038)
Agreeableness		-0.027		-0.002		-0.031		0.031		0.007
		(0.029)		(0.029)		(0.035)		(0.033)		(0.038)
Conscientiousness		-0.004		-0.035		-0.014		-0.035		-0.006
		(0.028)		(0.025)		(0.034)		(0.030)		(0.038)
CLASS average		0.021		0.041		0.078**		0.128***		0.059
		(0.030)		(0.033)		(0.039)		(0.034)		(0.037)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Parent controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Classroom controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Observations	2544	2544	2793	2793	2318	2318	2550	2550	2427	2427
R^2	0.189	0.227	0.158	0.223	0.183	0.255	0.179	0.257	0.172	0.268

Notes: Each column reports coefficients from OLS regressions estimated with clustered standard errors (in parentheses) at the school level. Columns (2), (4), (6), (8) and (10) control for the following student characteristics: baseline TVIP score, age, gender, attendance to preschool; parent characteristics: years of education and living standard conditions; classroom characteristics: class size, classroom averages of student and parent characteristics. CLASS stands for Classroom Assessment Scoring System. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 3.12 presents regression results for math learning. In contrast to the previous finding, the effect of *test-screened tenured* teachers on math is stronger for children at the middle-upper TVIP baseline quintiles, as shown in Columns (5) to (8).

We also examine the effects of *test-screened tenured* teachers among children from different socioeconomic backgrounds. We use our living standard indicator (LSI) standardized score to sort the student sample into quartiles starting with the lowest score. Subsequently, we estimate our regression model for each LSI quartile. Table 3.13 presents estimation results for language. Our results suggest that the effects of *test-screened tenured* teachers on language are stronger in size and significance for children in the lowest and highest LSI quartile. The effect for children in the lowest LSI quartile ranges from a 0.142 to a 0.200 standard deviation higher end-of-year test score, as presented in Columns (1) to (2). By contrast, we find no heterogeneous effect on math learning as shown in table 3.14.

Overall, our analysis of heterogeneous effects suggest that the impact of *test-screened tenured* teachers on language learning is stronger for vulnerable children who started the school year with lower TVIP scores or came from socioeconomically disadvantaged households. This is not the case for math learning. These results are in good agreement with the findings of Araujo P. (2019).

Table 3.13: Estimates of Effects of Test-Screened Tenured Teachers on Language by Household Living Standard Indicator Quintiles

Teacher	Language							
	LSI Q1		LSI Q2		LSI Q3		LSI Q4	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test-screened tenured	0.142*	0.200***	0.038	0.060	0.049	0.086	0.155*	0.152**
	(0.077)	(0.066)	(0.097)	(0.079)	(0.073)	(0.064)	(0.084)	(0.075)
Female teacher		0.202		0.196*		0.063		0.413**
		(0.234)		(0.104)		(0.177)		(0.187)
Years of experience		0.009***		0.000		0.004		0.000
		(0.003)		(0.003)		(0.003)		(0.003)
Years of education		0.007		0.010		-0.010		-0.017
		(0.013)		(0.011)		(0.009)		(0.016)
Cognitive skills		0.034		0.070***		0.030		-0.000
		(0.024)		(0.023)		(0.022)		(0.027)
Neuroticism		0.001		-0.013		0.005		0.032
		(0.017)		(0.021)		(0.023)		(0.026)
Extraversion		-0.010		0.013		-0.012		-0.021
		(0.021)		(0.021)		(0.025)		(0.029)
Openness		-0.017		-0.017		0.025		0.044*
		(0.025)		(0.025)		(0.028)		(0.026)
Agreeableness		-0.038		-0.013		-0.014		0.039
		(0.023)		(0.028)		(0.030)		(0.035)
Conscientiousness		0.031		-0.002		-0.020		-0.036
		(0.026)		(0.024)		(0.023)		(0.029)
CLASS average		0.066***		0.041**		0.011		0.062**
		(0.023)		(0.021)		(0.027)		(0.030)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	NO	YES	NO	YES	NO	YES
Parent controls	NO	YES	NO	YES	NO	YES	NO	YES
Classroom controls	NO	YES	NO	YES	NO	YES	NO	YES
Observations	3537	3537	3268	3268	3403	3403	2424	2424
R^2	0.152	0.450	0.162	0.469	0.193	0.453	0.191	0.461

Notes: Each column reports coefficients from OLS regressions estimated with clustered standard errors (in parentheses) at the school level. Columns (2), (4), (6) and (8) control for the following student characteristics: baseline TVIP score, age, gender, attendance to preschool; parent characteristics: years of education and living standard conditions; classroom characteristics: class size, classroom averages of student and parent characteristics. CLASS stands for Classroom Assessment Scoring System. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 3.14: Estimates of Effects of Test-Screened Tenured Teachers on Math by Household Living Standard Indicator Quintiles

Teacher	Math							
	LSI Q1		LSI Q2		LSI Q3		LSI Q4	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test-screened tenured	0.056	0.081	0.065	0.077	0.060	0.072	0.117	0.135
	(0.086)	(0.077)	(0.074)	(0.071)	(0.065)	(0.056)	(0.127)	(0.097)
Female teacher		-0.157		0.210		0.162		0.057
		(0.259)		(0.317)		(0.310)		(0.277)
Years of experience		0.007**		0.001		0.002		0.003
		(0.003)		(0.004)		(0.004)		(0.004)
Years of education		-0.005		0.010		-0.004		-0.020
		(0.011)		(0.012)		(0.010)		(0.018)
Cognitive skills		0.006		0.048*		-0.013		0.005
		(0.021)		(0.026)		(0.026)		(0.036)
Neuroticism		-0.008		0.015		0.046**		-0.009
		(0.023)		(0.025)		(0.023)		(0.035)
Extraversion		-0.017		0.015		0.028		0.025
		(0.023)		(0.021)		(0.022)		(0.034)
Openness		0.009		0.008		0.029		0.010
		(0.024)		(0.027)		(0.032)		(0.032)
Agreeableness		-0.005		-0.012		0.009		0.018
		(0.022)		(0.033)		(0.030)		(0.042)
Conscientiousness		-0.005		-0.008		-0.038		-0.034
		(0.028)		(0.027)		(0.024)		(0.038)
CLASS average		0.066***		0.059**		0.038		0.075*
		(0.025)		(0.023)		(0.033)		(0.044)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	NO	YES	NO	YES	NO	YES
Parent controls	NO	YES	NO	YES	NO	YES	NO	YES
Classroom controls	NO	YES	NO	YES	NO	YES	NO	YES
Observations	3537	3537	3268	3268	3403	3403	2424	2424
R^2	0.163	0.349	0.144	0.346	0.143	0.317	0.169	0.345

Notes: Each column reports coefficients from OLS regressions estimated with clustered standard errors (in parentheses) at the school level. Columns (2), (4), (6) and (8) control for the following student characteristics: TVIP score, age, gender, attendance to preschool; parent characteristics: years of education and living standard conditions; classroom characteristics: class size, classroom averages of student and parent characteristics. CLASS stands for Classroom Assessment Scoring System. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

3.6 Robustness Check

In this section, we show that our results are robust to using the subsample of schools that have at least one treated kindergarten classroom: in other words, a classroom taught by a *test-screened tenured* teacher.

Our original school sample comprises 204 schools and all of their kindergarten classrooms. Among them, 161 are purely control schools that have no kindergarten classroom taught by a *test-screened tenured* teacher, while another six schools are purely treated schools where all kindergarten classrooms are taught by *test-screened tenured* teachers. Finally, 37 schools in our sample have a combination of treated and control kindergarten classrooms. This is the subsample that we are interested in for our robustness check, although first we exclude a unique school whose control group is a classroom taught by a test-screened contract teacher, which guarantees that all of our control classrooms are taught by teachers who have not passed any stage of the new teacher recruitment process. Our final subsample comprises 36 schools and 84 teachers.

We repeat our original analysis with the subsample of 36 schools.[67] First, we evaluate the randomization of students into classrooms by regressing teacher *test-screened tenured* status on student and family predetermined covariates, conditioned on school fixed effects.

Our results are presented in table 3.15 and show that none of the student or family characteristics predict the likelihood that a child is assigned to a *test-screened tenured* teacher at the 5 percent significance level. Once again, parents' years of schooling is marginally correlated with assignment to a *test-screened tenured* teacher, albeit only at the 10 percent significance level and with a very small coefficient. Moreover, the F-test for the joint significance of all of the predetermined demographic variables is statistically insignificant (p=0.287). We conclude that the random assignment to *test-screened tenured* teachers was also successful for the subsample of 36 schools.

[67] We re-run all our econometric analyses for the subsample of 36 schools, but we present here the main estimations. All results were confirmatory and are available upon request.

Table 3.15: Randomization Test, School Subsample (Robustness Check)

	Test-screened tenured teachers
Children:	
Age (months)	-0.002
	(0.002)
Proportion female	0.004
	(0.018)
TVIP	0.000
	(0.001)
Proportion who attended preschool	-0.031
	(0.026)
Family:	
Parents' years of schooling	0.006*
	(0.003)
Living standard indicator	-0.003
	(0.015)
Observations	2393
R^2	0.057
F	1.15
p	0.353

Note: Subsample of 36 schools with treated and control classrooms. OLS model estimated with clustered standard errors (in parentheses) at the school level and school fixed effects. . TVIP stands for Test de Vocabulario en Imágenes Peabody, the Spanish version of the Peabody Picture Vocabulary Test (PPVT). * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Subsequently, we evaluate whether being a no-show, attritor or late enrollment is correlated with assignment to the treatment in our subsample of schools. The results are presented in table 3.16, again showing that there is no evidence that being randomly assigned to a *test-screened tenured* teacher has an effect on the decision to be a no-show, attritor or late enrollment.

Finally, we estimate our regression model as described in equation (3.1) for language and math end-of-year test scores and present them in table 3.17. The student sample size substantially decreases from 12,632 to 2,393 kindergarten children. Nonetheless, we still find that students randomly assigned to *test-screened tenured* teachers achieve at least a 0.102 standard deviation significantly higher end-of-year language test score, as shown in Columns (1) to (4). In math, the effect of *test-screened tenured* teachers is at least a 0.089 standard deviation higher end-of-year test score as shown in Columns (5) to (8), although its significance is only found at the 10 percent

level. In order to have a conservative estimation of our standard errors, all regression are clustered at the school level. However, when standard errors are clustered at the classroom level where treatment occurred, the significance level of the *test-screened tenured* teacher effect increases to 1 percent for language and math in all of our estimations.[68]

Table 3.16: No-Show, Attrition and Late Enrollment Tests (Robustness Check)

Teacher	No-shows		Attritors		Late enrollments	
	(1)	(2)	(3)	(4)	(5)	(6)
Test-screened tenured	-0.006	-0.005	-0.005	-0.005	-0.006	0.001
	(0.007)	(0.008)	(0.007)	(0.007)	(0.013)	(0.016)
Female		-0.077***		0.009		-0.105***
		(0.019)		(0.015)		(0.026)
Years of experience		-0.000		0.000		0.001
		(0.001)		(0.001)		(0.001)
Years of education		-0.002		0.002		-0.002
		(0.003)		(0.002)		(0.003)
Cognitive skills		0.007		0.005		-0.006
		(0.005)		(0.006)		(0.005)
Neuroticism		0.001		0.001		-0.003
		(0.004)		(0.003)		(0.005)
Extraversion		-0.000		0.006		-0.011
		(0.005)		(0.005)		(0.007)
Openness		0.006		-0.002		-0.007
		(0.005)		(0.005)		(0.009)
Agreeableness		-0.015***		-0.000		-0.006
		(0.004)		(0.005)		(0.013)
Conscientiousness		0.004		0.002		0.016*
		(0.006)		(0.005)		(0.009)
CLASS average		0.002		0.002		-0.006
		(0.004)		(0.005)		(0.009)
Observations	2688	2654	2589	2555	2681	2645
R^2	0.024	0.027	0.044	0.046	0.037	0.043
F	0.65	47.29	0.45	2.64	0.19	112.27
p	0.424	0.000	0.506	0.016	0.667	0.000

Note: Subsample of 36 schools with treated and control classrooms. OLS linear probability models estimated with clustered standard errors (in parentheses) at the school level and school fixed effects. No-shows is a dummy variable that takes the value of one if an enrolled student did not show up in the beginning of the school year. Attritors is a dummy variable that takes the value of one if the student dropped out of the school. Late enrollment is a dummy variable that takes the value of one if the student enrolled after the school year started. CLASS stands for Classroom Assessment Scoring System.* Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

[68] The results of our estimations for the subsample of 36 schools with standard errors clustered at the classroom level were confirmatory and are available upon request.

Table 3.17: Estimates of Effects of Test-Tenured Teachers, School Subsample (Robustness Check)

Teacher	Language				Math			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test-screened tenured	0.102**	0.137***	0.093***	0.102***	0.089*	0.107**	0.094**	0.105*
	(0.045)	(0.042)	(0.029)	(0.034)	(0.050)	(0.052)	(0.044)	(0.053)
Female		0.161***	0.357***	0.319***		0.104	0.362***	0.316**
		(0.055)	(0.086)	(0.102)		(0.089)	(0.127)	(0.156)
Years of experience		0.005**	0.008***	0.007***		-0.002	0.004	0.003
		(0.002)	(0.003)	(0.003)		(0.004)	(0.004)	(0.004)
Years of education		-0.005	0.002	0.002		-0.023*	-0.003	-0.003
		(0.011)	(0.011)	(0.011)		(0.013)	(0.013)	(0.012)
Cognitive skills			-0.010	-0.005			-0.075**	-0.069*
			(0.025)	(0.029)			(0.035)	(0.037)
Neuroticism			-0.012	-0.011			0.042	0.044
			(0.022)	(0.021)			(0.033)	(0.031)
Extraversion			0.072***	0.061**			0.074**	0.061
			(0.017)	(0.027)			(0.030)	(0.038)
Openness			0.037**	0.030			0.021	0.012
			(0.017)	(0.020)			(0.035)	(0.040)
Agreeableness			0.060**	0.054*			0.068*	0.061
			(0.028)	(0.029)			(0.036)	(0.039)
Conscientiousness			-0.072**	-0.062*			-0.087*	-0.075
			(0.027)	(0.034)			(0.044)	(0.051)
CLASS average				0.025				0.030
				(0.033)				(0.043)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	YES	YES	NO	YES	YES	YES
Parent controls	NO	YES	YES	YES	NO	YES	YES	YES
Classroom controls	NO	YES	YES	YES	NO	YES	YES	YES
Observations	2393	2393	2393	2393	2393	2393	2393	2393
R^2	0.107	0.426	0.430	0.431	0.045	0.262	0.267	0.267

Notes: Subsample of 36 schools with treated and control classrooms. Each column reports coefficients from OLS regressions estimated with clustered standard errors (in parentheses) at the school level. Columns (2)-(4) and (6)-(8) control for the following student characteristics: TVIP score, age, gender, attendance to preschool; parent characteristics: years of education and living standard conditions; classroom characteristics: class size, classroom averages of student and parent characteristics. CLASS stands for Classroom Assessment Scoring System.* Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

These results are practically identical to those found in our original analysis and have further strengthened our confidence in the significant and unbiased effect of Ecuadorian *test-screened tenured* teachers on kindergarten learning.

3.7 Conclusions

In this paper, we have assessed the effectiveness of Ecuador's new teacher recruitment policy, which from 2007 onwards required teacher candidates to pass mandatory skill and content knowledge tests before they were allowed to participate in merit-based selection competitions for tenure at public schools. For our identification strategy, we combine administrative teacher recruitment information from Ecuador's Ministry of Education with data provided by the "Closing Gaps" project, which randomly assigned a representative sample of kindergarten children to their classrooms and teachers in the 2012-2013 school year.

Our work has led us to conclude that teachers who passed national entry tests and won a competition for tenure in Ecuador have positive and significant effects on language learning in kindergarten, which persist even after controlling for teacher cognitive skill, personality and classroom practice. Children randomly assigned to a *test-screened tenured* teacher achieved between a 0.105 and a 0.125 standard deviation higher end-of-year language test score in the 2012-2013 school year. Moreover, *test-screened tenured* teachers have a robust effect of at least a 0.137 standard deviation compared with contract teachers. *Test-screened tenured* teachers also outperform teachers tenured before 2007 who were not required to pass national entry tests.

The evidence obtained in our study also suggests positive effects of *test-screened tenured* teachers on math learning in kindergarten. Children randomly assigned to these teachers achieved between a 0.085 and a 0.099 standard deviation higher end-of-year math test score in the 2012-2013 school year. When compared with a contract teacher, the effect of a *test-screened tenured* teacher is at least a 0.133 standard deviation, which is robust and highly significant. Likewise, the effects of *test-screened tenured* teachers are larger in size and significance than those of teachers tenured by previous processes.

Surprisingly, we do not find similar effects for teachers who have passed entry tests but have not won a completion for tenure, and therefore work with fixed-term contracts. These results suggest, on the one hand, that differences in skill and subject knowledge test scores among *test-screened* teachers might be a relevant indicator of

teacher quality. On the other hand, they indicate that the entire teacher recruitment process, which combines teacher test scores, background characteristics and classroom practice assessments, is able to identify more effective teachers. In addition, these results point to a positive association between job status and performance in Ecuador.

We also explore potential differences among teacher selection processes organized under three different regulations between 2007 and 2012. Our results suggest that competitions regulated by the original Ministerial Resolution of December 2007 recruited more effective teachers. There are two potential explanations behind this result. On the one hand, this is the regulation that assigns the highest weighting to the teacher demonstration class component. Our results, the findings of Araujo et al. (2016) as well as increasing international evidence (Kane *et al.*, 2011, 2013; Bacher-Hicks *et al.*, 2019) suggest that teacher classroom practices are good indicators of teacher effectiveness. Accordingly, it is possible that the results are driven by the weighting assigned to the classroom practice evaluation. On the other hand, it is possible that the results are driven by quality differences in skill and content knowledge tests among the processes organized under different regulations.

Remarkably, our study also confirms the potential effectiveness of *test-screened tenured* teachers in closing learning gaps between socioeconomically advantaged and disadvantaged children in Ecuador. We find that the effects of *test-screened tenured* teachers on language learning are stronger for children who started the school year with the lowest baseline TVIP scores or those who came from socioeconomically disadvantaged households. This is not the case for math learning.

In addition, as part of our robustness check, we conduct our entire analysis for the subsample of schools that have at least one classroom randomly assigned to a *test-screened tenured* teacher. Even though our student sample decreased to one-fifth of the original size, our estimation results are practically the same in size and significance as those of the full sample. These results emphasize the validity of our original estimations and further strengthen our confidence in our findings of the unbiased significant and positive causal effect of Ecuadorian *test-screened tenured* teachers on kindergarten learning.

We conclude by drawing some policy implications from our analysis. Our results show that between 2007 and 2012 the Ecuadorian teacher reform succeeded in recruiting more effective teachers. *Test-screened tenured* teachers were significantly more successful than their peers in raising kindergarten student learning in the 2012-2013 school year. It is likely that the mechanism behind the observed results is a combination of screening teacher candidates who demonstrate higher cognitive skills, greater content knowledge of the subject taught and better class practices, along with the provision of an economically attractive permanent job position. Under this scenario, it is desirable for policy-makers in other Latin American countries to introduce transparent, objective and highly competitive teacher recruitment processes to screen and select the best available candidates to improve teacher quality.

