WORKING PAPER Nº IDB-WP-01677

Using Screening Assessments to Recruit Effective Teachers: Experimental Evidence from Ecuador

María Daniela Araujo P.
Yyannú Cruz-Aguayo
Guido Heineck

Inter-American Development Bank
Office of Strategic Planning and Development Effectiveness
Social Sector

July 2025

Using Screening Assessments to Recruit Effective Teachers: Experimental Evidence from Ecuador

María Daniela Araujo P.
Yyannú Cruz-Aguayo
Guido Heineck

Inter-American Development Bank
Office of Strategic Planning and Development Effectiveness
Social Sector

July 2025

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library
Araujo P., María Daniela.
Using screening assessments to recruit effective teachers: experimental evidence from Ecuador / María Daniela Araujo, Yyannu Cruz-Aguayo, Guido Heineck.
p. cm. — (IDB Working Paper Series ; 1677)
Includes bibliographical references.
1. Teacher effectiveness-Ecuador. 2. Teachers-Recruiting-Ecuador. 3. Education and state-Ecuador. 4. Human capital-Ecuador. I. Cruz Aguayo, Yyanú. II. Heineck, Guido. III. Inter-American Development Bank. Office of Strategic Planning and Development Effectiveness. IV. Inter-American Development Bank. Social Sector. V. Title. VI. Series.
IDB-WP-1677

http://www.iadb.org

Using Screening Assessments to Recruit Effective Teachers: Experimental Evidence from Ecuador

MARÍA DANIELA ARAUJO P., YYANNÚ CRUZ-AGUAYO, AND GUIDO HEINECK*

July 2025

Since 2007, the Ecuadorian government has required teacher candidates to pass cognitive and knowledge tests before they are allowed to participate in merit-based competitions for tenured positions. We evaluate this policy by linking administrative teacher information to data from an experimental study that randomly assigned nearly 13,000 children to their teachers. We find that test-screened tenured teachers had a significant effect of at least 10.5 percent of a SD on language learning outcomes. Although the recruitment tests screened candidates with higher cognitive skills, the classroom practice instrument used in the competitions appears to have helped identify the most effective teachers.

JEL Classification: I20, I21, I25, I28, J45

Keywords: teacher quality, teacher recruitment policy, educational policy, Latin America.

---

* **María Daniela Araujo P.,** corresponding author (mariaara@iadb.org); Office of Strategic Planning and Development Effectiveness (SPD), Inter-American Development Bank, and University of Bamberg; 1300 New York Avenue, N.W. Washington, D.C. 20577, USA. **Yyannú Cruz-Aguayo** (yyannuc@iadb.org); Social Sector (SCL), Inter-American Development Bank; 1300 New York Avenue, N.W. Washington, D.C. 20577, USA. **Guido Heineck** (guido.heineck@uni-bamberg.de); University of Bamberg, IZA Bonn and LIfBi Bamberg; Feldkirchenstr. 21, 96052 Bamberg, Germany. **Acknowledgments**: We would like to thank the Inter-American Development Bank and the Ecuadorian Ministry of Education for their crucial collaboration and for providing the data for this study. Part of the data used in this study was collected with the support of the Inter-American Development Bank's Technical Cooperations EC-T1236, EC-T1275, and EC-T1281. This research was supported by the Bamberg Graduate School of Social Sciences (BAGSS), which is funded by the German Research Foundation (DFG) under the German Excellence Initiative (GSC1024). We also thank Pedro Carneiro, Eric A. Hanushek, David Post, Maria Caridad Araujo, Gloria Vidal Illingworth, Susana Araujo Fiallos, and seminar participants at the IAB Workshop on Field Experiments in Policy Evaluation, 5th IZA Workshop: The Economics of Education, ifo-WZB Workshop on Causal Analyses of School Reforms, 13th Annual Meeting of Impact Evaluation Network within the RIDGE Forum, 2021 ASSA Annual Meeting, 34th Annual ESPE Conference, 11th International Workshop on Applied Economics of Education, XXIX Meeting of the Economics of Education Association, 2021 Annual Conference of the Verein für Socialpolitik, and the University of Bamberg for helpful comments and discussions. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.