3.8 Appendix

Table A 3.1: Summary Statistics for Test-Screened Tenured, Other-Tenured and Test-Screened Contract Teachers

	Sample Mean (SD)					Difference to Test-screened tenured teachers (p-value)		
	Full Sample	Test-screened Tenured	Other-tenured	Test-screened contract	Contract	Other-tenured	Test-screened contract	Contract
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Proportion female	0.99	0.96	0.99	1.00	0.99	0.03	0.04	0.03
	(0.10)	(0.19)	(0.07)	(0.00)	(0.09)	(0.23)	(0.16)	(0.28)
Proportion Tenure	0.58	1.00	1.00	0.00	0.00	0.00	-1.00	-1.00
	(0.49)	(0.00)	(0.00)	(0.00)	(0.00)	(.)	(.)	(.)
Years of experience	14.74	12.84	20.19	9.48	9.34	7.35***	-3.36***	-3.51***
	(8.60)	(7.08)	(8.19)	(4.62)	(5.41)	(0.00)	(0.00)	(0.00)
Years of education	17.15	17.46	17.36	16.94	16.79	-0.11	-0.52	-0.67**
	(1.92)	(1.90)	(2.13)	(2.10)	(1.41)	(0.73)	(0.19)	(0.02)
University Degree	0.99	1.00	0.98	0.98	0.99	-0.02*	-0.02	-0.01
	(0.11)	(0.00)	(0.12)	(0.14)	(0.09)	(0.08)	(0.32)	(0.32)
Cognitive skills	86.46	89.93	87.73	88.86	82.18	-2.20	-1.07	-7.74***
	(9.53)	(9.73)	(9.32)	(8.78)	(8.61)	(0.14)	(0.56)	(0.00)
Neuroticism	43.85	43.22	43.16	44.89	44.76	-0.06	1.67	1.54
	(6.72)	(6.80)	(6.70)	(7.88)	(6.14)	(0.96)	(0.25)	(0.15)
Extraversion	45.65	48.55	44.42	47.12	45.74	-4.12***	-1.43	-2.80***
	(6.83)	(6.46)	(6.91)	(6.49)	(6.58)	(0.00)	(0.26)	(0.01)
Openness	50.82	52.97	51.00	50.91	49.60	-1.96*	-2.06	-3.37***
	(6.75)	(6.52)	(6.41)	(6.90)	(7.09)	(0.05)	(0.12)	(0.00)
Agreeableness	48.22	50.63	48.25	49.07	46.84	-2.38**	-1.57	-3.79***
	(7.56)	(7.02)	(7.60)	(6.10)	(7.99)	(0.03)	(0.23)	(0.00)
Conscientiousness	57.55	58.64	57.26	58.36	57.22	-1.38	-0.28	-1.42
	(8.15)	(6.99)	(8.31)	(8.08)	(8.41)	(0.22)	(0.85)	(0.24)
CLASS 2011-2012	3.62	3.80	3.71	3.61	3.42	-0.09	-0.19*	-0.38***
	(0.37)	(0.35)	(0.35)	(0.31)	(0.36)	(0.27)	(0.07)	(0.00)
CLASS 2012-2013	3.41	3.45	3.47	3.36	3.32	0.02	-0.09*	-0.13***
	(0.28)	(0.25)	(0.29)	(0.25)	(0.27)	(0.55)	(0.06)	(0.00)
CLASS Average	3.48	3.50	3.58	3.41	3.36	0.08*	-0.09*	-0.14***
	(0.28)	(0.27)	(0.27)	(0.26)	(0.26)	(0.06)	(0.08)	(0.00)
N	430	54	196	50	130			

Note: This table reports means and standard deviations (in parenthesis) of characteristics of test-screened tenured, other-tenured and test-screened contract teachers. Columns (6)-(8) display the difference between each group and test-screened tenured teachers, and the respective p-value (in parenthesis) from a t-test for equality. Cognitive skills were measured with the Spanish version of the Wechsler Adult Intelligence Scale (WAIS-III). The test is internationally normed so that 100 is the median score for the adult population. The Big Five personality trait scores (openness, conscientiousness, extraversion, agreeableness and neuroticism) were obtained with the NEO PI-R psychometric instrument. Each personality trait can be scored as very low (20-35), low (35-45), average (45-55), high (55-65) or very high (65-80). CLASS stands for Classroom Assessment Scoring System. CLASS domains can be scored as low (scores 1-2), medium (3-5) or high (6-7). * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table A 3.2: Summary Statistics for Test-Screened Tenured Teachers by Ministerial Resolution Competition

	Sample Mean (SD)					Difference to Non Test-screened tenured teachers (p-value)		
	Full Sample	Non Test-screened Tenured	Test-screen tenured AM 438-07	Test-screen tenured AM 018-10	Test-screen tenured AM 379-11	Test-screen tenured AM 438-07	Test-screen tenured AM 018-10	Test-screen tenured AM 379-11
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Proportion female	0.99	0.99	1.00	1.00	0.95	0.01	0.01	-0.05
	(0.10)	(0.07)	(0.00)	(0.00)	(0.23)	(0.16)	(0.16)	(0.21)
Years of experience	14.74	15.01	16.63	16.29	11.27	1.62	1.28	-3.75***
	(8.60)	(8.78)	(9.16)	(5.24)	(5.73)	(0.53)	(0.79)	(0.00)
Tenure	0.58	0.52	1.00	1.00	1.00	0.48***	0.48***	0.48***
	(0.49)	(0.50)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Years of education	17.15	17.11	18.43	17.50	17.11	1.32*	0.39	-0.00
	(1.92)	(1.92)	(2.71)	(2.12)	(1.41)	(0.09)	(0.84)	(1.00)
University Degree	0.99	0.99	1.00	1.00	1.00	0.01**	0.01**	0.01**
	(0.11)	(0.11)	(0.00)	(0.00)	(0.00)	(0.03)	(0.03)	(0.03)
Cognitive Skills (IQ)	86.46	85.96	90.57	82.50	90.08	4.61*	-3.46	4.12**
	(9.53)	(9.40)	(8.38)	(14.85)	(10.09)	(0.06)	(0.80)	(0.02)
Neuroticism	43.85	43.94	41.23	45.99	43.80	-2.71	2.04	-0.14
	(6.72)	(6.72)	(8.31)	(13.39)	(5.91)	(0.25)	(0.86)	(0.89)
Extraversion	45.65	45.24	47.97	46.92	48.85	2.73*	1.68	3.61***
	(6.83)	(6.79)	(5.44)	(1.95)	(6.99)	(0.09)	(0.43)	(0.00)
Openness	50.82	50.51	52.02	60.14	52.94	1.52	9.63	2.43**
	(6.75)	(6.73)	(5.73)	(14.30)	(6.38)	(0.35)	(0.52)	(0.03)
Agreeableness	48.22	47.87	50.18	49.06	50.88	2.31	1.18	3.01**
	(7.56)	(7.58)	(6.47)	(11.26)	(7.22)	(0.21)	(0.91)	(0.02)
Conscientiousness	57.55	57.39	59.64	52.83	58.58	2.24	-4.56	1.19
	(8.15)	(8.30)	(8.15)	(4.36)	(6.63)	(0.33)	(0.37)	(0.31)
CLASS 2011-2012	3.62	3.61	3.84		3.76	0.22*		0.15
	(0.37)	(0.37)	(0.39)		(0.31)	(0.09)		(0.17)
CLASS 2012-2013	3.41	3.41	3.47	3.53	3.44	0.06	0.12	0.03
	(0.28)	(0.29)	(0.29)	(0.18)	(0.24)	(0.46)	(0.51)	(0.42)
CLASS Average	3.48	3.48	3.58	3.53	3.47	0.10	0.04	-0.01
	(0.28)	(0.28)	(0.32)	(0.18)	(0.26)	(0.27)	(0.79)	(0.81)
N	430	376	14	2	38			

Note: This table reports means and standard deviations (in parenthesis) of characteristics of Test-Screened Tenured Teachers by Ministerial Resolution Competition. Columns (6)-(8) display the difference between each group and non-test-screened tenured teachers (Column 2), and the respective p-value (in parenthesis) from a t-test for equality. Cognitive skills were measured with the Spanish version of the Wechsler Adult Intelligence Scale (WAIS-III). The test is internationally normed so that 100 is the median score for the adult population. The Big Five personality trait scores (openness, conscientiousness, extraversion, agreeableness and neuroticism) were obtained with the NEO PI-R psychometric instrument. Each personality trait can be scored as very low (20-35), low (35-45), average (45-55), high (55-65) or very high (65-80). CLASS stands for Classroom Assessment Scoring System. CLASS domains can be scored as low (scores 1-2), medium (3-5) or high (6-7). * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Chapter 4 *Parents Can Tell! Evidence on Classroom Quality Differences in German Primary Schools*[69]

María Daniela Araujo P. and Johanna Sophie Quis

4.1 Introduction

In the past 20 years, a growing body of economic literature on teacher and classroom effects in the United States (US) has shown that high value-added teachers not only substantially contribute to student learning, but also positively influence later-life outcomes such as college attendance and earnings (Hanushek and Rivkin, 2012; Jackson, Rockoff and Staiger, 2014; Koedel, Mihaly and Rockoff, 2015; Strøm and Falch, 2020). It has also been found that easily quantifiable teacher characteristics are weakly, or not at all associated with individual teacher effects on student performance. This has led to the use of value-added measurements in processes of teacher recruitment, evaluation and dismissal (Koedel, Mihaly and Rockoff, 2015; Steinberg and Donaldson, 2016). Nonetheless, in the same period, there has been very little research on teacher effectiveness and its educational or economic impact in Germany.

We address this gap and examine to what extent individual teachers impact the mathematical and language competence development of their students in the first years of primary school in Germany. For this purpose, we first build a short teacher panel with grade 1 and 2 data from the Starting Cohort 2 (SC2) of the German National Educational Panel Study (NEPS). Then, for our estimation strategy, we show that there is no evidence for matching of students to teachers based on ability in German primary schools. Subsequently, we estimate a value-added to student competence development

[69] A version of this article has been submitted to *Labour Economics* (ISSN: 0927-5371), but not yet accepted.

model using classroom fixed-, as well as random effects, which are mainly driven by teacher quality differences across classrooms. Both model specifications apply empirical Bayes shrinkage to adjust the classroom effects by their level of precision. Our results show substantial individual classroom effects on math and language competence development in the first grades of primary school. One standard deviation increase in classroom quality is associated with at least a 0.12 standard deviation increase in student mathematical competence, and at least a 0.14 standard deviation increase in language competence, over a semester of instruction.

In addition, we examine the association between teacher characteristics and the estimated classroom effects. We find that almost none of the teacher characteristics analyzed, including gender, years of teaching experience, migration background, self-reported *Abitur* GPA, self-reported First State Examination grade, whether the teacher has passed the Second State Examination, teacher's constructivist beliefs, or exhaustion levels, are significantly associated with classroom quality, as measured by the individual classroom contribution to competence development. Remarkably, parental assessment of teacher quality is the only indicator that significantly explains the classroom effects on language competence.

This paper contributes to the literature in three ways. First, we present the first empirical estimations of classroom effects on mathematical and language competence development in primary school in Germany. Second, our results show that these classroom effects do not correlate with characteristics typically used in teacher recruitment and tenure processes in Germany, thus echoing previous findings in the US (Hanushek and Rivkin, 2012; Jackson, Rockoff and Staiger, 2014; Koedel, Mihaly and Rockoff, 2015). Nonetheless, we find that, for language competence development, parents seem to be able to identify more effective teachers and their classrooms, adding to the new and growing evidence on the association between parental and student evaluation and teacher quality (Araujo *et al.*, 2016; Bacher-Hicks *et al.*, 2019). Third, our estimations add to the evidence showing the robustness of teacher and classroom value-added estimates to different settings (Koedel, Mihaly and Rockoff, 2015; Strøm and Falch, 2020).

The remainder of the paper proceeds as follows. In Section 4.2, we provide a background on teacher and classroom effects' research and the German Educational System. Section 4.3 discusses the data. Section 4.4 presents our value-added model and estimation strategy. In Section 4.5, we present our results. Section 4.6 concludes.

4.2 Background and Evidence

4.2.1 Teacher and Classroom Effects

In economics, the study of teacher effects, also referred to as teacher value-added, evaluates the overall contribution of individual teachers to students' human capital accumulation in a specific time period (Hanushek and Rivkin, 2012; Jackson, Rockoff and Staiger, 2014; Koedel, Mihaly and Rockoff, 2015). The teacher value-added research naturally evolved from the education production function literature, where, among other factors, teachers and their characteristics are treated as inputs influencing students' achievement, measured generally through test scores. The value-added model specification differs from the regular education production function in the inclusion of a lagged or baseline achievement measure, which is taken to be a sufficient statistic for unobserved input histories, as well as the unobserved endowment of mental capacity (Todd and Wolpin, 2003). The value-added specification of the education production function estimates individual teacher effects via either fixed or random effects.

Most of the value-added literature stems from the US. Researchers have consistently found substantial individual teacher contribution to student achievement, and significant variation within this contribution (Rockoff, 2004; Nye, Konstantopoulos and Hedges, 2004; Rivkin, Hanushek and Kain, 2005; Aaronson, Barrow and Sander, 2007; Kane and Staiger, 2008; Kane, Rockoff and Staiger, 2008; Hanushek and Rivkin, 2010, 2012; Chetty, Friedman and Rockoff, 2014a, 2014b; Jackson, Rockoff and Staiger, 2014; Koedel, Mihaly and Rockoff, 2015). Estimations of the distribution of teacher effectiveness or value-added in the UShave generated an average standard deviation of 0.17 for math, and of 0.13 for reading, expressed in units

of normalized student achievement (Hanushek and Rivkin, 2010). [70] These estimates are relatively large compared to other interventions in educational production, and consequently, have provided evidence that teacher quality is an important determinant of short-term academic success (Koedel, Mihaly and Rockoff, 2015). Moreover, it has been shown that high value-added teachers positively affect later-life outcomes including college attendance, income[71], and teenage pregnancy (Chetty, Friedman and Rockoff, 2014b).

While a distribution in teacher effectiveness emerges from the value-added studies, the mechanisms by which good teachers outperform poor teachers are less clear. Most studies have shown that easily quantifiable teacher characteristics are consistently either weakly or not at all associated with teacher value-added (Hanushek and Rivkin, 2012; Jackson, Rockoff and Staiger, 2014; Strøm and Falch, 2020). In this context, the use of value-added estimations to evaluate teachers and improve teacher workforce quality is appealing, and hence is growing (Hanushek, 2011; Koedel, Mihaly and Rockoff, 2015; Steinberg and Donaldson, 2016). By 2014, about 80 percent of states implementing new teacher evaluation systems in the U.S. had incorporated one or more measures of teacher performance based on student test scores, and around 30 percent had implemented teacher value-added estimates (Steinberg and Donaldson, 2016).

Critics of value-added modeling have argued that resulting teacher effects' estimates may be biased due to non-random assignment of students to teachers (Rothstein, 2009, 2010; Paufler and Amrein-Beardsley, 2014; Guarino, Reckase and Wooldridge, 2015). Nonetheless, studies that compare teacher value-added estimates obtained in quasi-experimental or experimental[72] settings with those of non-experimental settings, have consistently found that teacher value-added measures are

[70] Most estimates rely on within-school variations (Hanushek and Rivkin, 2010) and have focused on elementary and middle school grades because of the availability of standardized testing data (Jackson, Rockoff and Staiger, 2014).

[71] Chetty, Friedman, and Rockoff (2014b) found that replacing a teacher whose value-added is in the bottom 5 percent of the distribution with an average teacher for one year, would increase the present value of students' lifetime income by approximately $250,000 per classroom.

[72] In experimental settings, students are randomly assigned to their teachers at the beginning of the school year.

unbiased predictors of teachers' impacts on student achievement, and that the scope for bias is quite small and statistically insignificant (Kane and Staiger, 2008; Kane *et al.*, 2013; Bacher-Hicks, Kane and Staiger, 2014; Chetty, Friedman and Rockoff, 2014a; Bacher-Hicks *et al.*, 2019). The inclusion of student baseline achievement measures seems to be the key behind the unbiased estimation of teacher effects (Kane and Staiger, 2008; Chetty, Friedman and Rockoff, 2014a).

Another central concern regarding teacher value-added estimations is their stability or real persistence over time (Koedel, Mihaly and Rockoff, 2015; Bitler *et al.*, 2019). Critics warn that if teacher effect estimates are not stable over time, their contribution to teacher quality and accountability policies should be limited. In this context, researchers have shown that increasing teacher-level sample sizes (students per teacher) and using multiple years of classroom data improves the predictive value of past teacher value-added over future value-added (McCaffrey *et al.*, 2009; Goldhaber and Hansen, 2013; Bitler *et al.*, 2019).[73] Moreover, the literature currently discriminates between the persistent teacher effect, estimated with at least two classrooms per teacher, and the teacher-classroom effect, also referred to as the classroom effect, estimated with only one year of classroom data per teacher (Chetty, Friedman and Rockoff, 2014a; Jackson, Rockoff and Staiger, 2014; Araujo *et al.*, 2016). Thus, the classroom effect includes not only differences in teacher effectiveness across classrooms, but also random classroom shocks.[74]

[73] Nonetheless, this improvement seems to be non-linear when including data from additional years, unless older data are properly down-weighted (Goldhaber and Hansen, 2013; Chetty, Friedman and Rockoff, 2014a).

[74] Classroom shocks could include particularly disruptive students or events in the specific classroom during the school year or the days in which students were tested.

4.2.2 Teacher and Classroom Effects in Germany

Research related to teacher and classroom effects in Germany is scarce. A major limitation has been the relatively recent introduction of standardized competence tests, which are comparable among federal states for specific grades in primary and secondary schools in Germany.[75] An additional problem has been the lack of publically available teacher panel data.

Jürges and Schneider (2007) attempt to estimate a first ranking of German teachers based on their individual contributions to students' reading performance in grade 4, using cross-sectional data from the Progress in International Reading Literacy Study (PIRLS) 2001.[76] The authors calculate individual teacher random effects by estimating a variance component model of an education production function that takes into account information on student socio-economical background. In addition, they implement a Hausman-Taylor estimator in order to account for possible endogeneity caused by potential non-random assignment of teachers to classrooms and students. Subsequently, the authors present a quality ranking of teachers that consists of teachers significantly above the average, those significantly below the average, and those indistinguishable from the average. Finally, Jürges and Schneider suggest that their model estimation of teacher quality could represent a first step in the development of performance-based payment schemes in Germany. A serious weakness of their study, however, is the lack of a student baseline test score, which is a fundamental measurement for the teacher value-added model and the estimation of reliable teacher effects.[77] In addition, because the authors' data had only one classroom per teacher,

[75] Starting in 2006, universal written comparison tests of math and language for students in grade 3 and grade 8 (*VERgleichsArbeiten [VERA]*) were introduced in Germany, as a consequence of the comprehensive strategy for educational monitoring adopted by the Conference of the Ministers of Education and Cultural Affairs (*Kultusministerkonferenz [KMK]*) (KMK, 2015). In addition, in 2011, the National Educational Panel Study (NEPS) started operating as the first large-scale panel study on educational decisions and outcomes in Germany (Blossfeld, Roßbach and von Maurice, 2011).

[76] PIRLS 2001 tested the reading literacy of students aged 9 to 10 in 35 countries, including Germany. The study sample of Jürges and Schneider (2007) consisted of 4,964 students and 279 teachers.

[77] Jürges and Schneider (2007) argue that they can attribute learning progress to the individual teachers in their sample, because in German primary schools, students typically stay with the same teacher for up to 4 years. The class teacher teaches most or all subjects, and school choice is very limited.

instead of a quality ranking of teachers, their estimates actually correspond to a quality ranking of classrooms driven by teacher contribution to student performance.

A small number of studies have investigated whether specific teacher characteristics can explain between-classroom variation in student achievement gains using multilevel structural equation models in the German school context (Baumert *et al.*, 2010; Kunter *et al.*, 2013). This between-classroom variation can also be interpreted as a random estimate of classroom effects measured in units of student achievement gains. Baumert et al. (2010) use a representative sample of grade 10 classes from the Cognitive Activating Instruction and Development of Students' Mathematics Literacy (COACTIV) study[78] to examine the influence of teachers' content knowledge[79] and pedagogical content knowledge[80] on student progress in math. For their estimation strategy, Baumert et al. implement a two-level structural equation model where the variance in math achievement is decomposed into a within-classroom or individual level component, and a between-classroom or classroom level component. At the individual level, the model takes into account student baseline achievement in math and reading (grade 9), as well as other cognitive and socioeconomic characteristics as explanatory variables. [81] Subsequently, the between-classroom variance is explained by

[78] The COACTIV study was conducted in Germany between 2003 and 2004 as an extension to the Programme for International Student Assessment (PISA) 2003 of the Organization for Economic Co-operation and Development (OECD). It extended the original PISA cross-sectional design to a grade-base study comprising a one-year period from the end of grade 9 to the end of grade 10. Students from the study sample were administered achievement tests at the end of grade 9 and 10, as well as questionnaires assessing their cognitive ability, mathematics instruction and family background. The COACTIV study also applied tests of content and pedagogical content knowledge to the math teachers of the study sample. A total of 181 teachers, 194 classrooms and 4,353 students participated in the study (Baumert *et al.*, 2010).

[79] Teachers' mathematical content knowledge was assessed with a paper-and-pencil test that covered conceptual topics that are compulsory from grade 5 to 10 (Baumert *et al.*, 2010).

[80] Teachers' mathematical pedagogical content knowledge was assessed in three dimensions: first, the "tasks" dimension which assessed teachers' ability to identify multiple solution paths; second, the "students" dimension which evaluated their ability to recognize students' misconceptions, difficulties, and solution strategies in the context of classroom situations; and third, the "instruction" dimension which assessed teachers' knowledge of different representations and explanations of standard mathematics problems within classroom situations (Baumert *et al.*, 2010).

[81] The authors acknowledge that by grade 10, students had already been allocated to academic and non-academic secondary tracks based on their performance and general ability in Germany. They therefore highlight the importance of introducing baseline achievement in the model to account for the sorting process.

classroom track (academic or non-academic), and teacher mathematical content knowledge and pedagogical content knowledge. Baumert et al. point out that controlling for academic track at the classroom level is highly relevant because, even though teachers are centrally assigned to schools by federal states, their allocation to school tracks is determined by their choice of teacher training program.[82] The authors' results show that, once student individual characteristics are taken into account, a maximum of 4.6 percent of the variance in math achievement can be explained by differences at the classroom level. Moreover, they find a significant and substantial positive effect of teacher content knowledge and pedagogical content knowledge on the between-classroom variation in students' math achievement gains, with pedagogical knowledge having the greater predictive power for student progress.[83]

Kunter et al. (2013) complement the study of Baumert et al. (2010) by examining, in addition to pedagogical content knowledge, the impact of teachers' constructivist beliefs,[84] enthusiasm for teaching,[85] and self-regulation[86] on student mathematical learning in grade 10. Their research also uses data from the COACTIV study and implements two-level structural equation models, which include student baseline achievement in math (grade 9). Surprisingly, the model does not take into account tracking into academic and non-academic secondary schools and classrooms. Kunter et al.'s findings indicate that students whose teachers had better pedagogical content knowledge, endorsed constructivist beliefs, and were enthusiastic about

[82] In Germany, universities offer different teacher education programs that correspond to the tracking system implemented after grade 4 (Baumert *et al.*, 2010; KMK, 2019).

[83] Teacher pedagogical content knowledge alone explained around 39 percent of the between-classroom variation in achievement gains at the end of grade 10.

[84] In the study, constructivist beliefs are described as conceptions that endorse the principals of active and constructive learning in the classroom. They contrast with the transmissive beliefs that tend to treat students as passive receivers of information. Constructivist beliefs were assessed using three subscales which measured the degree to which teachers understood mathematical knowledge as process, favored independent and insightful discursive learning, and thought it important to foster students' mathematical independence (Kunter *et al.*, 2013).

[85] Enthusiasm for teaching is defined as enjoyment of teaching activities. It was measured with on a short scale of two items developed by the COACTIV study (Kunter *et al.*, 2013).

[86] Self-regulation is described as teachers' ability to engage while simultaneously monitoring their behavior and copping with stressful situations. Self-regulatory style was measured using a procedure developed by Klusmann et al. (2008) based on eight subscales from the Occupational Stress and Coping Inventory (Kunter *et al.*, 2013).

teaching showed significantly higher achievement gains in mathematics. Thus, these characteristics were positively associated with the between-classroom variation in student achievement. Their analysis also shows that teachers' self-regulation had no direct effect on student outcomes. In addition, they find that teachers' general cognitive ability, measured by their self-reported grade point average (GPA) at the university entry qualification *Abitur*, was unrelated to student achievement.

Enzi (2017) reports a first attempt to estimate the distribution and average value-added of language and math teachers in German secondary schools. He uses three-year data of students and their teachers from the Starting Cohort 3 (SC3) of the NEPS. The study sample is limited to students that shared the same math or German language teacher in grades 5 and 6.[87] In his analysis, Enzi estimates a teacher value-added model where students' language and math competence scores in grade 7 are explained by two-year lagged student test scores (grade 5), contemporaneous student and family background inputs and teacher fixed effects. Using the teacher fixed effects' estimates, he generates distributions of teacher quality for math and language, and reports standard deviations of 0.134 and 0.155 respectively. Since competence tests for grade 7 were administered by the NEPS in the first semester of the school year, the teacher effects are attributed to teachers who taught math or language between grades 5 and 6. This is a serious weakness in the study because students in grade 7 had already been exposed to other math and language teachers for between two and five months (NEPS, 2019b). Thus, the effects of grade 6 and grade 7 teachers are unfortunately confounded. Another problem in the estimation is that it does not control for tracking of students into academic and non-academic secondary classrooms.