I. Introduction

One of the most important challenges for policymakers and researchers in the field of education policy is how to identify and recruit effective teachers who contribute significantly to student learning outcomes. In high-income countries, teachers' cognitive skills and knowledge, as measured by certification tests, are among the observable characteristics widely used to signal teacher quality for recruitment. Although several studies have found that skills and certification test scores can be individual predictors of future teacher effectiveness, the evidence is inconclusive (Angrist and Guryan 2008; Clotfelter, Ladd, and Vigdor 2007; Goldhaber 2007; Goldhaber and Anthony 2007; Goldhaber, Gratz, and Theobald 2017; Harris and Sass 2009; Kane, Rockoff, and Staiger 2008; Rockoff et al. 2011; Hanushek, Piopiunik, and Wiederhold 2019). More recently, structured classroom observations have been introduced as an instrument to evaluate teacher performance (Steinberg and Donaldson 2016). Here experimental research has shown that classroom observation scores can serve as unbiased predictors of how much teachers will contribute to student achievement (Kane and Staiger 2012; Bacher-Hicks et al. 2019).

The focus of current research in teacher hiring and personnel economics has shifted from a basic analysis of how teacher effectiveness can be attributed to individual teacher characteristics to an evaluation of whether screening assessments conducted during recruitment and hiring processes can predict teacher effectiveness. Studies in the United States (US) have found that screening scores and, particularly, performance on specific assessments, such as knowledge tests, interviews, and demonstration lessons, can strongly predict future teacher effectiveness (Bruno and Strunk 2019; Goldhaber, Grout, and Huntington-Klein 2017; Jacob et al. 2018). Nevertheless, these studies are non-experimental, and their results may be biased by potential non-random matching of students to teachers within schools.[1]

The question of how to identify and recruit effective teachers is particularly important in middle- and low-income countries. In these countries, enrollment in primary and secondary education has significantly increased in recent decades, but their learning outcomes continue to fall well short of those of high-income countries (OECD 2023). Latin American countries have not traditionally had certification processes for the teaching profession or used screening assessments for recruitment to ensure teacher quality. However, over the past twenty years, several governments in the region have implemented teacher recruitment policies that screen candidates by assessing their cognitive skills, knowledge, and classroom practice in order to improve student learning outcomes. For example, Colombia (2002), Ecuador (2007), Mexico (2008), and Peru (2012) all implemented policies that required teacher candidates to pass national standardized tests before they could apply for permanent positions at public schools (Bruns and Luque 2015; Elacqua et al. 2017; Cruz-Aguayo, Hincapié, and Rodríguez 2020). Evidence on the effectiveness of teacher screening and hiring processes is also highly relevant to Latin American governments because personnel costs take up the majority of their education budgets. As a result, these governments rely almost exclusively on policies to improve the quality of their teachers as a means of improving student learning outcomes.

[1] Existing research suggests that students are not randomly assigned to teachers, which is likely to produce bias in teacher coefficient estimates (Clotfelter, Ladd, and Vigdor 2006; Rothstein 2010)

In this study, we assess the effectiveness of Ecuador's competitive teacher recruitment policy as a quality screening device. Specifically, we evaluate whether teachers who passed national entrance exams covering cognitive skills, pedagogical and content knowledge, followed by assessments of classroom practice, and who were tenured through Ecuador's competitive recruitment policy (whom we refer to as test-screened tenured teachers) have positive effects on student learning outcomes. We use data from unique administrative teacher records and link it to the rich experimental data produced by the "Closing Gaps" project. This project assigned a representative cohort of Ecuadorian kindergarten children to teachers in the 2012–2013 school year using a rule that is as good as random (Araujo et al. 2016).[2] We confirm that our data also reflects successful random assignment of these kindergarten children to test-screened tenured teachers, which allows us to estimate teacher causal effects and prevent the potential biases of selectively matching teachers to students. Our results show that kindergarten students who were randomly assigned to test-screened tenured teachers scored 10.5 percent of a standard deviation higher on end-of-year language tests. The evidence from our study suggests a weaker effect on math learning outcomes. Children randomly assigned to test-screened tenured teachers scored 8.5 percent of a standard deviation higher end-of-year math tests, but only at the 10 percent significance level. The stronger effect of test-screened tenured teachers on language learning could be explained by the fact that in Ecuador, kindergarten instruction is focused on developing socioemotional and language skills rather than mathematical skills.

In addition to having an experimental design, the "Closing Gaps" project collected rich information on students, families, and teachers, which allows us to check the robustness of our experimental estimations. We had access to not only common observable teacher characteristics—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). For language learning outcomes, the positive and significant effects of teachers tenured through the competitive recruitment policy persist even after controlling for the full set of teacher characteristics.