In addition, Enzi stresses that his results are upper-bound estimates because he neither applies Empirical Bayes shrinkage to adjust the teacher effect estimates by their level of precision, nor takes into account classroom or peer effects, and only observes one teacher per classroom. Given the absence of a shrinkage process, Enzi does not attempt to explain the teacher value-added estimated with specific teacher characteristics and opts to introduce them instead of the teacher fixed effects in his

[87] The student sample consisted of 1,939 students for language and 2,329 students for math. The total teacher sample consisted of 211 language teachers and 197 math teachers (Enzi, 2017).

original model. As a result, he finds some evidence that teachers' self-reported *Abitur* GPA is associated with student competence gains in math, but only at the 10 percent significance level. In his nonlinearity analysis, Enzi also suggests that teachers' First and Second State Examination grades might be associated with competence gains in math for the best quartile of teachers, yet only at the 10 percent significance level. Nonetheless, these associations only hold when *Abitur* GPA and the First and Second State Examination grades are introduced in three independent regression models. Any potential effect disappears when all three grades are taken into account in the same model.

As shown, research on teacher and classroom effects in Germany has relied on cross-sectional data or relatively small student panel samples, which has imposed limitations to its development and potential contribution. In addition, the literature has mainly focused on the lower secondary level, when tracking into different school types based on students' cognitive skills and families' background has already taken place, with potential negative implications for the estimates of teacher and classroom effects. Our research, on the one hand, partially overcomes the data limitation by generating a rich short-panel of teachers and their students between grades 1 and 2 from the NEPS SC2. On the other hand, our research contributes to the existing literature by examining for the first time the distribution of classroom effects driven by teacher quality in the first years of the German primary school system. These are pre-tracking years, in which educational quality is particularly critical for the development of children's cognitive and non-cognitive skills, and consequentially later-life outcomes (Cunha, Heckman and Schennach, 2010; Heckman, Pinto and Savelyev, 2013; Elango *et al.*, 2016; García *et al.*, 2020). Finally, our research takes into account, for the first time, the effect of institutional differences among federal states on the estimation of the classroom effects.

4.2.3 The German Educational System[88]

In Germany, the 16 federal states determine education policies. The Conference of the Ministers of Education and Cultural Affairs (*Kultusministerkonferenz [KMK]*), a commission of the relevant ministers from the federal states, sets the framework within which the federal states then decide upon different policies. The following paragraphs give a broad explanation of the system, but it should be noted that in some aspects the number of federal states diverge from the description.

Full-time school attendance is compulsory for nine to ten years. Children normally start school aged six. Following comprehensive primary schooling, which typically encompasses four (but sometimes six) years, children are sorted into different tracks for secondary schooling. This tracking process is based on an overall school assessment of children's aptitudes, accompanied by consultations with their parents. Historically there have been three tracks in all federal states: the lower vocational track, the *Hauptschule*, an intermediate vocational track, the *Realschule*, and the academic track, the *Gymnasium*. In addition, most federal states have some form of comprehensive schooling, where more than one type of school-leaving certificate is offered. Only the Gymnasium and some comprehensive schools directly lead to the university entry qualification, the *Abitur*. The *Abitur* GPA summarizes the students' final grades from the last four semesters of schooling and from the exit examinations.

Prospective teachers have to attend a teacher training at a university or college. Typically, the course of studies already determines the school type at which the prospective teacher will work.[89] The federal states regulate the details of two stages of the teacher training, which consist of theoretical education at the university (including periods of practical training), and practical training in a school setting. The First State Examination, equivalent to Bachelor or Master's examinations, depending on the

[88] For a comprehensive explanation of most facets of the German Education System please refer to KMK (2019), the official publication used to develop this section.

[89] Primary school teachers attend training programs specialized in primary school, or primary and lower secondary school types.

federal state,[90] marks the end of the first stage of teacher training. The examination thus covers theoretical knowledge in educational science, subject knowledge, and pedagogics. After the First State Examination, prospective teachers proceed to the preparatory service (*Vorbereitungsdienst*), where they continue to train in teacher training institutes (*Studienseminare*) and simultaneously work increasingly independently as teachers at schools. Subsequently, teachers become fully qualified upon passing the Second State Examination.[91]

In a next step, young teachers apply for permanent employment in the public sector by sending their application to either the Ministry of Education,[92] or the relevant school supervisory authority in the federal state. Placement decisions are made centrally by the relevant authority based on vacancies and on the applicant's aptitude, qualifications and record of achievements. Sometimes specific schools advertise positions. In this case, the school might also be involved in the selection process, but the Ministry or school authority always hires the teacher. The demand for teachers differs by subjects, school types and across the different federal states. This implies no legal entitlement to a teacher position for qualified teachers. Most federal states appoint teachers as civil servants on probation, followed by a lifelong civil servant appointment after successful completion of the probation phase. Some federal states also employ teachers as regular salaried employees.[93] Berlin and Saxony only employ teachers, and do not appoint them as civil servants.

Once appointed as a civil servant or employed, most federal states only allow a promotion to a higher salary group if the teacher also takes on new responsibilities or a

[90] Each federal state decides whether the teacher training programs are concluded with a state examination at the Bachelor level, or if they follow the graduated structure of higher education studies, where the Master's degree replaces the First State Examination as a rule.

[91] The Second State Examination usually consists of four parts: (i) a written paper relating to educational theory, pedagogic psychology, or didactics of a subject studied; (ii) a practical teaching examination or demonstration class; (iii) an examination of educational theory, legislation or school administration; and (iv) an examination of didactic and methodological issues in the subjects studied.

[92] Full name: Ministry of Education and Cultural Affairs.

[93] This may be the case for substitute teachers, who are hired to cover for sickness or parental leave and thus are only hired temporarily. Some teachers also do not meet the requirements for becoming a civil servant, e.g. because they are not healthy enough. In this case, they can also be employed as salaried personnel.

new position. Changes to a different school within or across federal states are possible, but teachers need to ask for permission from the relevant Ministry of Education or school supervisory authority and the desired school needs to have a suitable vacant position. Therefore, teachers only have limited scope to choose their schools.[94]

4.3 Data

4.3.1 National Educational Panel Study (NEPS)

The NEPS is a large-scale panel study on educational decisions and outcomes in Germany (Blossfeld, Roßbach and von Maurice, 2011). In order to depict all age groups without waiting for an entire lifespan, the NEPS consists of six different starting cohorts from newborns to adults, each a representative sample of the relevant cohort.

In our analyses, we rely on data from the Kindergarten SC2.[95] The Kindergarten Cohort initially consists of a target population of kindergarten children at age four, who are longitudinally followed into primary school and beyond. The NEPS SC2 sample was drawn in a multi-stage approach, where institutions were drawn in a first step and children in a second. First, a nationally representative sample of German primary schools was chosen, which formed the basis for the subsequent grade 1 survey (wave 3). Then, these elementary schools were connected to all kindergartens from which first grade students typically came, and a random sample of these linked institutions was drawn for the first kindergarten survey (wave 1). Between the last kindergarten year and the first grade of primary school, there was substantial panel attrition and subsequent student resampling. Aiming to achieve a sufficiently large and representative sample, we refrain from using kindergarten data as a baseline, and instead focus our analyses on measurements in grades 1 and 2 of primary school, which correspond to waves 3 and 4 of the Kindergarten SC2.

[94] However, since placement decisions are partially determined by teacher qualifications, exceptionally good teachers might have better chances to be placed in a school or region of their liking.

[95] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Kindergarten, doi:10.5157/NEPS:SC2:9.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS has been carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

The NEPS data is well suited for our analyses because it contains children's competence measurements and survey information from the children, their parents, classrooms, teachers and schools. Participating children completed tests in various competence domains: math, grammar[96] and science in grade 1, and math and early reading in grade 2. Based on these tests, the NEPS provides weighted maximum likelihood estimates (WLE) as measures of children's competences for math, grammar and science, which are normally distributed and have been standardized by grade to have a zero mean and a unit standard deviation.[97] A raw measurement of early reading competence in grade 2 is also provided.[98] We standardized it to have a zero mean and a unit standard deviation.

Math and early reading competences in grade 2 are the outcome variables of our value-added estimates, and math, grammar and science competences in grade 1 are the baseline measurements.[99] The NEPS math competence tests in grade 1 and 2 were designed in such a way that scores derived in different waves relate to the same scale and allow an accurate competence measurement within each age group across grades; accordingly, tests are comparable across grades.[100] Early reading competence was not

[96] The grammar test corresponds to listening comprehension at sentence level for first grade children.

[97] Weighted maximum likelihood estimation (WLE) is an application of the Item Response Theory, which delivers unbiased estimates of competence parameters. The NEPS's WLE estimates are calculated following the procedure outlined by Warm (1989). The NEPS SC2 data provides uncorrected and corrected WLE estimates of math and science competence for grade 1, and of math competence for grade 2. Uncorrected WLE estimates of grammar competence for grade 1 are also provided. Corrected WLE estimates correspond to the uncorrected WLE estimates standardized by grade to have zero mean and a unit standard deviation. We normalize the uncorrected WLE estimates of grammar competence for grade 1 to obtain the corrected WLE estimates. Corrected and uncorrected WLE competence estimates have a Pearson correlation of 1.0 per grade, which thus means that they represent the same variable. NEPS recommends using uncorrected WLE estimates for longitudinal comparison of competence development between grades, and corrected WLE estimates for cross-sectional research questions (Schnittjer and Gerken, 2018). In our value-added regression analysis, nonetheless, the inclusion of corrected or uncorrected WLE estimates produce the same results. We opt to present results from corrected WLE estimates in order to facilitate interpretation and comparability with other studies.

[98] Early reading competence was measured using the ELFE test (Lenhard and Schneider, 2006). NEPS provides a sum scoring of the test following the test authors' recommendation.

[99] Based on the findings of Chetty, Friedman and Rockoff (2014a), Lockwood and McCaffrey (2014) among others, we use multiple baseline tests scores in the same and other subjects to increase the precision of our estimates.

[100] For technical details on the linking procedures of math competence see Fischer et al (2016), Schnittjer and Fischer (2018) and Schnittjer and Gerken (2018).

measured in grade 1. For that reason, we use the grammar test as the closest measurement of children's baseline language competence. Competence tests were administered during the second semester of the 2012-2013 school year in grade 1, and during the first semester of the 2013-2014 school year in grade 2, with a span of 6 to 9 months between tests for most of the students.

The NEPS survey data also provides information on child age, gender, migration background and number of siblings, parent years of education[101] and International Socio-Economic Index of Occupational Status (ISEI)[102]. In our analysis, migration background is a dummy variable that takes the value of one if at least one parent was not born in Germany. We generate the variables parent years of education and ISEI as the highest value among parents in the household.

In addition to the child and parental data, the NEPS provides rich information on classrooms and teachers, which is crucial for identifying potential factors defining teacher quality. The teacher characteristics included in our analysis are gender, years of teaching experience, migration background, self-reported *Abitur* GPA, self-reported First and Second State Examination grades, whether the teacher has passed the Second State Examination, constructivist beliefs, exhaustion levels, and parental evaluation of teacher quality.

We calculate teacher years of experience as the time difference between the NEPS survey year and the year of the First State Examination for each teacher.[103] In our analysis, a teacher has migration background if she gives a positive answer to this question and indicates that she or at least one of her parents was not born in Germany. In addition, self-reported *Abitur* GPA and First and Second State Examination grades are measured on a scale from 1.0 to 4.0, with 1.0 being the best possible grade, and 4.0

[101] Years of education are estimated by the NEPS as a function based on the Comparative Analysis of Social Mobility in Industrial Nations (CASMIN) (Zielonka and Pelz, 2015), which is an internationally comparable educational classification developed in Germany (König, Lüttinger and Müller, 1988; Lechert, Schroedter and Lüttinger, 2006).

[102] The International Socio-Economic Index of Occupational Status (ISEI-08) is estimated from the International Standard Classification of Occupations (ISCO-08).

[103] Alternatively, we estimated years of experience as the time between the panel survey year and teacher's *Abitur* year plus three years of university instruction. Both measurements have a correlation of 0.968, with the first measurement being our preferred estimation.

the minimum passing grade. We standardize these self-reported grades to have zero mean, and one unit standard deviation with respect to the full sample of teachers.

Following Kunter et al.'s (2013) findings, we build indicators of teacher constructivist beliefs and teacher exhaustion (as opposed to enthusiasm for teaching). Our indicator of teacher constructivist beliefs is based on four items[104] available in the NEPS classroom survey of grade 1 for this purpose, which are taken from the constructivist scale of the Teaching and Learning International Survey (TALIS) 2008 study (Demmer and von Saldern, 2010; Organisation for Economic Co-Operation and Development [OECD], 2010). Teachers indicated their level of agreement with all items on a 4-point Likert scale. We enter the answers into an index, which is standardized to have zero mean and unit standard deviation with respect to the full sample of teachers. In order to generate our indicator of teacher exhaustion, we use a short scale of two items available in the NEPS classroom survey of grade 2, which asked teachers whether they felt often exhausted at school and if their workload was too heavy. Teachers indicated their level of agreement with the two items on a 5-point Likert scale, which we use to generate an index standardized to have zero mean and unit standard deviation with respect to the full sample of teachers.

Finally, in the NEPS parent survey of grade 2, families were asked to indicate their level of agreement on whether their school's teachers tried to meet children's needs using a 4-point Likert scale. We use this information to generate a parental evaluation of teacher quality indicator, following recent literature on whether parents can discriminate between good and poor teachers (Araujo *et al.*, 2016), and taking into account the growing international policy efforts to incorporate parent perspectives into teacher quality assessments (Steinberg and Donaldson, 2016; Fernández, LeChasseur and Donaldson, 2018). Our indicator corresponds to the average classroom assessment of the parents for each teacher. We normalize it to have zero mean and unit standard deviation with respect to the full sample of students.

[104] Items corresponded to: (i) my role as a teacher is to make it easier for the students to investigate and explore things; (ii) students will learn best when they try to find solutions to problems independently; (iii) students should be given the possibility to reflect on solutions themselves before the teacher shows the approach to the solution; and (iv) thinking and reasoning processes are more important than specific content of the syllabus.

At the classroom level, we have access to data on classroom size and proportion of female students, based on information of the full classroom as opposed to the NEPS student sample. In addition, we calculate the average ISEI of children in the classroom based on the sample of parents who participated in the NEPS survey.

We include students in the analysis sample if we can link them to a classroom with a teacher unique identifier. Additionally, we require children to be taught by the same teacher in grade 1 and grade 2[105], for there to be at least 5 students per teacher for the value-added analyses, and no missing information on any of the variables used for value added estimations. We also exclude children with special needs. This results in an analysis sample of 1,843 students and 251 teachers in the math sample, and 1,753 students and 240 teachers in the language sample.

4.3.2 Descriptive Statistics

Descriptive statistics are provided in table 4.1 for students and in table 4.2 for teachers. In both tables, Column (1) depicts descriptive statistics for the full NEPS SC2 sample of 4,564 children, whom we can link to their respective teacher and classroom data, and 680 teachers.[106] Columns (2) and (6) show descriptive statistics for the dropout sample, for math and language respectively. Likewise, Columns (3) and (7) present descriptive statistics for the math and language analysis sample. Column (4) and (8) display the difference between the dropout and the analysis samples, and the respective p-value from a t-test for equality, for math and language respectively. Finally, columns (5) and (9) present the normalized difference as suggested by Imbens and Wooldridge (2009).[107]

[105] We apply this restriction because the NEPS competence tests in grade 1 and grade 2 of the SC2, were not applied right at the beginning or the end of the respective school year, but in the middle of it. In this context, competence growth can only be attributed to teachers who had the same group of students in grades 1 and 2.

[106] This implies a lower number of observations for a number of variables due to missing data in the columns for the full sample and the dropout sample. The number of observations is stable over all variables in the student analysis sample, as we require information on all variables in the analyses for inclusion.

[107] A t-test might imply a statistically significant difference between samples, because of sample size or variable scaling, even though the samples are not substantially different from each other (Imbens, 2015). The normalized difference frees the sample comparison from sample size and scale of the variables, by correcting the difference between samples by their respective standard deviation. Imbens and Wooldridge

Table 4.1: Descriptive Statistics Students

		Math				Language			
	Full sample (1)	Dropout sample (2)	Analysis sample (3)	Diff (p-value) (4)	Norm Diff (5)	Dropout sample (6)	Analysis sample (7)	Diff (p-value) (8)	Norm Diff (9)
Competence measures	0.04	-0.09	0.19	-0.29***	0.19	-0.10	0.21	-0.31***	0.20
G1: Math (WLE)	(1.09)	(1.11)	(1.06)	(0.00)		(1.10)	(1.06)	(0.00)	
G2: Math (WLE)	0.05	-0.06	0.19	-0.25***	0.16	-0.07	0.22	-0.29***	0.18
	(1.15)	(1.15)	(1.13)	(0.00)		(1.14)	(1.13)	(0.00)	
G1: Grammar (WLE)	0.05	-0.09	0.22	-0.31***	0.23	-0.08	0.23	-0.31***	0.23
	(0.97)	(0.96)	(0.95)	(0.00)		(0.96)	(0.95)	(0.00)	
G2: Early reading (Std)	0.02	-0.08	0.15	-0.23***	0.17	-0.08	0.15	-0.23***	0.16
	(0.98)	(0.96)	(1.00)	(0.00)		(0.96)	(1.00)	(0.00)	
Child demographics									
Age [Months]	92.67	92.82	92.46	0.37***	0.05	92.78	92.51	0.26*	0.06
	(4.48)	(4.62)	(4.27)	(0.01)		(4.64)	(4.22)	(0.05)	
Female	0.51	0.51	0.53	-0.02	0.03	0.50	0.53	-0.02	0.03
	(0.50)	(0.50)	(0.50)	(0.15)		(0.50)	(0.50)	(0.11)	
Migration background	0.20	0.22	0.19	0.03*	-0.06	0.21	0.19	0.02*	-0.06
	(0.40)	(0.41)	(0.39)	(0.06)		(0.41)	(0.39)	(0.08)	
Parental background									
Years of education	15.00	14.83	15.15	-0.32***	0.12	14.80	15.20	-0.40***	0.14
	(2.30)	(2.32)	(2.28)	(0.00)		(2.33)	(2.26)	(0.00)	
ISEI	59.56	57.93	61.01	-3.08***	0.14	57.63	61.47	-3.84***	0.16
	(19.00)	(19.21)	(18.70)	(0.00)		(19.23)	(18.58)	(0.00)	
Number of siblings	1.14	1.15	1.13	0.02	-0.01	1.15	1.13	0.02	-0.01
	(0.87)	(0.89)	(0.85)	(0.43)		(0.89)	(0.85)	(0.43)	
Number of Students	4564	2721	1843			2811	1753		

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes*: This table contains means and standard deviations (in parenthesis) of child characteristics for the full, dropout and analysis sample. The full sample contains children that we can link to respective teacher, classroom, and parent data, and who have no special needs. The analysis sample includes children who were taught by the same teacher in grades 1 and 2, who belonged to classrooms with at least five students per teacher in the grade 2 sample, and who had no missing information on any variable used for value-added estimations. Most variables have fewer observations than stated in the full and dropout sample. Diff displays the difference between analysis and dropout sample, and the respective p-value (in parenthesis) from a t-test for equality. Norm Diff displays normalized differences as suggested by Imbens and Wooldridge (2009), where the critical value typically is 0.25 or -0.25. Variables correspond to grade 2 unless stated otherwise. Math and grammar competences are measured as weighted maximum likelihood estimations (WLE) and standardized by grade to have a zero mean, and a one-unit standard deviation. Early reading competence is standardized to have a zero mean and a one-unit standard deviation. Parents' years of education are estimated as a function based on the Comparative Analysis of Social Mobility in Industrial Nations (CASMIN). ISEI corresponds to the International Socio-Economic Index of Occupational Status (ISEI-08) estimated from the International Standard Classification of Occupations (ISCO-08). Parents' years of education and ISEI correspond to the highest values among parents in the household.* Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

As indicated, table 4.1 depicts the descriptive statistics of students for the math and language analysis samples. The variables always correspond to grade 2 unless

(2009) suggest a substantial difference between the samples if the normalized difference exceeds 0.25 or -.25.

otherwise stated. Children in the analysis samples are roughly 7.5 years old, half are female, and a fifth have a migration background. The average highest years of education among parents in the household is about 15 years, equivalent to a vocational training degree after completion of *Abitur* in the CASMIN classification (Zielonka and Pelz, 2015). The mean highest ISEI among parents is around 61, which corresponds to a medium-high level. Children in the analysis samples have on average one sibling. While the t-tests show some differences between the analysis and the dropout samples, the criterion for a substantial difference according to Imbens and Wooldridge (2009) is not met by any of the variables in the math or language samples. It is, however, close to the cutoff for the difference in grammar competence in grade 1, which implies a slightly better qualified analysis sample.

Regarding the teacher descriptive statistics, table 4.2 provides the respective numbers for the math and language analysis samples. More than 90 percent of the primary school teachers in our analysis samples are female. While this number seems strikingly high at first sight, official numbers confirm that roughly 90 percent of primary school teachers in Germany are female (Statistisches Bundesamt, 2019). Teachers are on average 47 years old, they have around 22 years of experience, and practically all of them have the *Abitur*. The average self-reported *Abitur* GPA is around 2.4, equivalent to a good achievement in the German grading system. Self-reported First and Second Examination Grades are on average around 2.0, which also represents a good performance. Four out of five teachers in the samples have already passed their Second State Examination or equivalent. About six percent of teachers have a migration background in the math analysis sample, and seven percent in the language sample. The average non-standardized constructivist beliefs index has a rather high value, with 3.38 points out of 4 possible. On average, the non-standardized exhaustion index has a moderate value of 2.89 points out of 5 possible. Parental evaluation of teacher quality also is high, with an average of 3.60 out of 4. Finally, the mean class size is around 22 students. Once again, even though the t-tests show significant differences between the analysis and the dropout samples, the criterion for a substantial difference according to normalized differences is not met by any of the variables in math or language. Consequently, we can conclude that teachers in our math and language analysis samples are not substantially different from teachers in the dropout samples.