Our estimates are robust to several specifications. Although we first compare our sample of test-screened tenured teachers to all of their peer kindergarten teachers, we then test an alternative control group that accounts for teacher heterogeneity in hiring procedures and tenure status. We confirm that test-screened tenured teachers significantly outperform the pool of fixed-term contract teachers who have not passed the national entrance exams and who are working at the same schools. On average, kindergarten students taught by test-screened tenured teachers scored at least 14.8 percent of a standard deviation higher on end-of-year language tests and at least 12.9 percent of a standard deviation higher in math than students randomly assigned to contract teachers.

In contrast, we do not find similar effects for teachers who have passed entrance tests but have not yet won competitions for tenure and are therefore working on fixed-term contracts (whom we refer to as test-screened contract teachers). Our study suggests that performance in screening assessments, particularly the demonstration class, explains this difference. Test-

[2] In Ecuador, children start kindergarten at the age of five, which is considered the first year of general basic education.

screened tenured teachers and test-screened contract teachers have similar means and distributions of cognitive skills, as measured by the WAIS-III, which are significantly higher than those of contract teachers. However, test-screened tenured teachers have better classroom practices than test-screened contract teachers and contract teachers, as measured by the CLASS instrument. Accordingly, our study indicates that the recruitment tests successfully screened teacher candidates with higher cognitive skills; however, the classroom practice instrument appears to have helped identify the most effective teachers among the test-screened candidates. Although test-screened tenured teachers are significantly more experienced than test-screened contract teachers and contract teachers, experience does not appear to be the reason for the difference in effects.

This paper makes four key contributions to research on teacher effectiveness in the literature on the economics of education and on labor and personnel economics. First, we provide the first experimental estimates of the effects of teachers screened through skill, subject knowledge, and pedagogical practice assessments and tenure in merit-based competitions in a Latin American country. To date, impact evaluations of these teacher recruitment policies in Colombia (Brutti and Sánchez Torres 2022; Busso et al. 2024), Ecuador (Cruz-Aguayo, Ibarrarán, and Schady 2017; Araujo P. 2019), Mexico (Estrada 2019; Bedoya, De Hoyos, and Estrada 2023), and Peru (Bertoni et al. 2023) have been correlational or quasi-experimental studies that are undermined by the potential bias caused by matching students to teachers within or across schools. Our study overcomes this fundamental concern. Its results align with the findings of Brutti and Sánchez Torres (2022) for Colombia and Estrada (2019) for Mexico, whose estimates show that test- and merit-based teacher hiring has a positive and significant effect on student achievement at the school level. In contrast, our results differ from the null effects found in previous nonexperimental studies for Ecuador (Cruz-Aguayo, Ibarrarán, and Schady 2017; Araujo P. 2019), which use value-added models with merely three months between students' baseline and final tests. Our results also differ from the negative effects found in Colombia by Busso et al. (2024), who show that the reform decreased student achievement in public schools because many experienced contract teachers were replaced by novice teachers with high entrance exam scores. This difference in results is not surprising for three reasons. First, we show that in the Ecuadorian case, test-screened tenured teachers are not novices and have significantly more years of experience than the contract teachers who remained at the schools. Second, we find no significant association between teacher experience and student learning outcomes once classroom practices (CLASS) are taken into account. Third, the Busso et al. (2024) study may confound the effects of teacher experience and classroom practices because it does not observe classroom practices.

Our findings also complement those of Bruno and Strunk (2019), Jacob et al. (2018), and Goldhaber, Grout, and Huntington-Klein (2017) for the US. These studies have shown that screening instruments and measures in recruitment processes can predict future teacher effectiveness. Our advantage is that we are able to reach a similar conclusion in an experimental setting.

Second, we contribute to the ongoing debate about which teacher characteristics are associated with student learning outcomes in middle- and low-income countries. The evidence we present suggests that the mediator of the test-screened tenured teacher effect is classroom practices, which the competitive recruitment process assessed using a demonstration class instrument. Furthermore, we find that teachers' overall test scores on screening assessments—

which consisted of a logical-verbal reasoning test, a pedagogical knowledge test, and a content knowledge test—are positively and statistically significantly associated with students' language learning outcomes. Our conclusions are consistent with previous findings in developing countries that positively associate learning effects with teachers' classroom practices (Araujo et al. 2016), and with findings on the impact of teachers' subject knowledge on student achievement (Bau and Das 2020; Bertoni et al. 2023; Bietenbeck, Piopiunik, and Wiederhold 2018; Glewwe et al. 2020; Metzler and Woessmann 2012)