Table 4.2: Descriptive Statistics Teachers

		Math				Language			
Teacher	Full sample (1)	Dropout sample (2)	Analysis sample (3)	Difference (p-value) (4)	Norm Diff (5)	Dropout sample (6)	Analysis sample (7)	Difference (p-value) (8)	Norm Diff (9)
Female	0.93	0.94	0.93	0.01***	-0.05	0.94	0.93	0.02***	-0.05
	(0.25)	(0.24)	(0.26)	(0.00)		(0.24)	(0.26)	(0.00)	
Age	46.02	45.38	47.09	-1.15***	0.10	45.56	46.85	-1.26***	0.08
	(10.73)	(10.85)	(10.47)	(0.00)		(10.91)	(10.39)	(0.00)	
Experience	20.38	19.53	21.77	-0.85***	0.08	19.76	21.45	-0.88***	0.05
	(11.50)	(11.59)	(11.23)	(0.00)		(11.69)	(11.10)	(0.00)	
Has *Abitur*	0.94	0.93	0.94	-0.02***	0.02	0.93	0.95	-0.01	0.02
	(0.24)	(0.25)	(0.23)	(0.00)		(0.25)	(0.23)	(0.11)	
Abitur GPA	2.46	2.48	2.41	0.06***	-0.09	2.48	2.40	0.08***	-0.11
	(0.52)	(0.53)	(0.49)	(0.00)		(0.53)	(0.50)	(0.00)	
FSE grade	1.99	1.99	1.98	0.01	-0.02	2.00	1.96	0.05***	-0.08
	(0.47)	(0.49)	(0.42)	(0.59)		(0.49)	(0.43)	(0.00)	
Passed SEE	0.84	0.87	0.80	0.03***	-0.08	0.87	0.80	0.05***	-0.08
	(0.36)	(0.34)	(0.40)	(0.00)		(0.34)	(0.40)	(0.00)	
SEE grade	1.93	1.93	1.93	0.01	-0.02	1.95	1.90	0.07***	-0.08
	(0.57)	(0.59)	(0.55)	(0.36)		(0.59)	(0.54)	(0.00)	
Migration background	0.05	0.04	0.06	-0.03***	0.10	0.04	0.07	-0.04***	0.11
	(0.22)	(0.20)	(0.25)	(0.00)		(0.20)	(0.25)	(0.00)	
Constructivist beliefs	3.38	3.38	3.38	-0.00	-0.02	3.38	3.38	0.00	-0.01
	(0.39)	(0.39)	(0.39)	(0.59)		(0.39)	(0.38)	(0.63)	
Exhaustion	2.99	3.05	2.89	0.10***	-0.10	3.05	2.89	0.15***	-0.10
	(1.04)	(1.00)	(1.11)	(0.00)		(1.00)	(1.11)	(0.00)	
Parental evaluation	3.59	3.59	3.60	-0.03	0.04	3.58	3.61	-0.04**	0.05
	(0.36)	(0.41)	(0.26)	(0.19)		(0.41)	(0.26)	(0.03)	
Class size	21.92	21.70	22.27	-0.67***	0.18	21.76	22.20	-0.81***	0.16
	(3.42)	(3.33)	(3.55)	(0.00)		(3.34)	(3.55)	(0.00)	
Number of Teachers	680	429	251			440	240		

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This table contains means and standard deviations (in parenthesis) of teacher characteristics of the full sample, the dropout, and analysis samples for math and language. The full sample contains all teachers with an individual identification number who can be linked to a classroom. The analysis sample comprises teachers who taught the same group of children in grades 1 and 2, who had at least five students in the grade 2 sample, and whose students had no missing information on any variable used for the value-added estimations. Most variables have fewer observations than stated in the full, dropout, and analysis sample. Diff displays the difference between analysis and dropout sample, and the respective p-value (in parenthesis) from a t-test for equality. Norm Diff displays normalized differences as suggested by Imbens and Wooldridge (2009), where the critical value typically is 0.25 or -0.25. Self-reported Abitur GPA, First State Examination (FSE) and Second State Examination (SSE) grades have a scale that ranges from 1.0 to 4.0, with 1.0 being the best possible grade and 4.0 the minimum passing grade. The constructivist beliefs' index and the parental evaluation indicator are on a scale ranging from 1 to 4, with 4 being the highest possible score. The exhaustion index is on a scale from 1 to 5, with 5 being the highest possible score. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

4.4 Estimation Strategy

Our model is derived from the value-added specification of the regular education production function formalized by Todd and Wolpin (2003), but rooted in the longstanding empirical education production literature (Ben-Porath, 1967; Hanushek, 1971, 1979). We apply a lagged-score specification of a value-added model, which places baseline test scores on the right-hand-side.[108] Subsequently, given that we have one teacher per classroom, we estimate individual teacher-classroom effects derived from the value-added specification using adjusted fixed effects, as well as random effects.

4.4.1 Adjusted Fixed Effects

In our first estimation strategy, we implement a two-step or "average residuals" value-added model (Koedel, Mihaly and Rockoff, 2015). Specifically, in a first step we estimate the following equation using OLS:

$$Y_{isjt} = \alpha_o + Y_{isjt-1}\beta_1 + X_{isjt}\beta_2 + C_{sjt}\beta_3 + n_{isjt} \tag{4.1}$$

where Y_{isjt} is the test score in math or language competence for student *i* at school *s* with teacher *j* in year *t,* Y_{isjt-1} is a vector of lagged competence test scores (math, grammar and science), X_{isjt} is a vector of child characteristics (age, gender, migration background and time between tests) and family background (parents' years of schooling, ISEI and number of siblings), C_{isjt} is a vector of classroom characteristics (classroom size, proportion of female students, average parents' ISEI), and α_o is a federal state fixed effect. We introduce federal state fixed effects in our model to

[108] Even though the lagged test score parameter may be poorly estimated in the regression, we can consistently estimate the teacher-classroom effects with a lagged-score specification under two conditions. First, past shocks to learning decay at the same rate as learning from family and school-related sources (common factor restriction), and, therefore, errors are serially uncorrelated (Guarino, Reckase and Wooldridge, 2015). Second, the baseline tests scores serve as a good proxy for unobservable individual characteristics (Guarino, Reckase and Wooldridge, 2015). Empirical evidence has shown that the lagged-score specification of the teacher value-added model is the most robust, and even performs better than the gain-score specification in the estimation of teacher effects (Guarino, Reckase and Wooldridge, 2015; Koedel, Mihaly and Rockoff, 2015).

capture specificities of the educational systems, and school quality at the state level, since education is a competence of the federal states in Germany. Standard errors are clustered at the student level.[109]

In equation (4.1), n_{isjt} is a composed error term attributed to individual teacher effects and classroom shocks, and unobserved school-level or student-level effects. In a second step, the composed error term n_{isjt} is averaged among the individual teacher-classroom fixed effects:

$$n_{isjt} = \theta_{sjt} + e_{isjt} \tag{4.2}$$

The vector θ in equation (4.2) contains the individual classroom effects, which are mainly driven by teacher quality differences across classrooms. The error term e_{isjt} is composed of the unobserved school-level or student-level effects, which are expected to be uncorrelated to the classroom effect in German primary schools.

We acknowledge that in the presence of non-random assignment of students to teachers, unobserved student characteristics might be correlated to the classroom effects, and consequently, their estimations could be biased. In addition, a matching of teachers and schools within states based on unobserved quality factors could also bias the estimations. However, we argue that our value-added model specifications have the potential to lead to unbiased estimators of classroom effects, mainly driven by teacher quality differences across classrooms, because matching of students to teachers is not prevalent in primary schools in Germany. On the one hand, students are not subject to any tracking based on their ability in the first four years of primary school, and most of them must attend the nearest public school to their homes (KMK, 2019). [110] On the

[109] We rerun our analysis with standard errors clustered at the classroom level. Results do not change and are available upon request.

[110] According to the German Conference of the Ministers of Education and Cultural Affairs (2019), in order to complete general compulsory schooling, pupils generally must attend the local primary school. The exceptions to this rule are the states of Nordrhein-Westfalen and Schleswig-Holstein, where parents are free to enroll their child in a primary school other than that nearest their home. In Berlin, enrolment in a primary school other than the nearest to the home may take place subject to place availability.

other hand, teachers are centrally allocated to schools at the federal state level, based on the teaching subjects required at the schools, as opposed to teacher or school preferences (Baumert *et al.*, 2010; KMK, 2019).

In addition, vast empirical evidence has shown that education production function models that take into account baseline student performance have small and statistically insignificant scope for bias in the estimation of teacher effects, even in the presence of non-random assignment of students to teachers (Kane and Staiger, 2008; Kane *et al.*, 2013; Bacher-Hicks, Kane and Staiger, 2014; Chetty, Friedman and Rockoff, 2014a; Bacher-Hicks *et al.*, 2019). In other words, lagged or baseline performance empirically seems to be a sufficient statistic for unobserved student and family histories, as well as unobserved endowment of mental capacity or ability. Furthermore, our estimates take into account potential student peer effects because we control for classroom average characteristics in equation (4.1).

As a final step, we implement a procedure known as Empirical Bayes (EB) shrinkage to adjust our classroom effects' estimates by their level of precision, which is commonly done in research and policy applications (Koedel, Mihaly and Rockoff, 2015). The implementation of EB shrinkage recognizes that value-added estimates of teachers matched to fewer students are less precise because one or two students with unusually high or low achievement growth can more heavily influence these estimates (Herrmann, Walsh and Isenberg, 2016). Accordingly, the EB shrinkage procedure places less weight on imprecise initial value-added estimates (fewer students) and greater weight on more precise ones.[111]

We follow Herrmann, Walsh and Isenberg (2016)[112] in the implementation of the EB shrinkage procedure outlined by Morris (1983), according to which the classroom's adjusted fixed effect can be written as follows:

[111] If class size were constant across teachers and time, the EB estimates would be identical to the original classroom effects estimates produced by our model specification (Guarino, Reckase and Wooldridge, 2015).

[112] We apply the Stata program developed by the Mathematica Policy Research Educator Impact Laboratory, version 1.00 -25Feb2016.

$$\hat{\theta}_j^{EB} \approx \left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2+\hat{\sigma_j}^2}\right)\hat{\theta}_j$$

(4.3)

Where $\hat{\theta}_j^{EB}$ is the classrom's EB estimate, $\hat{\theta}_j$ is the pre-shrinkage classroom point estimate for teacher j from the value-added regression model, $\hat{\sigma_j}^2$ is the heteroskedasticity-robust variance estimate of $\hat{\theta}_j$, and $\hat{\sigma}$ is an estimate of the standard deviation of the classroom effects, which is purged of sampling error and constant for all classrooms.

Subsequently, we attempt to explain the classroom adjusted fixed effect (FE) using our rich vector of teacher observable characteristics τ_j (gender, experience, migration background, *Abitur* GPA, First State Examination grade, passed Second State Examination, constructivist beliefs, exhaustion and parental evaluation) as explanatory variables in the following OLS regression:

$$\hat{\theta}_j^{EB} = \beta_0 + \tau_j\beta_1 + u_j$$

(4.4)

Under this scenario, the shrinkage procedure is particularly valuable because it reduces attenuation bias (Koedel, Mihaly and Rockoff, 2015).

4.4.2 Random effects

The classrooms' EB estimates can also be directly obtained from a value-added multilevel model, where classroom effects are estimated as random intercepts (Ballou, Sanders and Wright, 2004; Rabe-Hesketh and Skrondal, 2012; Guarino, Reckase and Wooldridge, 2015). We model equation (4.1) as a two-level variance-component model:

$$Y_{isjt} = \alpha_o + Y_{isjt-1}\beta_1 + X_{isjt}\beta_2 + C_{isjt}\beta_3 + \zeta_{sjt} + \epsilon_{isjt},$$

(4.5)

where ζ_{sjt}, and ϵ_{isjt} are the error components assumed to have zero mean and to be mutually uncorrelated, so that their variances add up to the total variance. Specifically, ζ_{sjt} is a random intercept for teacher-classroom *j* at school *s*, and ϵ_{isjt} is an idiosyncratic component for student *i*. The level-2 variance σ_j^2 of the random intercept ζ_{sjt} is the between-classroom variance, and the level-1 variance σ_i^2 of the residuals ϵ_{isjt} can be interpreted as the between-student, within-classroom variance.

We implement a maximum likelihood estimation to identify equation (4.5). Then, we apply EB prediction to estimate the random intercepts ζ_{sjt} for individual classrooms, in other words, the classroom effects:

$$\hat{\zeta}_j^{EB} = \left(\frac{\hat{\sigma}_j^2}{\hat{\sigma_j}^2 + {\hat{\sigma}_i^2}/{n_i}} \right) \hat{\zeta}_j \tag{4.6}$$

In this process, the EB prediction is shrunk toward zero (the mean of the prior). As mentioned earlier, shrinkage is desirable in our application because it only affects clusters (classrooms) that provide little information, and it effectively reduces their influence, borrowing strength from other classrooms (Rabe-Hesketh and Skrondal, 2012).

Once we have calculated the individual classroom random effects (RE), we implement a simple OLS regression to explain it with our vector of teacher observable characteristics, τ_j:

$$\hat{\zeta}_j^{EB} = \beta_0 + \tau_j \beta_1 + u_j \tag{4.7}$$

4.5 Results

4.5.1 Random Assignment of Students to Teachers

We claim that our model specifications have the potential to lead to unbiased estimators of classroom effects, which are mainly driven by teacher quality differences across classrooms, because random assignment of students to teachers is prevalent in the first four years of primary school in Germany. In this section, we present evidence that shows no systematic matching of students to teachers within schools in our data.

In order to check for non-random assignment within schools, we regress individual teacher observable characteristics on our vector of student and family characteristics in grades 1 and 2. Our goal is to assess whether these covariates can systematically explain teachers' observable characteristics, including gender, experience, *Abitur* GPA, First and Second State Examination grades, whether she passed the Second State Examination, and the indicator of constructivist beliefs.[113] We use all observations and information available for first and second graders in the NEPS SC2 data. In addition, all regressions include school fixed effects.

Table 4.3 presents the regression results for children and their teachers' characteristics in grade 1. From the F-test of the joint significance of our vector of child and family characteristics, we observe that they cannot significantly explain teachers' observable characteristics, excepting only whether the teacher passed the Second State Examination. Teachers who passed the Second State Examination are more likely to be assigned to students with migration background, who are older, and who belong to families with higher ISEI. Nevertheless, the point estimates for the last two characteristics are close to zero, with migration background being the only relevant association. Children with a migration background might be expected to face greater academic challenges and, therefore, be more likely to be assigned to fully certified teachers.

[113] We exclusively include teacher characteristics observable before the beginning of the first school year (pre-treatment characteristics). Therefore, teacher exhaustion and parental evaluation are not taken into account.

Table 4.3: Associations of Teacher and Student Observable Characteristics for Grade 1

	Teacher						
	Gender	Experience	Abitur GPA	FSE grade	SSE passed	SSE grade	Constructivist beliefs
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Child demographics:							
Age in months	0.001	-0.025	0.002	0.000	0.003**	-0.002	0.000
	(0.001)	(0.037)	(0.002)	(0.002)	(0.001)	(0.002)	(0.003)
Female	0.013**	0.101	-0.011	-0.012	0.013	-0.012	0.011
	(0.006)	(0.253)	(0.014)	(0.012)	(0.009)	(0.014)	(0.021)
Migration background	0.003	0.369	0.021	-0.020	0.056***	0.011	-0.030
	(0.009)	(0.464)	(0.028)	(0.020)	(0.018)	(0.026)	(0.036)
Parental background:							
Years of education	0.000	0.061	-0.004	0.004	-0.003	-0.001	0.005
	(0.002)	(0.087)	(0.005)	(0.004)	(0.003)	(0.005)	(0.008)
ISEI	0.000	0.004	0.001*	-0.001*	0.001***	0.000	-0.001
	(0.000)	(0.011)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)
Siblings	0.004	-0.222	-0.016	0.003	-0.005	0.008	-0.007
	(0.003)	(0.195)	(0.011)	(0.008)	(0.007)	(0.011)	(0.014)
Constant	0.914***	9.096***	2.528***	2.697***	0.738***	3.173***	-0.424
	(0.072)	(3.216)	(0.160)	(0.159)	(0.125)	(0.170)	(0.287)
Number of students	3995	2655	2074	2181	2718	2073	3934
R^2	0.503	0.670	0.668	0.657	0.616	0.711	0.602
F	1.54	0.75	1.41	1.28	2.93	0.40	0.38
p	0.163	0.610	0.212	0.267	0.009	0.880	0.893

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* OLS regressions estimated with school fixed effects. Standard errors (in parentheses) are clustered at the school level. Total number of observation correspond to the full sample of students whose teachers provided the respective information on their characteristics. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 4.4 shows our random assignment test for the panel children in grade 2. We add competence scores in math, grammar and science from grade 1 in the vector of individual student characteristics. Remarkably, we find no significant association of baseline test scores with any of the teacher observable characteristics, which means that decisions on the assignment of students to teachers for grade 2 within schools are indeed not based on student ability. This is quite important for our study because in the following regression analyses, we use the subsample of students assigned to the same teacher in grades 1 and 2. Moreover, once again we observe that other student and family characteristics are not systematically correlated with teachers' observable characteristics, again with the only exception of whether the teacher passed the Second State Examination.

Our findings from this section confirm that, within schools, children in the first years of primary school in Germany are neither systematically matched to their teachers based on their ability, nor are they matched on other socio-economic characteristics other than migration background.

Table 4.4: Associations of Teacher and Student Observable Characteristics for Grade 2

	Teacher						
	Gender (1)	Experience (2)	Abitur GPA (3)	FSE grade (4)	SSE passed (5)	SSE grade (6)	Constructivist beliefs (7)
Child competences:							
Lagged Math	0.002	0.021	-0.007	-0.001	-0.002	-0.002	0.004
	(0.005)	(0.201)	(0.010)	(0.009)	(0.008)	(0.012)	(0.018)
Lagged Scientific	0.002	-0.064	-0.010	-0.006	-0.016	0.010	-0.009
	(0.007)	(0.271)	(0.013)	(0.012)	(0.010)	(0.015)	(0.028)
Lagged Grammar	-0.008*	-0.284	0.014	-0.005	0.009	-0.009	-0.005
	(0.004)	(0.238)	(0.013)	(0.011)	(0.009)	(0.011)	(0.019)
Child demographics:							
Age (months)	0.001	0.012	0.001	0.000	0.003**	-0.003	-0.003
	(0.001)	(0.044)	(0.002)	(0.002)	(0.001)	(0.002)	(0.004)
Female	0.019**	0.389	-0.021	-0.010	0.010	-0.005	0.010
	(0.008)	(0.295)	(0.016)	(0.014)	(0.011)	(0.017)	(0.028)
Migration background	0.014	0.731	0.058*	-0.014	0.061***	-0.001	-0.060
	(0.010)	(0.480)	(0.034)	(0.021)	(0.020)	(0.026)	(0.045)
Parental background:							
Years of education	-0.001	0.206*	-0.001	0.008*	-0.003	0.006	0.005
	(0.002)	(0.109)	(0.006)	(0.004)	(0.004)	(0.006)	(0.008)
ISEI	0.000	0.002	0.001	-0.001*	0.001**	-0.000	0.000
	(0.000)	(0.013)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)
Siblings	0.005	-0.272	-0.015	0.013*	-0.001	0.023**	-0.004
	(0.004)	(0.208)	(0.010)	(0.007)	(0.007)	(0.011)	(0.018)
Constant	0.930***	8.772*	1.578***	1.054***	0.674***	1.288***	0.138
	(0.110)	(4.572)	(0.201)	(0.179)	(0.151)	(0.223)	(0.434)
Number of students	2920	2485	1993	2091	2554	1998	2672
R^2	0.583	0.666	0.675	0.645	0.607	0.664	0.653
F	1.57	1.43	0.99	1.16	2.59	0.84	0.50
p	0.124	0.174	0.446	0.322	0.007	0.577	0.875

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* OLS regressions estimated with school fixed effects. Standard errors (in parentheses) are clustered at the school level. Total number of observation correspond to the full sample of students whose teachers provided the respective information on their characteristics. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

4.5.2 The Distribution of Classroom Effects

In table 4.5 we report the results of the first step specifications of our value-added model to student math competence in grade 2 of primary school, estimated with

classroom FE and RE. Column (1) presents results of the OLS regression described in equation (4.1). In column (2), we add the individual classroom FE to the original specification to assess changes in the explained variance of student math competence due to their inclusion. The adjusted R^2 increases from 0.485 to 0.525, which means that adding classroom FE into the model increases the explained variance by about four percentage points. The classroom FE are also jointly significant according to an F-test. Column (4) displays the results of our two-level variance-component model according to equation (4.5), and column (3) contains the results of an analogous model where the classroom random intercept is omitted. We observe that the inclusion of classroom RE in column (4) is statistically significant and also accounts for around four percentage points of the unexplained level-1 variance in column (3). Thus, the between-classroom variation is about four percent, which is very close to the variation found by Baumert et al. (2010). In addition, from the results displayed in columns (1), (2), (3) and (4), we can conclude that, aside from the classroom FE or RE, the variables that have more explanatory power for student math competence are baseline math, science and grammar competence scores, time between tests (months), gender, and to some extent, parents' ISEI. It is also remarkable that the point estimates of the covariates in the FE and RE specifications are very similar.

In table 4.6 we present the first step results of our classroom value-added to student language competence, estimated with classroom FE and RE. From the adjusted R^2 reported in columns (1) and (2), we observe that the inclusion of classroom FE increases the explained variance of the model by around six percentage points. Likewise, column (4) shows that around five percentage points of the unexplained variance of our RE specification in column (3) can be attributed to the classroom random intercept. In addition, the regression outputs in columns (1), (2), (3) and (4) indicate that the variables significantly and consistently associated with student language competence are prior grammar and math competence scores, time between tests (months), gender, migration background, and parents' years of education. Once again, it is remarkable that the point estimates of the covariates in the FE and RE specifications are very similar.

Table 4.5: Value-Added to Math Competence with and without Classroom Effects

	OLS (Fixed Effects)		HML (Random Effects)	
	(1)	(2)	(3)	(4)
Child competences:				
Lagged Math	0.520***	0.497***	0.520***	0.513***
	(0.025)	(0.026)	(0.023)	(0.023)
Lagged Scientific	0.195***	0.265***	0.195***	0.216***
	(0.032)	(0.034)	(0.031)	(0.031)
Lagged Grammar	0.160***	0.143***	0.160***	0.155***
	(0.028)	(0.029)	(0.025)	(0.025)
Child demographics:				
Age	0.003	0.001	0.003	0.002
	(0.005)	(0.005)	(0.005)	(0.005)
Female	-0.209***	-0.206***	-0.209***	-0.208***
	(0.039)	(0.040)	(0.039)	(0.038)
Migration background	0.062	0.105**	0.062	0.071
	(0.047)	(0.052)	(0.050)	(0.051)
Time between tests	0.111***	0.044	0.111***	0.109***
	(0.013)	(0.174)	(0.013)	(0.015)
Parental background:				
Years of education	0.007	0.018	0.007	0.010
	(0.012)	(0.012)	(0.012)	(0.012)
ISEI	0.003**	0.002	0.003**	0.003**
	(0.001)	(0.001)	(0.001)	(0.001)
Siblings	-0.024	-0.016	-0.024	-0.023
	(0.022)	(0.022)	(0.022)	(0.022)
Classroom:				
Class size	-0.002	-0.117	-0.002	-0.003
	(0.006)	(0.146)	(0.006)	(0.007)
Female proportion	-0.000	0.038	-0.000	-0.001
	(0.002)	(0.054)	(0.002)	(0.002)
Average ISEI	0.000	-0.027	0.000	-0.000
	(0.002)	(0.065)	(0.002)	(0.003)
Constant	-1.159**	1.548	-1.159**	-1.094**
	(0.541)	(3.661)	(0.533)	(0.556)
var(Classroom)				0.044***
				(0.012)
var(Residual)			0.645***	0.601***
			(0.021)	(0.021)
Federal State Effect	YES	YES	YES	YES
Classroom Effect	NO	YES	NO	YE
Number of students	1843	1843	1843	1843
Number of teachers/classrooms	251	251	251	251
R^2	0.493	0.592		
Adjusted R^2	0.485	0.525		

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* Columns (1) and (2) report coefficients from OLS regressions estimated without and with classroom effects, respectively. Columns (3) and (4) report coefficients from a hierarchical multilevel (mixed) model without and with classroom random effects, respectively. Standard errors (in parentheses) are clustered at the student level in the OLS regressions. *Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 4.6: Value-Added to Language Competence with and without Classroom Effects

	OLS (Fixed Effects)		HML (Random Effects)	
	(1)	(2)	(3)	(4)
Child competences:				
Lagged Grammar	0.211***	0.221***	0.211***	0.215***
	(0.031)	(0.032)	(0.028)	(0.028)
Lagged Math	0.258***	0.217***	0.258***	0.245***
	(0.026)	(0.028)	(0.025)	(0.025)
Lagged Scientific	0.007	0.031	0.007	0.013
	(0.034)	(0.036)	(0.034)	(0.034)
Child demographics:				
Age	0.006	0.004	0.006	0.005
	(0.005)	(0.005)	(0.005)	(0.005)
Female	0.190***	0.168***	0.190***	0.181***
	(0.043)	(0.044)	(0.043)	(0.042)
Migration background	0.155***	0.154***	0.155***	0.153***
	(0.054)	(0.059)	(0.055)	(0.056)
Time between tests	0.071***	0.128	0.071***	0.072***
	(0.015)	(0.154)	(0.014)	(0.017)
Parental background:				
Years of education	0.036***	0.037***	0.036***	0.036***
	(0.013)	(0.013)	(0.013)	(0.013)
ISEI	0.003*	0.003	0.003*	0.003*
	(0.002)	(0.002)	(0.002)	(0.002)
Siblings	-0.007	-0.012	-0.007	-0.007
	(0.026)	(0.027)	(0.025)	(0.025)
Classroom:				
Class size	0.003	0.322***	0.003	0.003
	(0.007)	(0.117)	(0.007)	(0.008)
Female proportion	0.002	-0.112***	0.002	0.002
	(0.002)	(0.040)	(0.002)	(0.003)
Average ISEI	-0.005**	0.093**	-0.005**	-0.005
	(0.003)	(0.041)	(0.003)	(0.003)
Constant	-1.829***	-7.803**	-1.829***	-1.786***
	(0.561)	(3.232)	(0.589)	(0.619)
var(Classroom)				0.054***
				(0.014)
var(Residual)			0.735***	0.681***
			(0.025)	(0.025)
Federal State Effect	YES	YES	YES	YES
Classroom Effect	NO	YES	NO	YE
Number of students	1753	1753	1753	1753
Number of teachers/classroom.	240	240	240	240
R^2	0.263	0.411		
Adjusted R^2	0.251	0.314		

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* Columns (1) and (2) report coefficients from OLS regressions estimated without and with classroom fixed effects, respectively. Columns (3) and (4) report coefficients from a hierarchical multilevel (mixed) model without and with classroom random effects, respectively. Standard errors (in parentheses) are clustered at the student level in the OLS regressions. *Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table 4.7: Estimates of Classroom Effects on Math Competence

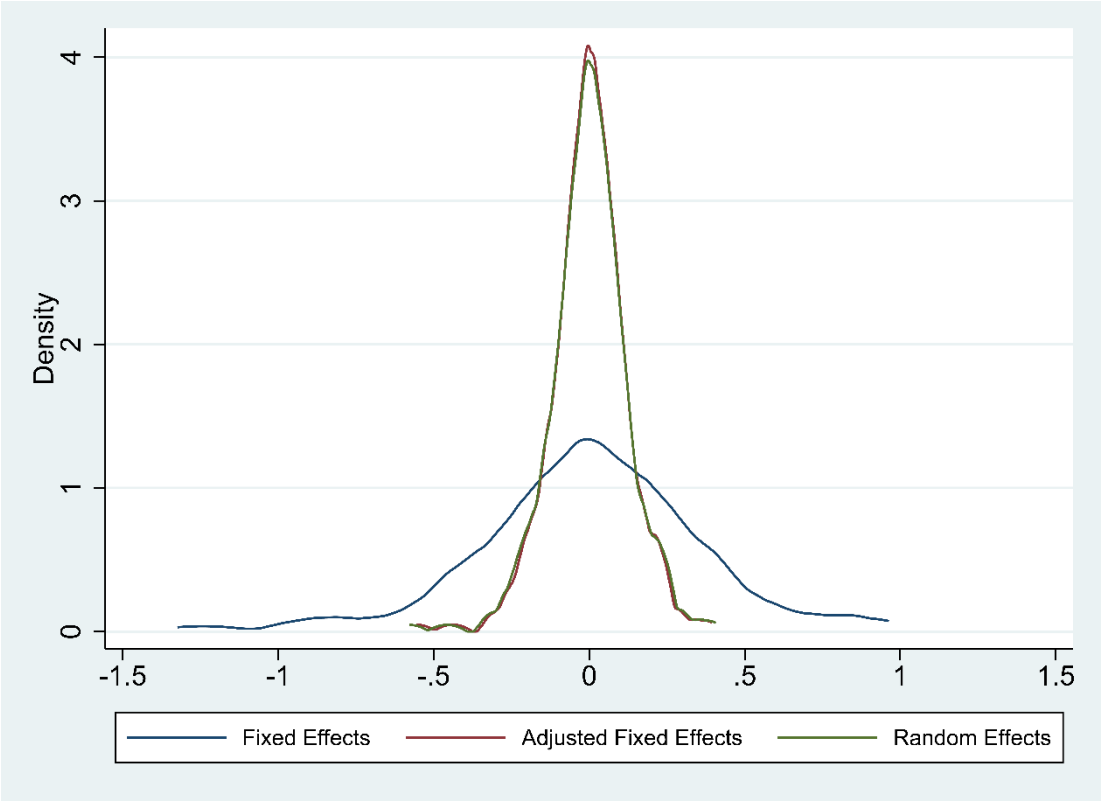
	(1)	(2)	(3)	(4)	(5)
Classroom Fixed Effects (FE):					
Standard deviation	0.364	0.362	0.360	0.360	0.360
Adjusted EB standard deviation	0.119	0.121	0.120	0.120	0.120
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Classroom Random Effects (RE):					
EB Standard deviation	0.122	0.124	0.124	0.124	0.124
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Included covariates:					
Federal State effects	YES	YES	YES	YES	YES
Lagged test scores	YES	YES	YES	YES	YES
Student characteristics	NO	YES	YES	YES	YES
Parental background	NO	NO	YES	YES	YES
Classroom size	NO	NO	NO	YES	YES
Classroom averages	NO	NO	NO	NO	YES
Number of teachers/classrooms	251	251	251	251	251
Number of students threshold	5	5	5	5	5

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This table presents standard deviations of distributions of classroom effects on math competence estimated as FE using OLS regressions, and RE using hierarchical multilevel (mixed) regressions. Empirical Bayes (EB) shrinkage was implemented to obtain the adjusted FE standard deviation. All results are based on regressions of math competence test scores in grade 2 on lagged math, grammar and science competence test scores, time between tests and federal state fixed effects. Columns (2)-(5) control for the following student characteristics: age, gender, migration background; for parental background: highest years of education, highest ISEI, number of siblings; and for classroom averages: proportion of females, average ISEI.

In table 4.7, we present the distribution of the classroom effects on student math competence estimated as adjusted classroom FE following equations (4.1), (4.2) and (4.3), and as classroom RE according to equations (4.5) and (4.6). The classroom effects can also be interpreted as indicators of classroom quality in terms of the individual classroom contribution to student competence development. Individual teacher effects, as explained earlier, mainly drive our classroom effects' estimates. Column (1) reports the standard deviations of the distributions of classroom effects estimated with adjusted FE and RE after controlling only for federal state effects, lagged competence scores and time between tests in equations (4.1) and (4.5), respectively. The adjusted FE specification shows that one standard deviation change in classroom quality is associated with a 0.119 standard deviation higher student math competence score. The RE specification estimates a slightly higher standard deviation of 0.122. In Column (5), we present our estimations after controlling for a full set of child characteristics, family socio-economical background, classroom size, and additional classroom averages. Results remain practically the same, with one standard deviation change in classroom

quality associated with a 0.120 standard deviation higher student math competence score in the classroom adjusted FE specification, and with a 0.124 standard deviation higher score in the classroom RE specification. The estimated distributions of classroom effects on math competence are also presented in figure 4.1. Notably, the adjusted FE and RE distributions practically overlap.

Figure 4.1: Distribution of Classroom Effects on Math Competence



Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This figure presents kernel density distributions of classroom effects on students' math competence development, estimated as FE using OLS regressions, and RE using hierarchical multilevel (mixed) regressions. Empirical Bayes (EB) shrinkage was implemented to obtain the adjusted FE distribution. All distributions are based on regressions of math competence test scores in grade 2 on lagged math, grammar and science competence test scores, time between tests, child age, gender, migration background and number of siblings, parents' highest years of education and highest ISEI, classroom averages and federal state fixed effects.

We present our estimates of the distribution of the classroom effects on language competence in table 4.8, following the same structure of table 4.7. The standard deviation of the classroom quality distribution estimated with the adjusted FE specification ranges from 0.149 in column (1) to 0.142 in column (5), when a full set to

controls are introduced in equation (4.1). The standard deviation of the classroom quality distribution estimated with RE is virtually the same. It decreases from 0.148 in Column (1) to 0.140 in Column (5), when a full set of controls are taken into account in equation (4.5). Accordingly, we can conclude that a one standard deviation increase in classroom quality is associated with about a 0.140 standard deviation higher student language competence score. Figure 4.2 displays the distributions of the classroom adjusted FE and RE on language competence.

Table 4.8: Estimates of Classroom Effects on Language Competence

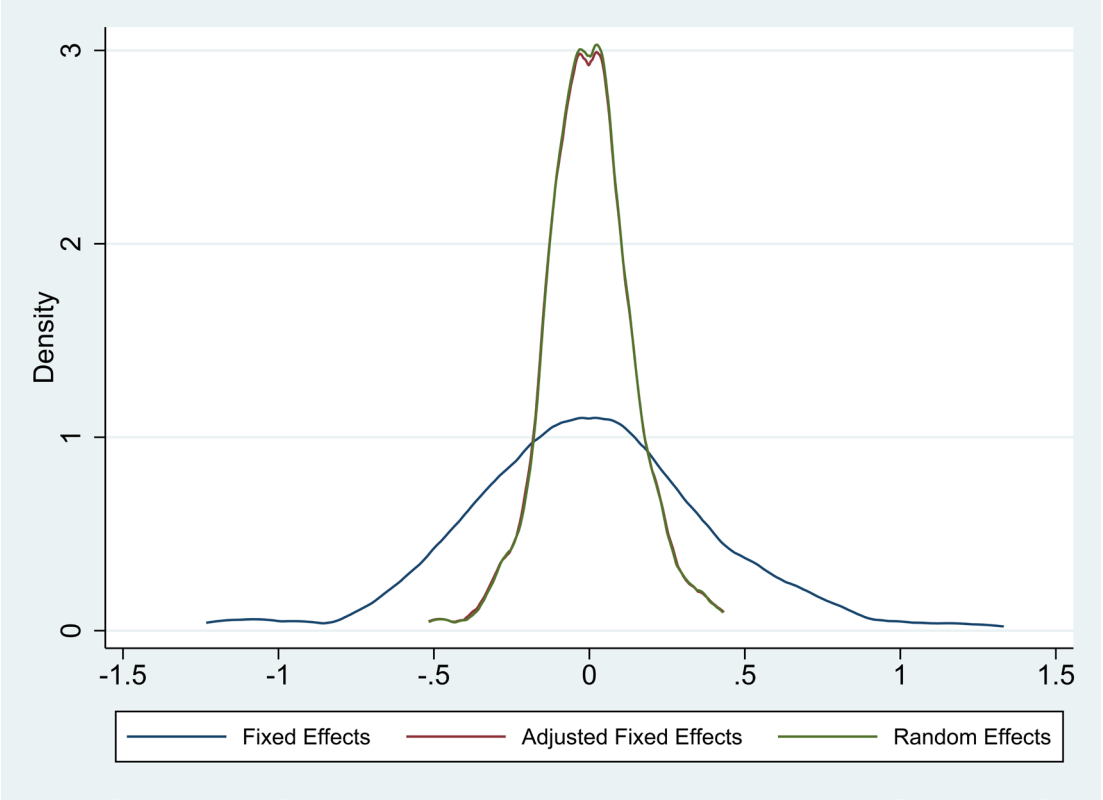
	(1)	(2)	(3)	(4)	(5)
Classroom Fixed Effects (FE):					
Standard deviation	0.403	0.398	0.397	0.397	0.396
Adjusted EB standard deviation	0.149	0.142	0.146	0.145	0.142
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Classroom Random Effects (FE):					
EB Standard deviation	0.148	0.141	0.145	0.145	0.140
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Included covariates:					
Federal State effects	YES	YES	YES	YES	YES
Lagged test scores	YES	YES	YES	YES	YES
Student characteristics	NO	YES	YES	YES	YES
Family background	NO	NO	YES	YES	YES
Classroom size	NO	NO	NO	YES	YES
Classroom averages	NO	NO	NO	NO	YES
Number of teachers/classrooms	240	240	240	240	240
Number of students threshold	5	5	5	5	5

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This table presents standard deviations of distributions of classroom effects on language competence estimated as FE using OLS regressions, and RE using hierarchical multilevel (mixed) regressions. Empirical Bayes (EB) shrinkage was implemented to obtain the adjusted FE standard deviation. All results are based on regressions of early reading competence test scores in grade 2 on lagged math, grammar and science competence test scores, time between tests and federal state fixed effects. Columns (2)-(5) control for the following student characteristics: age, gender, migration background; for parental background: highest years of education, highest ISEI, number of siblings; and for classroom averages: proportion of females, average ISEI.

Interestingly, our classroom adjusted FE and RE's estimates are practically the same, and do not change with the inclusion of additional controls once we have taken into account lagged competence scores and time between tests. This aligns with previous research in the US, which has found that controlling for lagged test scores is key to obtaining unbiased value-added estimates, since most of the sorting of students to teachers relevant for future achievement is captured by them (Chetty, Friedman and

Rockoff, 2014a). In addition, our estimates are comparable in size to classroom effects estimated for primary school in the US. With respect to the distributions of classroom value-added obtained by Chetty, Friedman and Rockoff (2014a), our standard deviations are smaller for math (0.166 SD) and greater for language (0.117 SD). Thus, in terms of student competence development, the quality differences among teachers and their classrooms in Germany are smaller for math and larger for language.

Figure 4.2: Distribution of Classroom Effects on Language Competence



Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This figure presents kernel density distributions of classroom effects on students' language competence development, estimated as FE using OLS regressions, and RE using hierarchical multilevel (mixed) regressions. Empirical Bayes (EB) shrinkage was implemented to obtain the adjusted FE distribution. All distributions are based on regressions of early reading competence test scores in grade 2 on lagged math, grammar and science competence test scores, time between tests, child age, gender, migration background and number of siblings, parents' highest years of education and highest ISEI, classroom averages and federal state fixed effects.

Our estimates of the adjusted classroom FE and RE can be also used to build quality rankings of classrooms based on their individual contribution to competence development. We present classroom rankings of predicted value-added to student math

competence derived from their individual adjusted FE and RE in figure A 4.1 of the Appendix. We also display rankings of classroom predicted value-added to student language competence derived from the adjusted FE and the RE estimations in figure A 4.2 of the Appendix. Even though the individual classroom contributions are nosily predicted because of our small student sample size, it is clear that some classrooms, and primarily their teachers, significantly outperform or underperform compared to the average classroom's contribution to learning.

In addition, it is relevant that our classroom effect estimations assume that one teacher is responsible for teaching all main subjects in the classroom, because the NEPS SC2 data provides information on classroom teachers for the primary school grades, as opposed to subject teachers.[114] Indeed, having classroom teachers in primary school is common practice in Germany, and consistent with teaching careers at the primary school level (KMK, 2019). In this context, we also estimate the correlation between the math and language classroom effects for the adjusted FE and RE specifications. We find a positive correlation of 0.208 for the adjusted FE specification and of 0.205 for the RE specification[115], which suggests that higher math value-added classrooms also tend to be higher language value-added classrooms. In addition, as a robustness check, we rerun our analysis for subsamples of teachers who explicitly declared they were responsible for math and/or language instruction in grade 2 in the NEPS SC2 data. Results are very similar and presented in the Appendix of this Chapter.

4.5.3 Explaining Classroom Effects with Teacher Characteristics

In this subsection, we report regression results of the association between the estimated classroom effects on student competence scores and observed teacher characteristics, according to equation (4.4) for the adjusted FE specification and equation (4.7) for the RE specification.

[114] NEPS SC2 classroom data is not divided into math and language classrooms as it is done in NEPS SC3 for grade 5 and up.

[115] These results are based on classrooms for which we are able to estimate both math and language effects. The sample size decreases to 234 classrooms.

Our value-added estimations are based on all classrooms linked to teacher unique identifiers, regardless of whether a teacher has answered specific questions on her characteristics in the NEPS surveys. Accordingly, from the original math sample of 251 teachers, and language sample of 240 teachers, we have full information on the characteristics of 147 and 141 teachers, respectively. Thus, the reduction in the teacher-classroom sample size could pose a concern of sample selectivity and representability of the results. In order to test whether there is an association between teacher willingness to disclose professional information and our classroom effects' estimations, we calculate an index based on the number of questions answered by each teacher. Then, we calculate its correlation with our adjusted classroom FE and RE estimations. We found correlations virtually equal to zero for both estimations in math and language.[116] Therefore, we conclude that our reduced sample is not positively or negatively selected with respect to teacher quality.

Table 4.9 presents results for math and language competence. Column (1) shows the association of teacher characteristics with the classroom effects on math competence development estimated with adjusted FE and column (2) with RE. As reported by previous research, our rich set of teacher covariates explain very little of the variance of the classroom effects on math competence, just about five percent in both model specifications. Moreover, we identify no significant correlation.[117]

[116] Results available upon request.

[117] Alternatively, we investigate the association between teacher characteristics and student math competence development by directly introducing these observables into equation (4.1) and estimating an OLS regression with clustered standard errors at the classroom level. We find a negative correlation with teacher female gender, which is significant only at the 10 percent level. All the other teacher characteristics remain uncorrelated with math competence development. We argue that the indirect associations with the classroom effects are more robust, because the number of students per teacher might influence the statistical significance of the direct associations estimated with OLS. Results available upon request.

Table 4.9: Association of Teacher Characteristics and Classroom Effects on Math and Language Competence

	Math		Language	
Teacher	EB Adjusted Fixed Effect (1)	EB Random Effect (2)	EB Adjusted Fixed Effect (3)	EB Random Effect (4)
Female	-0.067	-0.068	0.010	0.008
	(0.041)	(0.042)	(0.044)	(0.043)
Years of experience	0.001	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Abitur GPA	-0.012	-0.012	0.025*	0.025*
	(0.011)	(0.012)	(0.014)	(0.014)
FSE Grade	-0.001	-0.002	-0.030*	-0.030*
	(0.010)	(0.011)	(0.018)	(0.018)
SSE Passed	-0.006	-0.005	0.023	0.023
	(0.026)	(0.026)	(0.039)	(0.039)
Migration background	0.036	0.037	-0.022	-0.020
	(0.033)	(0.034)	(0.058)	(0.057)
Constructivist beliefs	0.010	0.010	0.017	0.017
	(0.012)	(0.013)	(0.011)	(0.011)
Exhaustion	-0.004	-0.005	-0.012	-0.012
	(0.009)	(0.010)	(0.013)	(0.013)
Parental evaluation	-0.001	-0.001	0.026**	0.025**
	(0.010)	(0.010)	(0.012)	(0.012)
Constant	0.044	0.043	-0.058	-0.055
	(0.050)	(0.051)	(0.061)	(0.061)
Number of teacher with observables	147	147	141	141
R^2	0.049	0.049	0.102	0.102

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This table presents OLS regressions of classroom value-added to math and language competence on teacher characteristics. Self-reported *Abitur* GPA and First State Examination (FSE) grade originally were on a scale that went from 1.0 to 4.0, with 1.0 being the best possible grade and 4.0 the minimum passing grade. These self-reported grades were standardized to have zero mean and a one unit standard deviation with respect to the full sample of teachers. Standard errors in parentheses. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Columns (3) and (4) of table 4.9 present the association of teacher observables with classroom effects on language competence development estimated as classroom adjusted FE and RE, respectively. For language competence, the observable teacher characteristics explain about 10 percent of the variance in the classroom effects in both specifications. Surprisingly, only average parental evaluation of teacher quality is significantly associated with classroom effects on language competence at the 5 percent significance level. One standard deviation increase in the average parental evaluation is associated with a 0.026 standard deviation higher student language competence score

in the adjusted FE specification, and with a 0.025 standard deviation higher score in the RE specification. This association aligns with previous experimental value-added research conducted in primary schools (Araujo *et al.*, 2016). In addition, we find marginal associations with Abitur GPA and the First State Examination grade in both specifications at the 10 percent significance level.[118]

4.5.4 Heterogeneity by Teacher Gender

In our analysis samples, more than 90 percent of teachers are female. Given that we probably have a highly selective group of male teachers, we replicate our entire analysis exclusively for the female sample.

Mirroring the structure of table 4.7, in table 4.10 we report the distribution of classroom quality, or effects on student math competence estimated as adjusted FE and RE for the teacher female sample. The adjusted FE specification produces a standard deviation of 0.107 in the classroom effects distribution, which does not change when a full set of controls are introduced in Column (5). The standard deviation estimated with RE is exactly the same, but slightly increases from 0.107 in Column (1) to 0.108 in Column (5). Thus, we observe that the variance of the classroom quality distribution is smaller for the teacher female sample. One standard deviation increase in classroom quality is associated with at least a 0.107 standard deviation higher student math competence score.