Third, our results provide new evidence on the link between teacher qualifications, job status, and performance in middle- and low-income countries. Contrary to existing evidence that contract teachers have a positive effect on learning outcomes (Duflo, Dupas, and Kremer 2015; Muralidharan and Sundararaman 2013), our estimates show that in Ecuador, teachers who are recruited and tenured through an assessment-based screening process significantly outperform contract teachers. Duflo, Dupas, and Kremer's (2015) randomized evaluation of the Extra Teacher Program (ETP) in Kenya finds that fully qualified teachers who were offered short-term, low-pay contracts in ETP schools exerted greater effort and improved student learning outcomes more effectively than civil service teachers. Nonetheless, Duflo, Dupas, and Kremer (2015) and the subsequent work of Bold et al (2018), which shows that an ETP scale-up led by the national government had no impact on student achievement, also highlight that the effectiveness of hiring contract teachers is sensitive to the institutional context. In the case of Kenya, the national teachers' union pressured the government to guarantee permanent civil service employment to the new contract teachers participating in the scaled-up ETP, without any quality or performance screening (Duflo, Dupas, and Kremer 2015; Bold et al. 2018). Ecuador had the opposite institutional and policy context. After a long freeze on permanent hiring, the Ecuadorian government opened thousands of permanent teacher positions through transparent, assessment- and merit-based public competitions, and it reduced any discretionary power exercised by the teachers' union. As a result, our evidence shows that not only did the most qualified candidates obtain tenure or civil service positions, but also that these teachers significantly outperformed the remaining contract teachers in the system. In this regard, our study also adds to concerns about the external validity of experimental evaluations of policies and programs in a specific country context and about generalizing them to other middle- or low-income countries (Pritchett and Sandefur 2015; Hanushek 2021; Williams 2020).

Finally, our study confirms the potential effectiveness of better-qualified teachers in closing early learning gaps between children of different abilities and socioeconomic backgrounds in a Latin American country. We find that test-screened tenured teachers helped close the learning gap for children who already started their first year of school with significantly lower language development, came from poorer households, and had parents with fewer years of education. In addition, we show that the effect of test-screened tenured teachers on language is significantly stronger for children with lower executive function levels, meaning that these teachers were more effective for children with lower self-control, attention, and, consequently, ability to learn. These findings are particularly important for Ecuador and other Latin American countries, given

the large and persistent differences in cognitive development between children of high and low socioeconomic status in the region (Schady et al. 2015).[3]

The remainder of the paper is structured as follows. In Section 2, we provide background on the Ecuadorian education system and the competitive teacher recruitment policy. In Section 3, we provide an overview of our experimental design and data. We present our model, estimation strategy, and results in Section 4, and in Section 5, we describe a heterogeneous effects analysis. We present potential mechanisms in Section 6 and share our conclusions in Section 7.

II. Background and Evidence

A. Ecuador's Education System

Ecuador is a middle-income country in South American with compulsory schooling from ages 5 to 15. The Ecuadorian education system is organized into three levels: early education, general basic education, and high school (Asamblea Nacional del Ecuador 2011). Early education serves children under age 5 (equivalent to ISCED[4] level 0). General basic education starts at age 5 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 consists of three years of upper secondary education (ISCED level 3).

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 that these reforms successfully expanded education: while net enrollment in kindergarten and primary school remained above 90 percent, net enrollment in secondary school increased from 62 percent in 2005 to 85 percent in 2017 (World Bank 2019). Nonetheless, questions about the reforms' impact on the system's quality remain. These reforms particularly targeted the teaching profession, the prestige and quality of which 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).

B. Competitive Teacher Recruitment Policy in Ecuador

In November 2007, the Ecuadorian government issued Executive Order No. 708, which required teacher candidates to pass mandatory national skill and content knowledge tests before they were allowed to participate in merit-based selection competitions for tenure at public schools.[5] The new regulation was applied exclusively to teachers seeking tenure from

---

[3] Schady et al. (2015) find large and statistically significant differences in language development between richer and poorer children in five Latin American countries (Chile, Colombia, Ecuador, Peru, and Nicaragua). The authors highlight that differences across income quartiles are largest in urban Colombia (1.23 standard deviations) and rural Ecuador (1.21 standard deviations).

[4] ISCED stands for International Standard Classification of Education.

[5] The Executive Order No. 708 reformed the Implementing Regulations of 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 Executive Order No. 708 was issued in 2007.