[118] We also estimated the direct association between teacher characteristics and student language competence development in equation (4.1) using an OLS regression with clustered standard errors at the classroom level. We find a stronger association with average parental evaluation of teacher quality, which is significant at the 1 percent level. The associations with Abitur GPA and the First State Examination grade have the same direction, but only Abitur is significant at the 10 percent level. A positive association with the constructivist beliefs index becomes significant, but only at the 10 percent level. We argue that the indirect associations with the classroom effects are more robust, since the number of students per teacher might influence the statistical significance of the direct associations estimated with OLS. Results available upon request.

Table 4.10: Estimates of Classroom Effects on Math Competence, Female Teacher Sample

	(1)	(2)	(3)	(4)	(5)
Classroom Fixed Effects (FE):					
Standard deviation	0.357	0.355	0.354	0.354	0.354
Adjusted EB standard deviation	0.107	0.109	0.108	0.107	0.107
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Classroom Random Effects (RE):					
EB Standard deviation	0.107	0.109	0.108	0.108	0.108
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Included covariate:					
Federal State effects	YES	YES	YES	YES	YES
Lagged test scores	YES	YES	YES	YES	YES
Student characteristics	NO	YES	YES	YES	YES
Family background	NO	NO	YES	YES	YES
Classroom size	NO	NO	NO	YES	YES
Classroom averages	NO	NO	NO	NO	YES
Number of teachers/classrooms	233	233	233	233	233
Number of students threshold	5	5	5	5	5

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This table presents standard deviations of female classroom effects on math competence distributions estimated as FE using OLS regressions, and RE using hierarchical multilevel (mixed) regressions. Empirical Bayes (EB) shrinkage was implemented to obtain the adjusted FE standard deviation. All results are based on regressions of math competence test scores in grade 2 on lagged math, grammar and science competence test scores, time between tests and federal state fixed effects. Columns (2)-(5) control for the following student characteristics: age, gender, migration background; for parental background: highest years of education, highest ISEI, number of siblings; and for classroom averages: proportion of females, average ISEI.

Table 4.11 presents the distribution of classroom effects on student language competence, following the same structure as table 4.7. Once again, we find that the variance of the classroom quality distribution is smaller for the female sample. The standard deviations of both the classroom adjusted FE, and the RE specifications, range from 0.134 in Column (1) to 0.128 in Column (5), when a full set of controls are introduced. Accordingly, we observe that one standard deviation increase in classroom quality is associated with at least a 0.128 standard deviation higher student language competence score.

Table 4.11: Estimates of Classroom Effects on Language Competence, Female Teacher Sample

	(1)	(2)	(3)	(4)	(5)
Classroom Fixed Effects (FE):					
Standard deviation	0.393	0.389	0.387	0.387	0.385
Adjusted EB standard deviation	0.134	0.129	0.132	0.132	0.128
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Classroom Random Effects (RE):					
EB Standard deviation	0.134	0.128	0.132	0.132	0.128
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Included covariate:					
Federal State effects	YES	YES	YES	YES	YES
Lagged test scores	YES	YES	YES	YES	YES
Student characteristics	NO	YES	YES	YES	YES
Family background	NO	NO	YES	YES	YES
Classroom size	NO	NO	NO	YES	YES
Classroom averages	NO	NO	NO	NO	YES
Number of teachers/classrooms	223	223	223	223	223
Number of students threshold	5	5	5	5	5

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This table presents standard deviations of female classroom effects on language competence distributions estimated as FE using OLS regressions, and RE using hierarchical multilevel (mixed) regressions. Empirical Bayes (EB) shrinkage was implemented to obtain the adjusted FE standard deviation. All results are based on regressions of early reading competence test scores in grade 2 on lagged math, grammar and science competence test scores, time between tests and federal state fixed effects. Columns (2)-(5) control for the following student characteristics: age, gender, migration background; for parental background: highest years of education, highest ISEI, number of siblings; and for classroom averages: proportion of females, average ISEI.

We also look at the association between classroom effects on student math and language competence scores and observed teacher characteristics in the female teacher sample. Results are reported in table 4.12 using the same structure as table 4.9. Similar to the findings in the full teacher sample, column (1) and column (2) show that there is no significant association between any of the individual teacher characteristics and the classroom effects on math competence, estimated either as adjusted FE or as RE for the female sample.[119] Likewise, parental evaluation of teacher quality is the only teacher characteristic significantly and positively associated with classroom value-added to language competence in grade 2, estimated as adjusted FE in column (3), or as RE in column (4) for the female sample. The size of the association is virtually the same as

[119] We find the same results in our alternative specification, which directly introduces teacher characteristics into equation (4.1) and estimates their association with student math competence development using an OLS regression with standard errors clustered at the classroom level. Results available upon request.

that of the full teacher sample, a 0.024 standard deviation higher language competence score in both specifications. Moreover, the previous marginal associations with *Abitur* GPA or First State Examination grade become insignificant, which suggest that they were probably driven by the male sample.[120]

Table 4.12: Association of Teacher Characteristics and Classroom Effects on Math and Language Competence, Female Teacher Sample

	Math		Language	
Teacher	EB Adjusted Fixed Effect (1)	EB Random Effect (2)	EB Adjusted Fixed Effect (3)	EB Random Effect (4)
Years of experience	0.001	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Abitur GPA	-0.012	-0.012	0.015	0.014
	(0.010)	(0.011)	(0.013)	(0.013)
FSE Grade	-0.003	-0.003	-0.027	-0.027
	(0.010)	(0.010)	(0.018)	(0.018)
SSE Passed	-0.014	-0.014	0.033	0.033
	(0.024)	(0.024)	(0.039)	(0.039)
Migration background	0.039	0.039	-0.019	-0.018
	(0.032)	(0.032)	(0.053)	(0.053)
Constructivist beliefs	0.007	0.007	0.015	0.015
	(0.011)	(0.012)	(0.011)	(0.011)
Exhaustion	-0.007	-0.007	-0.008	-0.008
	(0.009)	(0.009)	(0.013)	(0.013)
Parental evaluation	-0.001	-0.001	0.024**	0.024**
	(0.009)	(0.009)	(0.011)	(0.011)
Constant	-0.018	-0.019	-0.062	-0.059
	(0.029)	(0.030)	(0.044)	(0.044)
Number of teacher with observables	135	135	129	129
R^2	0.046	0.046	0.097	0.096

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This table presents OLS regressions of female classroom value-added to math and language competence on teacher characteristics. Self-reported *Abitur* GPA and First State Examination (FSE) grades originally were on scale from 1.0 to 4.0, with 1.0 being the best possible grade and 4.0 the minimum passing grade. These self-reported grades were standardized to have zero mean and a one-unit standard deviation with respect to the full sample of teachers. Standard errors in parentheses. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

[120] When estimating the direct association between teacher characteristics and student language competence development in equation (4.1) using an OLS regression with standard errors clustered at the classroom level, we also find a significant association with average parental evaluation of teacher quality at the 1 percent level. No other teacher characteristic is significantly correlated to language competence development. Results available upon request.

4.5.5 Parental Behavioral Response

Our results show that German primary school parents give higher evaluation scores to teachers who exhibit higher classroom effects on language competence development; in other words, our results suggest that parents can identify better teachers for language competence development. In the value-added model framework, simultaneous parental behavioral responses to augment or offset the effect of being assigned to a better or worst teacher are included in the classroom effect estimations, as stressed by Todd and Wolpin (2003). Accordingly, in this section, we analyze these parental responses.

The NEPS SC2 survey asked parents how much time they spend helping their children with their homework and other school exercises in a typical school week, and whether their children have received private tutoring in grade 2. Using the student language sample,[121] we analyze the association of these measures of parental behavioral response with the individual parental evaluation of teacher quality in grade 2.[122]About 94 percent of the parents in the student language sample provided detailed answers on time spent helping with homework in hours and minutes. In addition, virtually all parents provided information on whether their children had private tutoring; nonetheless, only around a three percent gave an affirmative answer.

Table 4.13 reports regression results for parental time spent helping with homework, in column (1) accounting only for federal state fixed effects as control, and in column (2) including a full set of child, family, and classroom controls including child lagged competence test scores. We find that higher parental perception of teacher quality is associated with less time spent helping with homework; however, these associations are statistically insignificant. Columns (3) and (4) present regression results for whether the child had private tutoring as a dependent variable, respectively without and with a full set of controls. Results show negative and significant

[121] We also run our analysis using the math student sample and the results (available upon request) are virtually the same.

[122] As indicated in section 3, this indicator comes from parents' assessment on whether school teachers tried to meet children's needs on a 4-point Likert scale. The Likert scale has the following categories: 1. Does not apply, 2. Does rather not apply, 3. Does rather apply and 4. Does apply. Due to a small number of observations in categories 1. and 2., we combine them into one category "Does not/rather not apply".

associations with parental perception of teacher quality. A child is about eight percent less likely to receive private tutoring if the parental evaluation of teacher quality corresponds to the highest possible category, as displayed in column (4). Nonetheless, as mentioned, a very small and probably highly selective percentage of children receive private tutoring in our sample. Finally, columns (5) and (6) display regression results on whether the child had private tutoring specifically in the subject of German language,[123] respectively without and with a full set of controls. Once again, we observe that the higher the parental evaluation of teacher quality, the lower the probability of receiving private language tutoring for a child, about seven percent significantly lower for the highest evaluation of teacher quality. However, the percentage of children expose to private language tutoring is even smaller, at about one percent of the student language sample.

Table 4.13: Parental Evaluation of Teacher Quality and Behavioral Responses

	Time Helping with Homework (h)		Private Tutoring		Private Tutoring (German)	
	(1)	(2)	(3)	(4)	(5)	(6)
Teacher meets child's needs:						
Does rather apply	-0.369	-0.181	-0.083**	-0.076**	-0.064**	-0.061**
	(0.308)	(0.294)	(0.036)	(0.035)	(0.031)	(0.031)
Does apply	-0.221	-0.086	-0.089**	-0.082**	-0.073*	-0.071**
	(0.299)	(0.283)	(0.035)	(0.035)	(0.031)	(0.031)
Included covariates:						
Federal State effects	YES	YES	YES	YES	YES	YES
Lagged test scores	NO	YES	NO	YES	NO	YES
Student characteristics	NO	YES	NO	YES	NO	YES
Family background	NO	YES	NO	YES	NO	YES
Classroom size	NO	YES	NO	YES	NO	YES
Classroom averages	NO	YES	NO	YES	NO	YES
N	1652	1652	1752	1752	1752	1752
R^2	0.015	0.049	0.026	0.052	0.037	0.049

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* Column (1) and (2) results are based on OLS regressions of parental time spent helping with children's homework in a typical school week (hours) on parental evaluation of teacher quality measured on a 3-point Likert scale (base category: "Does not/rather not apply") in grade 2. Column (3) and (4) results are based on OLS regressions of whether the child receives private tutoring on parental evaluation of teacher quality in grade 2. Column (5) and 6) results are based on OLS regressions of whether the child receives private tutoring for language (German) on parental evaluation of teacher quality in grade 2. Columns (2), (4), and (6) control for the following student characteristics: lagged math, language and science competence, age, gender, migration background, parental background, highest years of education, highest ISEI, number of siblings, classroom averages, proportion of females, average ISEI. Standard errors (in parentheses) clustered at the individual level. Total number of observations corresponds to valid parental answers to the dependent variables in the language student sample. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

[123]Among the topics covered in language private tutoring are: reading and understanding texts, speaking and oral comprehension, spelling and writing,

Our results suggest, on the one hand, that parents do not generally respond to a perceived higher teacher quality by spending significantly less time helping their children with their homework. On the other hand, parents seem to decrease their investment in private tutoring when teacher quality is perceived as higher, but this seems to affect a highly selective sample of students.

4.6 Conclusion

In this paper, we have provided the first estimations of substantial classroom quality differences on competence development in German primary schools. One standard deviation increase in classroom quality is associated with at least 12 percent of a standard deviation increase in student mathematical competence score, and at least 14 percent of a standard deviation increase in language competence score. These classroom effects are driven by unbiased teacher quality differences across classrooms, since our estimations take into account student baseline competence scores and peer effects, and there is no systematic matching of students to teachers based on ability and other socio-economic factors in the German primary school.

We have managed to build a short panel of teachers and their students that covers math and language competence development between grades 1 and 2 using data from the SC2 of the NEPS. However, we observe only one classroom per teacher, and therefore, have not been able to estimate persistent teacher effects. Nonetheless, based on previous empirical research on classroom and teacher effects conducted in primary schools (Chetty, Friedman and Rockoff, 2014a; Araujo *et al.*, 2016), we are confident that persistent teacher effects are at most one to two percentage points smaller than our classroom effects' estimates.

Our research has also confirmed that easily quantifiable teacher characteristics explain very little of the variance of the classroom effects on math and language competence in the German primary school. Moreover, we have found no association between our estimates of classroom quality and most of the teacher characteristics analyzed, including gender, years of teaching experience, migration background, self-reported *Abitur* GPA, self-reported First State Examination grade, whether the teacher has passed the Second State Examination, constructivist beliefs and exhaustion levels.

Remarkably, our indicator of parental evaluation of teacher quality is the only covariate that is significantly and positively associated with our estimates of classroom effects on language competence development. This result suggests that parents can identify effective teachers in the first years of primary school. In addition, we find that a selective group of parents exhibits behavioral responses to differences in perceived teacher and classroom quality.

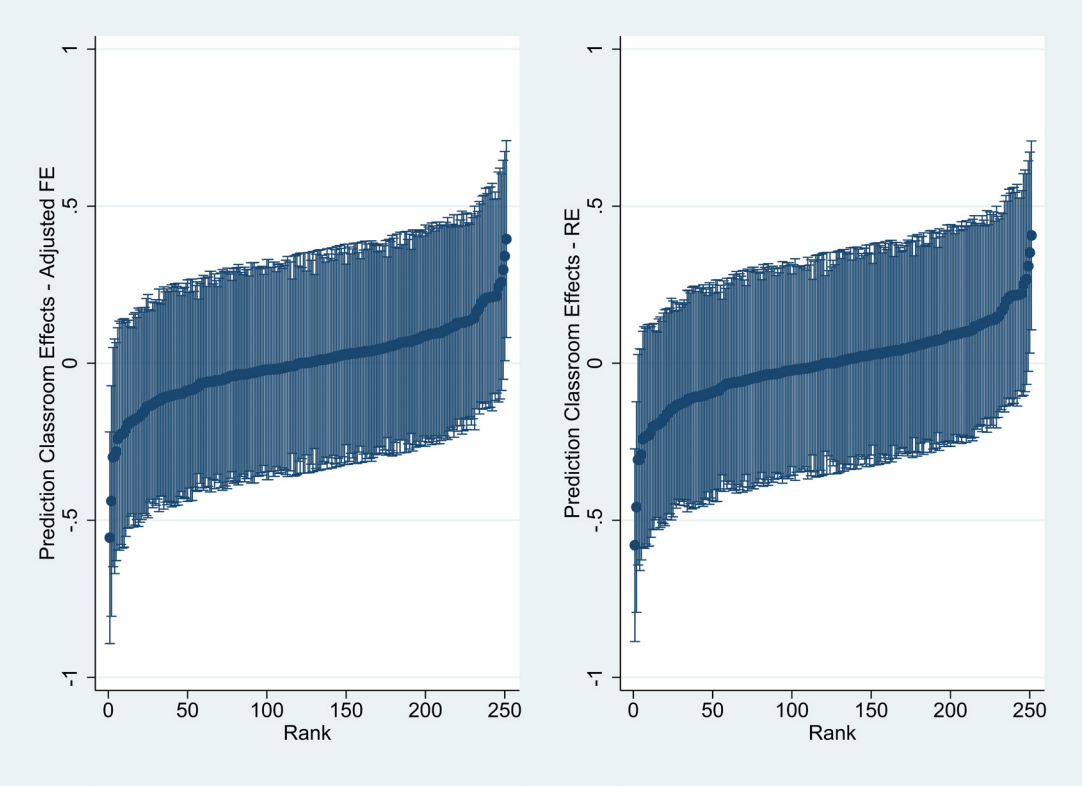
In the last 20 years, research in the US and around the world has consistently found that teacher value-added is an educationally and economically meaningful measure. This study is the first step toward the estimation of persistent teacher effects and their determinants in German primary schools. The implementation of a national panel study of teachers is urgently needed for the development of future research and policy applications.

We conclude with some policy recommendations. Our research suggests that policy makers should consider teacher and classroom value-added measures as powerful tools for evaluating and improving teacher workforce quality in Germany, given that the observable characteristics typically used in teacher recruitment processes explain very little of the variance in teacher effectiveness. Quality rankings of teachers and their classrooms, based on their individual contribution to competence development, could be used to incentivize top performers, or to identify and dismiss teachers who are permanently at the bottom of the quality distribution. Moreover, we present evidence that the inclusion of parent perspectives in teacher quality assessments is meaningful and worth of consideration in primary schools.

4.7 Appendix

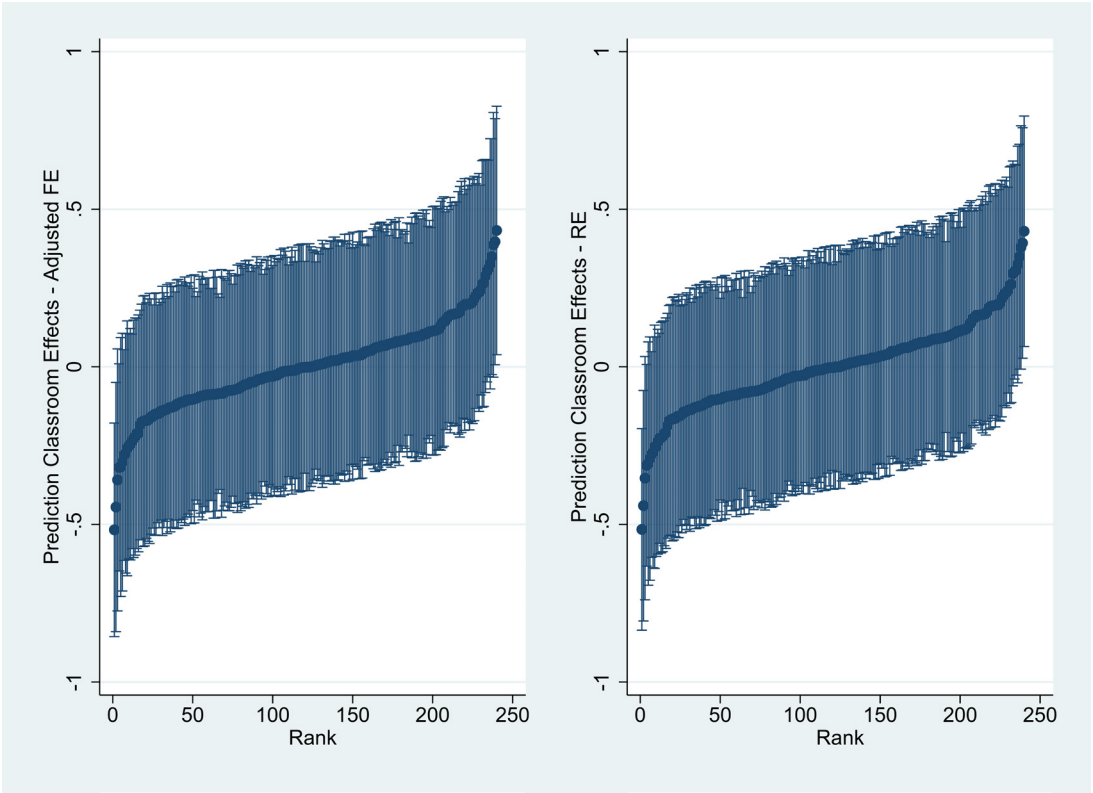
4.7.1 Figures

Figure A 4.1: Classroom Ranking by Effects on Math (Adjusted Fixed and Random Effects)



Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This figure presents classroom rankings of predicted individual effects on students' math competence development, estimated as FE using OLS regressions, and RE using hierarchical multilevel (mixed) regressions. Empirical Bayes (EB) shrinkage was implemented to obtain the adjusted FE. All distributions are based on regressions of math competence test scores in grade 2 on lagged math, grammar and science competence test scores, time between tests, child age, gender, migration background and number of siblings, parents' highest years of education and highest ISEI, classroom averages and federal state fixed effects.

Figure A 4.2: Classroom Ranking by Effects on Language (Adjusted Fixed and Random Effects)



Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This figure presents classroom rankings of predicted individual effects on students' language competence development, estimated as FE using OLS regressions, and RE using hierarchical multilevel (mixed) regressions. Empirical Bayes (EB) shrinkage was implemented to obtain the adjusted FE. All distributions are based on regressions of early reading competence test scores in grade 2 on lagged math, grammar and science competence test scores, time between tests, child age, gender, migration background and number of siblings, parents' highest years of education and highest ISEI, classroom averages and federal state fixed effects.

4.7.2 Robustness Check

The NEPS SC2 data provides information on classroom teachers for the primary school grades assuming that one teacher is responsible for teaching all main subjects in the classroom, which is common practice in the German school system and consistent with teaching careers at the primary school level (KMK, 2019). Nonetheless, after a closer look at the classroom questionnaires, we found a subsample of teachers who explicitly declared to be responsible for math and/or language instruction in grade 2. Under this scenario, our math sample is reduced to 1,326 students and 182 teachers, and our language sample to 1,542 students and 211 teachers. We rerun our analysis for these subsamples of teachers and present the estimations of the classroom effects on student math competence in table A 4.1, and on language competence in table A 4.2, following the structure of table 4.7.

Table A 4.1: Estimates of Classroom Effects on Math Competence, Declared Math Teachers

	(1)	(2)	(3)	(4)	(5)
Classroom Fixed Effects (FE):					
Standard deviation	0.365	0.364	0.363	0.362	0.362
Adjusted EB standard deviation	0.121	0.126	0.125	0.124	0.124
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Classroom Random Effects (RE):					
EB Standard deviation	0.124	0.129	0.129	0.129	0.129
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Included covariate:					
Federal State effects	YES	YES	YES	YES	YES
Lagged test scores	YES	YES	YES	YES	YES
Student characteristics	NO	YES	YES	YES	YES
Family background	NO	NO	YES	YES	YES
Classroom size	NO	NO	NO	YES	YES
Classroom averages	NO	NO	NO	NO	YES
Number of teachers/classroom	182	182	182	182	182
Number of students threshold	5	5	5	5	5

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This table presents standard deviations of classroom value-added to math competence distributions estimated as teacher FE using OLS regressions, and RE using hierarchical multilevel (mixed) regressions for the declared math teacher sample. Empirical Bayes (EB) shrinkage was implemented to obtain the adjusted FE standard deviation. All results are based on regressions of math competence test scores in grade 2 on lagged math, grammar and science competence test scores, time between tests and federal state fixed effects. Columns (2)-(5) control for the following student characteristics: age, gender, migration background; for parental background: highest years of education, highest ISEI, number of siblings; and for classroom averages: proportion of females, average ISEI.

Table A 4.2: Estimates of Classroom Effects on Language Competence, Declared Language Teachers

	(1)	(2)	(3)	(4)	(5)
Classroom Fixed Effects (FE):					
Standard deviation	0.391	0.386	0.387	0.387	0.384
Adjusted EB standard deviation	0.131	0.124	0.130	0.129	0.124
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Classroom Random Effects (RE):					
EB Standard deviation	0.131	0.124	0.131	0.130	0.123
p-value, F-test of classroom effects	0.000	0.000	0.000	0.000	0.000
Included covariate:					
Federal State effects	YES	YES	YES	YES	YES
Lagged test scores	YES	YES	YES	YES	YES
Student characteristics	NO	YES	YES	YES	YES
Family background	NO	NO	YES	YES	YES
Classroom size	NO	NO	NO	YES	YES
Classroom averages	NO	NO	NO	NO	YES
Number of teachers/classrooms	211	211	211	211	211
Number of students threshold	5	5	5	5	5

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This table presents standard deviations of classroom value-added to math competence distributions estimated as teacher FE using OLS regressions, and RE using hierarchical multilevel (mixed) regressions for the declared language teacher sample. Empirical Bayes (EB) shrinkage was implemented to obtain the adjusted FE standard deviation. All results are based on regressions of math competence test scores in grade 2 on lagged math, grammar and science competence test scores, time between tests and federal state fixed effects. Columns (2)-(5) control for the following student characteristics: age, gender, migration background; for parental background: highest years of education, highest ISEI, number of siblings; and for classroom averages: proportion of females, average ISEI.

Results in table A 4.1 show that we obtain practically the same adjusted FE and RE distributions of classroom quality or effects on student math competence, compared to the original sample. Our preferred model specification, which includes a full set of control variables in column (5), shows that a one standard deviation increase in classroom quality is associated with a 0.124 standard deviation higher student math competence score when estimated with adjusted FE, and with a 0.129 standard deviation when estimated with RE. By contrast, we find slightly smaller standard deviations in the distributions of classroom effects on student language competence, displayed in table A 4.2. Our preferred estimations of the classroom adjusted FE and RE distributions presented in column (5) correspond to 0.124 and 0.123 standard deviations respectively, which are about two percentage points smaller than those found in the original sample.

Table A 4.3: Association of Teacher Characteristics and Classroom Effects on Math and Language Competence, Declared Math or Language Teachers

	Math		Language	
Teacher	EB Adjusted Fixed Effect (1)	EB Random Effect (2)	EB Adjusted Fixed Effect (3)	EB Random Effect (4)
Female	-0.112**	-0.118**	-0.005	-0.007
	(0.049)	(0.051)	(0.038)	(0.037)
Experience	0.001	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Abitur GPA	-0.031**	-0.032*	0.021	0.022*
	(0.016)	(0.016)	(0.013)	(0.013)
FSE Grade	0.014	0.013	-0.029	-0.029
	(0.014)	(0.014)	(0.018)	(0.018)
SSE Passed	0.033	0.036	0.003	0.003
	(0.032)	(0.033)	(0.038)	(0.038)
Migration background	0.070*	0.073*	0.005	0.006
	(0.038)	(0.040)	(0.059)	(0.059)
Constructivist beliefs	0.000	-0.000	0.011	0.011
	(0.016)	(0.016)	(0.011)	(0.011)
Exhaustion	-0.008	-0.008	-0.010	-0.010
	(0.012)	(0.012)	(0.013)	(0.013)
Parental evaluation	-0.004	-0.003	0.019*	0.019*
	(0.010)	(0.011)	(0.011)	(0.011)
Constant	0.045	0.041	-0.021	-0.021
	(0.057)	(0.059)	(0.054)	(0.053)
Number of teacher with observables	107	107	130	130
R^2	0.090	0.090	0.080	0.081

Data: NEPS SUF, SC2 8.0.1, own calculations. *Notes:* This table presents OLS regressions of classroom value-added to math and language competence on teacher characteristics for the declared math or language teacher sample. Self-reported *Abitur* GPA and First State Examination (FSE) grades originally were on a scale from 1.0 to 4.0, with 1.0 being the best possible grade and 4.0 the minimum passing grade. These self-reported grades were standardized to have zero mean and a one-unit standard deviation with respect to the full sample of teachers. Standard errors in parentheses. * Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

Table A 4.3 show the association between our vector of teacher characteristics and the estimated classroom effects on student math and language competences, following the same structure of table 4.9. It should be noted that the number of observations with full information is thus reduced to 107 math teachers and 130 language teachers, which implies higher selectivity in the sample. Columns (1) and (2) show a consistent negative association between classroom value-added to math competence and being a female teacher, which is statically significant at the five percent significance level for the adjusted FE and RE specifications. Moreover, there is also a

negative association with *Abitur* GPA at the 5 percent significance level for the adjusted FE estimation and at the 10 percent for the RE, which actually means that higher math classroom value-added is associated with higher percentiles, or better *Abitur* GPA. With respect to classroom effects on language competence, we confirm a positive association with parental evaluation, but in this case at the 10 percent significance level as shown in columns (3) and (4). Given that all these findings are based on a limited number of observations, the results should be treated with considerable caution.

References

Aaronson, D., Barrow, L. and Sander, W. (2007) 'Teachers and student achievement in the Chicago public high schools', *Journal of Labor Economics*, 25(1), pp. 95–135. doi: 10.1086/508733.

Alkire, S., Conconi, A., Robles, G., Roche, J. M., Santos, M. E., Seth, S. and Vaz, A. (2016) *The Global Multidimensional Poverty Index (MPI): 5-year methodological note*, *OPHI Briefing*. 37. Oxford. Available at: https://www.ophi.org.uk/wp-content/uploads/MPI_Methodology_2010-2015_Jan2016.pdf.

Almlund, M., Duckworth, A. L., Heckman, J. J. and Kautz, T. (2011) 'Personality Psychology and Economics', in Hanushek, E. A., Machin, S., and Woessmann, L. (eds) *Handbook of The Economics of Education*. Volume 4. Amsterdam: Elsevier, pp. 1–181. doi: 10.1016/B978-0-444-53444-6.00001-8.

Angrist, J. D. and Guryan, J. (2008) 'Does teacher testing raise teacher quality? Evidence from state certification requirements', *Economics of Education Review*, 27, pp. 483–503. doi: 10.1016/j.econedurev.2007.03.002.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y. and Schady, N. (2016) 'Teacher Quality and Learning Outcomes in Kindergarten', *The Quarterly Journal of Economics*, 131(3), pp. 1415–1453. doi: 10.1093/qje/qjw016.

Araujo, M. C., Bosch, M. and Schady, N. (2019) 'Can Cash Transfers Help Households Escape an Intergenerational Poverty Trap?', in Barrett, C. B. et al. (eds) *The Economics of Poverty Traps*. University of Chicago Press, pp. 357–382. doi: doi:10.7208/9780226574448-015.

Araujo P., M. D. (2019) *Measuring the Effect of Competitive Teacher Recruitment on Student Achievement: Evidence from Ecuador*, *BERG Working Paper Series*. No. 150. Bamberg. Available at: https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/sowi_faecher/vwl/BERG/BERG_150.pdf.

Araujo P., M. D. and Bramwell, D. (2015) ‘Cambios en la política educativa en Ecuador desde el año 2000’, in *Paper commissioned for the EFA Global Monitoring Report 2015, Education for All 2000-2015: achievements and challenges*. Paris: UNESCO, p. 34. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000232430.

Asamblea Nacional del Ecuador (2011) *LEY ORGÁNICA DE EDUCACIÓN INTERCULTURAL*, *Registro Oficial Nº 417*. Ecuador.

Bacher-Hicks, A., Chin, M. J., Kane, T. J. and Staiger, D. O. (2019) ‘An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys’, *Economics of Education Review*, 73, p. 101919. doi: 10.1016/j.econedurev.2019.101919.

Bacher-Hicks, A., Kane, T. J. and Staiger, D. O. (2014) *Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles*, *NBER Working Paper Series*. 20657. Cambridge, MA. doi: 10.3386/w20657.

Bakos, D. S., Denburg, N., Fonseca, R. P. and de Mattos Pimenta Parente, M. A. (2010) ‘A cultural study on decision making: Performance differences on the Iowa Gambling Task between selected groups of Brazilians and Americans.’, *Psychology & Neuroscience*, 3(1), pp. 101–107. doi: 10.3922/j.psns.2010.1.013.

Ballou, D., Sanders, W. and Wright, P. (2004) ‘Controlling for Student Background in Value-Added Assessment of Teachers’, *Journal of Educational and Behavioral Statistics*, 29(1), pp. 37–65. doi: 10.3102/10769986029001037.

Barrick, M. R. and Mount, M. K. (1991) ‘The Big Five Personality Dimensions and Job Performance: a Meta-Analysis’, *Personnel Psychology*, 44(1), pp. 1–26. doi: 10.1111/j.1744-6570.1991.tb00688.x.

Bastian, K. C. (2019) ‘A degree above? The value-added estimates and evaluation ratings of teachers with a graduate degree’, *Education Finance and Policy*, 14(4), pp. 652–678. doi: 10.1162/edfp_a_00261.

Bau, N. and Das, J. (2020) 'Teacher Value Added in a Low-Income Country', *American Economic Journal: Economic Policy*, 12(1), pp. 62–96. doi: 10.1257/pol.20170243.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M. and Tsai, Y.-M. (2010) 'Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress', *American Educational Research Journal*, 47(1), pp. 133–180. doi: 10.3102/0002831209345157.

Becker, G. S. (1962) 'Investment in Human Capital: A Theoretical Analysis', *Journal of Political Economy*, 70(5), pp. 9–49. Available at: http://www.jstor.org/stable/1829103.

Becker, G. S. (1964) *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*. First. New York : Columbia University Press. Available at: https://www.nber.org/books-and-chapters/human-capital-theoretical-and-empirical-analysis-special-reference-education-first-edition (Accessed: 11 May 2021).

Ben-Porath, Y. (1967) 'The Production of Human Capital and the Life Cycle of Earnings', *Journal of Political Economy*, 75(4), pp. 352–365. Available at: https://www.jstor.org/stable/1828596 (Accessed: 24 May 2021).

Bietenbeck, J., Piopiunik, M. and Wiederhold, S. (2018) 'Africa's skill tragedy: Does teachers' lack of knowledge lead to low student performance?', *Journal of Human Resources*, 53(3), pp. 553–578. doi: 10.3368/jhr.53.3.0616-8002R1.

Bitler, M., Corcoran, S., Domina, T. and Penner, E. (2019) *Teacher Effects on Student Achievement and Height: A Cautionary Tale*, *NBER Working Paper Series*. 26480. Cambridge, MA. doi: 10.3386/w26480.

Blossfeld, H.-P., Roßbach, H.-G. and von Maurice, J. (2011) 'Education as a lifelong process: the German National Educational Panel Study (NEPS)', *Zeitschrift für Erziehungswissenschaft*, 14(Special Issue).

Borman, G. D. and Kimball, S. M. (2005) 'Teacher Quality and Educational Equality: Do Teachers with Higher Standards-Based Evaluation Ratings Close Student Achievement Gaps?', *The Elementary School Journal*, 106(1), pp. 3–20. doi: 10.1086/496904.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J. and Wyckoff, J. (2008) 'The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools', *Journal of Policy Analysis and Management*, 27(4), pp. 793–818. doi: 10.1002/pam.20377.

Bruno, P. and Strunk, K. O. (2019) 'Making the Cut: The Effectiveness of Teacher Screening and Hiring in the Los Angeles Unified School District', *Educational Evaluation and Policy Analysis*, 41(4), pp. 426–460. doi: 10.3102/0162373719865561.

Bruns, B. and Luque, J. (2015) *Great Teachers: How to raise student learning in Latin American and the Caribbean*. Washington D.C.: The World Bank. doi: 10.1596/978-1-4648-0151-8.

Brutti, Z. and Sánchez, F. (2017) *Does Better Teacher Selection Lead to Better Students? Evidence from a Large Scale Reform in Colombia*, *Documentos CEDE*. 11. Bogotá: Centro de Estudios sobre Desarrollo Económico. Available at: https://ideas.repec.org/p/col/000089/015350.html (Accessed: 3 January 2018).

Caliendo, M. and Kopeinig, S. (2008) 'Some practical guidance for the implementation of propensity score matching', *Journal of Economic Surveys*, 22(1), pp. 31–72. doi: 10.1111/j.1467-6419.2007.00527.x.

Card, D. (1999) 'The causal effect of education on earnings', in Ashenfelter, O. and Card, D. (eds) *Handbook of Labor Economics*. Elsevier B.V., pp. 1801–1863. doi: 10.1016/S1573-4463(99)03011-4.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W. and Yagan, D. (2011) 'How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star', *The Quarterly Journal of Economics*, 126(4), pp. 1593–1660. doi: 10.1093/qje/qjr041.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014a) 'Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates', *American Economic Review*, 104(9), pp. 2593–2632. doi: 10.1257/aer.104.9.2593.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014b) 'Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood', *American Economic Review*, 104(9), pp. 2633–2679. doi: 10.1257/aer.104.9.2633.

Chingos, M. M. and Peterson, P. E. (2011) 'It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness', *Economics of Education Review*, 30(3), pp. 449–465. doi: 10.1016/j.econedurev.2010.12.010.

Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (2006) 'Teacher-Student Matching and the Assessment of Teacher Effectiveness', *Journal of Human Resources*, 41(4), pp. 778–820. doi: 10.2307/40057291.

Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (2007a) *How and Why do Teacher Credentials Matter for Student Achievement?*, *NEPS Survey Papers*. 12828. Cambridge, MA. doi: 10.3386/w12828.

Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (2007b) 'Teacher credentials and student achievement: Longitudinal analysis with student fixed effects', *Economics of Education Review*, 26(6), pp. 673–682. doi: 10.1016/j.econedurev.2007.10.002.

Costa Jr., P. T. and McCrae, R. R. (2008) 'The Revised NEO Personality Inventory (NEO-PI-R).', in *The SAGE handbook of personality theory and assessment, Vol 2: Personality measurement and testing.* Thousand Oaks, CA, US: Sage Publications, Inc, pp. 179–198. doi: 10.4135/9781849200479.n9.

Cruz-Aguayo, Y., Ibarrarán, P. and Schady, N. (2017) 'Do tests applied to teachers predict their effectiveness?', *Economics Letters*, 159, pp. 108–111. doi: 10.1016/J.ECONLET.2017.06.035.

Cunha, F., Heckman, J. J. and Schennach, S. M. (2010) 'Estimating the Technology of Cognitive and Noncognitive Skill Formation', *Econometrica*, 78(3), pp. 883–931. doi: 10.3982/ECTA6551.

D'Agostino, J. V. and Powers, S. J. (2009) 'Predicting Teacher Performance With Test Scores and Grade Point Average: A Meta-Analysis', *American Educational Research Journal*, 46(1), pp. 146–182. doi: 10.3102/0002831208323280.

Demmer, M. and von Saldern, M. (2010) *„Helden des Alltags" Erste Ergebnisse der Schulleitungs- und Lehrkräftebefragung (TALIS) in Deutschland*. First. Edited by M. Demmer and M. von Saldern. Münster: Waxmann.

Dubeck, M. M. and Gove, A. (2015) 'The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations', *International Journal of Educational Development*, 40, pp. 315–322. doi: 10.1016/j.ijedudev.2014.11.004.

Duflo, E., Dupas, P. and Kremer, M. (2015) 'School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools', *Journal of Public Economics*, 123, pp. 92–110. doi: 10.1016/j.jpubeco.2014.11.008.

Duflo, E., Glennerster, R. and Kremer, M. (2008) 'Using Randomization in Development Economics Research: A Toolkit', in Schultz, T. P. and Strauss, J. (eds) *Handbook of Development Economics*. 1st edn. Amsterdam and New York: North Holland.

Dunn, Lloyd, Padilla, E., Lugo, D. and Dunn, Leota (1986) *Test de Vocabulario en Imagenes Peabody*. Circle Pines, MN: American Guidance Service.

Elacqua, G., Hincapié, D., Vegas, E. and Alfonso, M. (2017) *Profesión: Profesor en América Latina. ¿Por qué se perdió el prestigio docente y cómo recuperarlo?* Washington, D.C.: Inter-American Development Bank. doi: 10.18235/0000901.

Elango, S., García, J. L., Heckman, J. J. and Hojman, A. (2016) 'Early Childhood Education', in Moffitt, R. A. (ed.) *Economics of Means-Tested Transfer Programs in the United States*. University of Chicago Press, pp. 235–297. doi: 10.7208/chicago/9780226392523.001.0001.

Enzi, B. (2017) *Microeconometric Analyses of Cognitive Achievement Production*, *ifo Beiträge zur Wirtschaftsforschung*. No. 75. München: ifo Institut - Leibniz-Institut für Wirtschaftsforschung an der Universität München. Available at: https://www.econstor.eu/handle/10419/172967 (Accessed: 27 June 2019).

Estrada, R. (2019) 'Rules versus Discretion in Public Service: Teacher Hiring in Mexico', *Journal of Labor Economics*, 37(2), pp. 545–579. doi: 10.1086/700192.

Fasfous, A. F., Hidalgo-Ruzzante, N., Vilar-López, R., Catena-Martínez, A. and Pérez-García, M. (2013) 'Cultural Differences in Neuropsychological Abilities Required to Perform Intelligence Tasks', *Archives of Clinical Neuropsychology*, 28(8), pp. 784–790. doi: 10.1093/arclin/act074.

Fernández, E., LeChasseur, K. and Donaldson, M. L. (2018) 'Responses to including parents in teacher evaluation policy: A critical policy analysis', *Journal of Education Policy*, 33(3), pp. 398–413. doi: 10.1080/02680939.2017.1370135.

Fischer, L., Rohm, T., Gnambs, T. and Carstensen, C. H. (2016) *Linking the Data of the Competence Tests*, *NEPS Survey Papers*. 1. Bamberg. doi: 10.5157/NEPS:SP01:1.0.

García, J. L., Heckman, J. J., Leaf, D. E. and Prados, M. J. (2020) 'Quantifying the Life-Cycle Benefits of an Influential Early-Childhood Program', *Journal of Political Economy*, 128(7), pp. 2502–2541. doi: 10.1086/705718.

Glewwe, P., Hanushek, E. A., Humpage, S. D. and Ravina, R. (2014) 'School Resources and Educational Outcomes in Developing Countries: A review of the literature from 1990 to 2010', in Paul Glewwe (ed.) *Education Policy in Developing Countries*. Chicago and London: The University of Chicago Press, pp. 13–64. doi: 10.3386/w17554.

Glewwe, P., Shen, R., Sun, B. and Wisniewski, S. (2020) 'Teachers in developing countries', in Bradley, S. and Green, C. B. T. (eds) *The Economics of Education*. Second. Elsevier, pp. 371–389. doi: 10.1016/b978-0-12-815391-8.00027-6.

Goldhaber, D. (2007) ‘Everyone’s Doing It, but What Does Teacher Testing Tell Us about Teacher Effectiveness?’, *The Journal of Human Resources*, 42(4), pp. 765–794. doi: 10.2307/40057329.

Goldhaber, D. and Anthony, E. (2007) ‘Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching’, *The Review of Economics and Statistics*, 89(1), pp. 134–150. doi: 10.1162/rest.89.1.134.

Goldhaber, D., Gratz, T. and Theobald, R. (2017) ‘What’s in a teacher test? Assessing the relationship between teacher licensure test scores and student STEM achievement and course-taking’, *Economics of Education Review*, 61, pp. 112–129. doi: 10.1016/j.econedurev.2017.09.002.

Goldhaber, D., Grout, C. and Huntington-Klein, N. (2017) ‘Screen twice, cut once: Assessing the predictive validity of applicant selection tools’, *Education Finance and Policy*, 12(2), pp. 197–223. doi: 10.1162/EDFP_a_00200.

Goldhaber, D. and Hansen, M. (2013) ‘Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance’, *Economica*, 80(319), pp. 589–612. doi: 10.1111/ecca.12002.

Guarino, C. M., Reckase, M. D. and Wooldridge, J. M. (2015) ‘Can value-added measures of teacher performance be trusted?’, *Education Finance and Policy*, 10(1), pp. 117–156. doi: 10.1162/EDFP_a_00153.

Gunderson, M. and Oreopolous, P. (2020) ‘Returns to education in developed countries’, in Bradley, S. and Green, C. (eds) *The Economics of Education: A Comprehensive Overview*. Second. Academic Press, pp. 39–51. doi: https://doi.org/10.1016/B978-0-12-815391-8.00003-3.

Hampf, F., Wiederhold, S. and Woessmann, L. (2017) ‘Skills, earnings, and employment: exploring causality in the estimation of returns to skills’, *Large-Scale Assessments in Education*, 5(1), pp. 1–30. doi: 10.1186/s40536-017-0045-7.

Hamre, B. K., La Paro, K. M. and Pianta, R. C. (2007) *Classroom assessment scoring system (CLASS) manual*. Baltimore: Paul H. Brookes Publishing Co., Inc.

Hanushek, E. A. (1971) ‘Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data’, *The American Economic Review*, 61, pp. 280–288. doi: 10.2307/1817003.

Hanushek, E. A. (1979) ‘Conceptual and Empirical Issues in the Estimation of Educational Production Functions’, *The Journal of Human Resources*, 14(3), pp. 351–388. doi: 10.2307/145575.

Hanushek, E. A. (2003) ‘The Failure of Input-based Schooling Policies’, *The Economic Journal*, 113(485), pp. F64–F98. doi: 10.1111/1468-0297.00099.

Hanushek, E. A. (2011) ‘The economic value of higher teacher quality’, *Economics of Education Review*, 30(3), pp. 466–479. doi: 10.1016/j.econedurev.2010.12.006.

Hanushek, E. A., Schwerdt, G., Wiederhold, S. and Woessmann, L. (2015) ‘Returns to skills around the world: Evidence from PIAAC’, *European Economic Review*, 73, pp. 103–130. doi: 10.1016/j.euroecorev.2014.10.006.

Hanushek, E. A. (2020) ‘Education production functions’, in Bradley, S. and Green, C. (eds) *The Economics of Education: A Comprehensive Overview*. Second. Elsevier, pp. 161–170. doi: 10.1016/B978-0-12-815391-8.00013-6.

Hanushek, E. A., Peterson, P., Talpey, L. and Woessmann, L. (2020) *Long-Run Trends in the U.S. SES-Achievement Gap*, *NBER Working Paper Series*. 26764. Cambridge, MA. doi: 10.3386/w26764.

Hanushek, E. A. and Kimko, D. D. (2000) ‘Schooling, Labor Force Quality, and the Growth of Nations’, *American Economic Review*, 90(5), pp. 1184–1208. doi: 10.1257/aer.90.5.1184.

Hanushek, E. A., Piopiunik, M. and Wiederhold, S. (2019) ‘The Value of Smarter Teachers’, *Journal of Human Resources*, 54(4), pp. 857–899. doi: 10.3368/jhr.54.4.0317.8619R1.

Hanushek, E. A. and Rivkin, S. G. (2006) ‘Teacher Quality’, in Hanushek, E. and Welch, F. (eds) *Handbook of the Economics of Education*. Elsevier, pp. 1052–1075. doi: 10.1016/S1574-0692(06)02018-6.

Hanushek, E. A. and Rivkin, S. G. (2010) 'Generalizations about using value-added measures of teacher quality', in *American Economic Review: Papers & Proceedings*, pp. 267–271. doi: 10.1257/aer.100.2.267.

Hanushek, E. A. and Rivkin, S. G. (2012) 'The Distribution of Teacher Quality and Implications for Policy', *Annual Review of Economics*, 4(1), pp. 131–157. doi: 10.1146/annurev-economics-080511-111001.

Hanushek, E. A. and Woessmann, L. (2008) 'The role of cognitive skills in economic development', *Journal of Economic Literature*, 46(3), pp. 607–668. doi: 10.1257/jel.46.3.607.

Hanushek, E. A. and Woessmann, L. (2011) 'The Economics of International Differences in Educational Achievement', in Hanushek, E. A., Machin, S., and Woessmann, L. (eds) *Handbook of the Economics of Education*. First. Elsevier, pp. 89–200. doi: 10.1016/B978-0-444-53429-3.00002-8.

Hanushek, E. A. and Woessmann, L. (2012a) 'Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation', *Journal of Economic Growth*. doi: 10.1007/s10887-012-9081-x.

Hanushek, E. A. and Woessmann, L. (2012b) 'Schooling, educational achievement, and the Latin American growth puzzle', *Journal of Development Economics*, 99, pp. 497–512. doi: 10.1016/j.jdeveco.2012.06.004.

Hanushek, E. A. and Zhang, L. (2009) 'Quality-consistent estimates of international schooling and skill gradients', *Journal of Human Capital*, 3(2), pp. 107–143. doi: 10.1086/644780.

Harmon, C., Oosterbeek, H. and Walker, I. (2003) 'The returns to education: Microeconomics', *Journal of Economic Surveys*, 17(2), pp. 115–156. doi: 10.1111/1467-6419.00191.

Harris, D. N. and Sass, T. R. (2009) 'The effects of NBPTS- Certified Teachers on Student', *Journal of Policy Analysis and Management*, 28(1), pp. 55–80. doi: 10.2307/29738986.

Harris, D. N. and Sass, T. R. (2011) 'Teacher training, teacher quality and student achievement', *Journal of Public Economics*, 95(7–8), pp. 798–812. doi: 10.1016/J.JPUBECO.2010.11.009.

Heckman, J. J., Humphries, J. E. and Veramendi, G. (2018) 'Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking', *Journal of Political Economy*, 126(S1), pp. S197–S246. doi: 10.1086/698760.

Heckman, J. J., Lochner, L. J. and Todd, P. E. (2006) 'Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond', in Hanushek, E. A. and Welch, F. (eds) *Handbook of the Economics of Education*. Elsevier, pp. 307–458. doi: https://doi.org/10.1016/S1574-0692(06)01007-5.

Heckman, J. J., Pinto, R. and Savelyev, P. (2013) 'Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes', *The American Economic Review*, 103(6), pp. 2052–2086. doi: 10.1257/aer.103.6.2052.

Heineck, G. and Anger, S. (2010) 'The returns to cognitive abilities and personality traits in Germany', *Labour Economics*, 17(3), pp. 535–546. doi: 10.1016/j.labeco.2009.06.001.

Herrmann, M., Walsh, E. and Isenberg, E. (2016) 'Shrinkage of Value-Added Estimates and Characteristics of Students with Hard-to-Predict Achievement Levels', *Statistics and Public Policy*, 3(1), pp. 1–10. doi: 10.1080/2330443X.2016.1182878.

Hill, H. C., Rowan, B. and Loewenberg Ball, D. (2005) 'Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement', *American Educational Research Journal; Summer*, 42(2), pp. 371–406. doi: 10.3102/00028312042002371.

Hoekstra, M. (2020) 'Returns to education quality', in Bradley, S. and Green, C. (eds) *The Economics of Education: A Comprehensive Overview*. Second. Academic Press, pp. 65–73. doi: https://doi.org/10.1016/B978-0-12-815391-8.00005-7.

Imbens, G. W. (2015) ‘Matching methods in practice: Three examples’, *Journal of Human Resources*, 50(2), pp. 373–419. doi: 10.3368/jhr.50.2.373.

Imbens, G. W. and Wooldridge, J. M. (2009) ‘Recent developments in the econometrics of program evaluation’, *Journal of Economic Literature*, 47(1), pp. 5–86. doi: 10.1257/jel.47.1.5.

Instituto Ecuatoriano de Estadísticas y Censos (INEC) (2020) *Serie Histórica IPC*, *Ecuador en Cifras*. Available at: https://www.ecuadorencifras.gob.ec/historicos-ipc/ (Accessed: 28 April 2020).

Jackson, C. K., Rockoff, J. E. and Staiger, D. O. (2014) ‘Teacher Effects and Teacher-Related Policies’, *Annual Review of Economics*, 6, pp. 801–25. doi: 10.1146/annurev-economics-080213-040845.

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B. and Rosen, R. (2018) ‘Teacher applicant hiring and teacher performance: Evidence from DC public schools’, *Journal of Public Economics*, 166, pp. 81–97. doi: 10.1016/j.jpubeco.2018.08.011.

James, J. and Wyckoff, J. (2020) ‘Teacher labor markets: An overview’, in Bradley, S. and Green, C. B. T. (eds) *The Economics of Education*. Elsevier, pp. 355–370. doi: 10.1016/b978-0-12-815391-8.00026-4.

Jürges, H. and Schneider, K. (2007) ‘Fair ranking of teachers’, *Empirical Economics*, 32, pp. 411–431. doi: 10.1007/s00181-006-0112-3.

Kane, T. J., Taylor, E. S., Tyler, J. H. and Wooten, A. L. (2011) ‘Identifying Effective Classroom Practices Using Student Achievement Data’, *The Journal of Human Resources*, 46(3), pp. 587–613. doi: 10.3368/jhr.46.3.587.

Kane, T. J., Mccaffrey, D. F., Miller, T. and Staiger, D. O. (2013) *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment*, *MET Project Reseach Paper*. Bill & Melinda Gates Foundation. Available at: https://usprogram.gatesfoundation.org/-/media/dataimport/resources/pdf/2016/12/met-validating-using-random-assignment-research-paper.pdf (Accessed: 1 April 2020).

Kane, T. J., Rockoff, J. E. and Staiger, D. O. (2008) 'What does certification tell us about teacher effectiveness? Evidence from New York City', *Economics of Education Review*, 27, pp. 615–631. doi: 10.1016/j.econedurev.2007.05.005.

Kane, T. J. and Staiger, D. O. (2008) *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*, *NBER Working Paper Series*. 14607. Cambridge, MA. doi: 10.3386/w14607.

Kane, T. J. and Staiger, D. O. (2012) *Gathering Feedback for Teaching Combining High-Quality Observations with Student Surveys and Achievement Gains*, *MET Project Reseach Paper*. Bill & Melinda Gates Foundation. Available at: www.metproject.org (Accessed: 1 April 2020).

Klusmann, U., Kunter, M., Trautwein, U., Lüdtke, O. and Baumert, J. (2008) 'Teachers' Occupational Well-Being and Quality of Instruction: The Important Role of Self-Regulatory Patterns', *Journal of Educational Psychology*, 100(3), pp. 702–715. doi: 10.1037/0022-0663.100.3.702.

Koedel, C. and Betts, J. R. (2011) 'Does student sorting invalidate value-added models of teacher effectiveness? an extended analysis of the rothstein critique', *Education Finance and Policy*, 6(1), pp. 18–42. doi: 10.1162/EDFP_a_00027.

Koedel, C., Mihaly, K. and Rockoff, J. E. (2015) 'Value-added modeling: A review', *Economics of Education Review*, 47, pp. 180–195. doi: 10.1016/J.ECONEDUREV.2015.01.006.

Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) (2019) *The Education System in the Federal Republic of Germany 2017/2018*. Edited by T. Eckhardt. Bonn: Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). Available at: https://www.kmk.org/fileadmin/Dateien/pdf/Eurydice/Bildungswesen-engl-pdfs/dossier_en_ebook.pdf (Accessed: 28 February 2021).

König, W., Lüttinger, P. and Müller, W. (1988) *A Comparative Analysis of the Development and Structure of Educational Systems. Methodological foundations and the construction of a comparative educational scale*. 12. Mannheim. Available at: https://www.gesis.org/fileadmin/upload/dienstleistung/tools_standards/mikrodaten_tools/CASMIN/Koenig_Casmin.pdf (Accessed: 25 March 2021).

Krueger, A. B. (1999) 'Experimental estimates of education production functions', *The Quarterly Journal of Economics*, 114(2), pp. 497–532. Available at: https://academic.oup.com/qje/article-abstract/114/2/497/1844226 (Accessed: 11 December 2017).

Kuncel, N. R., Ones, D. S. and Sackett, P. R. (2010) 'Individual differences as predictors of work, educational, and broad life outcomes', *Personality and Individual Differences*. doi: 10.1016/j.paid.2010.03.042.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T. and Hachfeld, A. (2013) 'Professional competence of teachers: Effects on instructional quality and student development.', *Journal of Educational Psychology*, 105(3), pp. 805–820. doi: 10.1037/a0032583.

Lechert, Y., Schroedter, J. and Lüttinger, P. (2006) *Die Umsetzung der Bildungsklassifikation CASMIN für die Volkszählung 1970, die Mikrozensus-Zusatzerhebung 1971 und die Mikrozensen 1976-2004*, *ZUMA-Methodenbericht 2006/12*. Mannheim. Available at : https://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2006/06_12_lechert.pdf (Accessed: 25 March 2021).

Lenhard, W. and Schneider, W. (2006) *ELFE 1-6: Ein Leseverständnistest für Erst- bis Sechstklässler*. First. Göttingen: Hogrefe.

Lockwood, J. R. and McCaffrey, D. F. (2014) 'Correcting for Test Score Measurement Error in ANCOVA Models for Estimating Treatment Effects', *Journal of Educational and Behavioral Statistics*, 39(1), pp. 22–52. doi: 10.3102/1076998613509405.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R. and Mihaly, K. (2009) 'The Intertemporal Variability of Teacher Effect Estimates', *Education Finance and Policy*, 4(4), pp. 572–606. doi: 10.1162/edfp.2009.4.4.572.

McMahon, W. W. (2017) *Higher Learning, Greater Good. The Private and Social Benefits of Higher Education*. Second. Baltimore: Johns Hopkins University Press Books.

Metzler, J. and Woessmann, L. (2012) 'The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation', *Journal of Development Economics*, 99(2), pp. 486–496. doi: 10.1016/J.JDEVECO.2012.06.002.

Mincer, J. (1970) 'The Distribution of Labor Incomes: A Survey With Special Reference to the Human Capital Approach', *Journal of Economic Literature*, 8(1), pp. 1–26. Available at: http://www.jstor.org/stable/2720384.

Mincer, J. A. (1974) *Schooling, Experience, and Earnings*. New York: Columbia University Press.

Ministerio de Educación del Ecuador (2007) *Acuerdo Ministerial 438-07: NORMAS DE LOS CONCURSOS DE MÉRITOS Y OPOSICIÓN PARA EL INGRESO AL MAGISTERIO NACIONAL*, *Acuerdos Ministeriales*. Quito, Ecuador: Ministerio de Educación del Ecuador (MinEduc).

Ministerio de Educación del Ecuador (2008) *Acuerdo Ministerial 363: NORMAS PARA LLENAR VACANTES A TRAVÉS DEL SISTEMA DE RUEDAS DE CAMBIOS*, *Acuerdos Ministeriales*. Ecuador: Ministerio de Educación del Ecuador (MinEduc).

Ministerio de Educación del Ecuador (2010) *Acuerdo Ministerial 018-10: NORMATIVA PARA LOS CONCURSOS DE MÉRITOS Y OPOSICIÓN PARA LLENAR VACANTES DEL MAGISTERIO NACIONAL*, *Acuerdos Ministeriales*. Ecuador: Ministerio de Educación del Ecuador (MinEduc).

Ministerio de Educación del Ecuador (2011) *Acuerdo Ministerial 379-11: NORMATIVA DE CONCURSOS DE MÉRITOS Y OPOSICIÓN PARA LLENAR VACANTES DE DOCENTES EN EL SECTOR PÚBLICO*, *Acuerdos Ministeriales*. Ecuador: Ministerio de Educación del Ecuador (MinEduc).

Ministerio de Educación del Ecuador (2012) 'Ecuador 2007-2012 a Transformation in Education. Presentation at the Inter-American Development Bank'. Washington D.C.: Ministerio de Educación del Ecuador (MinEduc).

Ministerio de Educación del Ecuador (2020) *AMIE-Registros Administrativos 2011-2012 (FiN)*, *Estadísticas Educativas*. Available at: https://educacion.gob.ec/amie/.

Morris, C. N. (1983) 'Parametric Empirical Bayes Inference: Theory and Applications', *Journal of the American Statistical Association*, 78(381), pp. 47–55. doi: 10.2307/2287098.

Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S. and Mather, N. (2005) *Batería III Woodcock-Muñoz*. Itasca, IL: Riverside Publishing.

Muralidharan, K. and Sundararaman, V. (2013) *Contract Teachers: Experimental Evidence from India*, *NBER Working Papers*. 19440. doi: 10.3386/w19440.

Murnane, R. J. (1975) *The impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger Publishing Company.

National Education Panel Study (NEPS) (2019) *Study Overview. NEPS Starting Cohort 3 — Grade 5. Paths Through Lower Secondary School — Educational Pathways of Students in Grade 5 and Higher. Waves 1 to 9*. Bamberg.

Nye, B., Konstantopoulos, S. and Hedges, L. V. (2004) 'How Large Are Teacher Effects?', *Educational Evaluation and Policy Analysis*, 26(3), pp. 237–257. doi: 10.3102/01623737026003237.

Organisation for Economic Co-Operation and Development (OECD) (2010) *TALIS 2008 Technical Report: Teaching And Learning International Survey*. Paris. Available at: http://www.oecd.org/education/school/44978960.pdf (Accessed: 15 January 2021).

Organisation for Economic Co-Operation and Development (OECD) (2014) *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition, February 2014)*. Paris: OECD (PISA). doi: 10.1787/efaa764e-en.

Organisation for Economic Co-Operation and Development (OECD) (2016) *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD (PISA). doi: 10.1787/9789264266490-en.

Papay, J. P. and Kraft, M. A. (2015) 'Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement', *Journal of Public Economics*, 130, pp. 105–119. doi: 10.1016/j.jpubeco.2015.02.008.

Patrinos, H. A. and Psacharopoulos, G. (2020) 'Returns to education in developing countries', in Bradley, S. and Green, C. (eds) *The Economics of Education: A Comprehensive Overview*. Second. Academic Press, pp. 53–64. doi: https://doi.org/10.1016/B978-0-12-815391-8.00004-5.

Paufler, N. A. and Amrein-Beardsley, A. (2014) 'The Random Assignment of Students Into Elementary Classrooms', *American Educational Research Journal*, 51(2), pp. 328–362. doi: 10.3102/0002831213508299.

Phillips, K. J. R. (2010) 'What Does "Highly Qualified" Mean for Student Achievement? Evaluating the Relationships between Teacher Quality Indicators and At-Risk Students' Mathematics and Reading Achievement Gains in First Grade', *Elementary School Journal*, 110(4), pp. 464–493. doi: 10.1086/651192.

Platas, L. M., Ketterlin-Gellar, L., Brombacher, A. and Sitabkhan, Y. (2014) *Early Grade Mathematics Assessment (EGMA) Toolkit*. Research Triangle Park, NC.

Pontificia Universidad Católica del Ecuador (PUCE) (2012) *Informe Final: Levantamiento de Datos en Establecimientos Educacionales para Estudio sobre Prácticas Pedagógicas, Características y Actitudes del Docente y Mejoras en el Proceso Educativo del Ecuador*. Quito.

Psacharopoulos, G. and Patrinos, H. A. (2018) 'Returns to investment in education: a decennial review of the global literature', *Education Economics*, 26(5), pp. 445–458. doi: 10.1080/09645292.2018.1484426.

Rabe-Hesketh, S. and Skrondal, A. (2012) *Multilevel and longitudinal modeling using Stata*. Third. College Station, Texas: Stata Press Publication.

Rivkin, S. G., Hanushek, E. A. and Kain, J. F. (2005) 'Teachers, Schools, and Academic Achievement', *Econometrica*, 73(March, 2005), pp. 417–458. doi: 10.2307/3598793.

Rockoff, J. E. (2004) 'The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data', *The American Economic Review*, 94(2), pp. 247–252. doi: 10.1257/0002828041302244.

Rockoff, J. E., Jacob, B. A., Kane, T. J. and Staiger, D. O. (2011) 'Can You Recognize an Effective Teacher When You Recruit One?', *Education Finance and Policy*, 6(1), pp. 43–74. doi: 10.1162/EDFP_a_00022.

Rosenbaum, P. R. and Rubin, D. B. (1983) 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, 70(1), pp. 41–55. doi: 10.1093/biomet/70.1.41.

Rosenbaum, P. R. and Rubin, D. B. (1985) 'Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score', *The American Statistician*, 39(1), p. 38. doi: 10.2307/2683903.

Rosselli, M. and Ardila, A. (2003) 'The impact of culture and education on non-verbal neuropsychological measurements: A critical review', *Brain and Cognition*, 52(3), pp. 326–333. doi: 10.1016/S0278-2626(03)00170-2.

Rothstein, J. (2009) 'Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables', *Education Finance and Policy*, 4(4), pp. 537–571. doi: 10.1162/edfp.2009.4.4.537.

Rothstein, J. (2010) 'Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement', *The Quarterly Journal of Economics*, 125(1), pp. 175–214. doi: 10.1162/qjec.2010.125.1.175.

RTI International (2009a) *Early Grade Mathematics Assessment (EGMA): A Conceptual Framework Based on Mathematics Skills Development in Children*. Washington, DC: United States Agency for International Development. Available at: https://www.globalpartnership.org/content/early-grade-mathematics-assessment-egma-conceptual-framework-based-mathematics-skills (Accessed: 12 December 2017).

RTI International (2009b) *Early Grade Reading Assessment Toolkit*. Research Triangle Park, NC: RTI International. Available at: https://globalreadingnetwork.net/eddata/early-grade-reading-assessment-toolkit-2009 (Accessed: 12 December 2017).

Schady, N., Behrman, J. R., Araujo, M. C., Azuero, R., Bernal, R., Bravo, D., Lopez-Boo, F., Macours, K., Marshall, D., Paxson, C. and Vakis, R. (2015) 'Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries', *Journal of Human Resources*, 50(2), pp. 446–463. doi: 10.3368/jhr.50.2.446.

Schneider, B. R., Cevallos Estarellas, P. and Bruns, B. (2019) 'The Politics of Transforming Education in Ecuador: Confrontation and Continuity, 2006–17', *Comparative Education Review*, 63(2), pp. 259–280. doi: 10.1086/702609.

Schnittjer, I. and Fischer, L. (2018) 'NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 2 for Grade 1', *NEPS Survey Papers*. doi: 10.5157/NEPS:SP46:1.0.

Schnittjer, I. and Gerken, A.-L. (2018) *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 2 for Grade 2*, *NEPS Survey Papers*. 47. Bamberg. doi: 10.5157/NEPS:SP47:1.0.

Statistisches Bundesamt (2019) 'Bildung', in *Statistisches Jahrbuch 2019*. Statistisches Bundesamt, p. 95. Available at: https://www.destatis.de/DE/Themen/Querschnitt/Jahrbuch/jb-bildung.pdf?__blob=publicationFile.

Steinberg, M. P. and Donaldson, M. L. (2016) 'The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era', *Education Finance and Policy*, 11(3), pp. 340–359. doi: 10.1162/EDFP_a_00186.

Strøm, B. and Falch, T. (2020) 'The role of teacher quality in education production', in Bradley, S. and Green, C. (eds) *The Economics of Education: A Comprehensive Overview*. Second. Elsevier, pp. 307–319. doi: 10.1016/b978-0-12-815391-8.00022-7.

Todd, P. E. and Wolpin, K. I. (2003) 'On the Specification and Estimation of the Production Function for Cognitive Achievement', *The Economic Journal*, 113(485), pp. F3–F33. doi: 10.1111/1468-0297.00097.

VandenBos, G. R. (Ed. . (2007) *APA Dictionary of Psychology*. Washington, DC, US: American Psychological Association.

Warm, T. A. (1989) 'Weighted likelihood estimation of ability in item response theory', *Psychometrika*, 54(3), pp. 427–450. doi: 10.1007/BF02294627.

Wayne, A. and Youngs, P. (2003) 'Teacher Characteristics and Student Achievement Gains: A Review', *Review of Educational Research*, 73(1), pp. 89–122. doi: 10.3102/00346543073001089.

Wechsler, D. and Psychological Corporation (1997) *WAIS-III: Administration and Scoring Manual: Wechsler Adult Intelligence Scale*. Third. San Antonio, TX: Psychological Corporation.

World Bank (2019) *Education Statistics (EdStats)*, *The World Bank Group*. Available at: http://datatopics.worldbank.org/education/country/ecuador (Accessed: 12 December 2019).

World Bank (2021) *World Bank Country and Lending Groups*, *The World Bank Group*. Available at: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (Accessed: 7 June 2021).

Zielonka, M. and Pelz, S. (2015) *NEPS Technical Report: Implementation of the ISCED-97, CASMIN and Years of Education Classification Schemes in SUF Starting Cohort 2*. Bamberg. Available at: https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/3-0-0/TR_Derived_Educational_Variables_SC2.pdf (Accessed: 25 March 2021).