December 2007 onwards. [6] Teachers who already had tenure were exempted from the regulation. [7] The policy reform was institutionalized in 2011, when Ecuador's new Intercultural Education Law (*Ley Orgánica de Educación Intercultural, LOEI*) ratified national entrance exams as a mandatory requirement for teacher candidates (Asamblea Nacional del Ecuador 2011).

Prior to December 2007, teacher recruitment processes were locally organized by Provincial Education Directorates with no national standards other than academic degree requirements. Competitions for permanent or tenured teacher positions were based on academic qualifications, experience, and professional training. Teachers were not required to pass cognitive, content-knowledge, or classroom practice assessments. In general, a public competition was opened to fill a permanent vacancy when a teacher retired. School expansion, particularly for non-universally covered levels, was carried out with fixed-term contract teachers. [8] Schools and Provincial Directorates had a great deal of discretion in selecting contract teachers. Moreover, the national teachers' union influenced teacher selection in Ecuador in non-transparent ways (Bruns and Luque, 2015; Schneider, Cevallos Estarellas and Bruns, 2019).

Since December 2007, Ecuador's Ministry of Education has organized centralized processes to recruit permanent teachers. The Ministry has regulated these recruitment processes through Ministerial Resolutions (*Acuerdos Ministeriales*, *AM*). The first three processes were organized into three stages or components (Ministerio de Educación del Ecuador 2007; 2008; 2010; 2011a). In the first stage, teacher candidates had to pass a logical-verbal reasoning test, a pedagogical knowledge test, and a subject-specific knowledge test. Teacher candidates who achieved the required minimum score became eligible to compete for a tenured position. These test scores were part of each candidate's overall competition score. The second stage of the recruitment process involved evaluating eligible candidates' credentials: academic degrees, teaching experience, additional training courses, and academic publications. Credentials were graded and added to the competition score. Finally, teachers were also required to give a demonstration class before a school board, which was also evaluated and added to the total competition score. [9] When the policy was first implemented, the demonstration class was part of the first stage, but afterward it became the third stage. In each competition, the candidate with the highest overall score earned a tenured position. The different stages of the recruitment process were highly automated and transparent. [10]

---

[6] In Ecuador, tenured teachers are civil servants who hold permanent teaching positions at public schools.

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

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

[9] A school board consisted of the school principal or the assistant principal, a teacher, and two parents elected by the school's general assembly. For positions in lower and upper secondary education, a student was also included.

[10] Teacher candidates were required to open an account on the Ministry's online recruitment platform to register for eligibility tests. In the first stage, standardized exams were scanned and scored using a specialized software. Afterwards, test scores were automatically uploaded and published on the Ministry's platform. In the second stage,

The Ministry of Education changed the way the recruitment process stages were weighted slightly between 2007 and 2013, but it always gave test scores the most weight, increasing from 45 to 55 percent of the total competition score. Table 1 describes the ministerial resolutions applied from the beginning of the process to the first semester of 2013, during which three competition rounds were held.[11] Our analysis focuses on this period because it covers all possible recruitment processes for teachers tenured after December 2007 and employed in the 2012–2013 school year.

TABLE 1. REGULATIONS AND COMPONENTS OF ECUADOR'S COMPETITIVE TEACHER RECRUITMENT

	Ministerial Resolutions					
	AM No. 438-07 December 2007 (First-round)		**AM No. 018-10 January 2010** (Second-round)		**AM No. 379-11 November 2011** (Third-round)	
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% on eligibility tests		– 60% on eligibility tests		– 60% on logical-verbal and pedagogical knowledge tests – 70% on subject-specific knowledge test	
Issued	Dec-07		Jan-10		Nov-11	
Abolished	Jan-10		Nov-11		May-13	

Source: Prepared by the authors based on Araujo P. (2019), and 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.

eligible candidates could apply for a permanent position from among all open vacancies in public schools through the Ministry's platform. They were also required to upload their credentials and certificates. Although staff at Provincial Education Directorates were responsible for verifying the accuracy of the information and certificates, the Ministry's platform included a software program that scored candidates' credentials. In the third stage, school boards were required to use a standardized instrument provided by the Ministry to assess the demonstration classes. Assessment sheets were scanned, automatically scored, and uploaded to the Ministry's platform. Teacher candidates were able to follow the recruitment process at each stage on the platform and had access to the scores of all candidates participating in their specific competitions. Moreover, teacher candidates had the right to appeal their results at any stage of the process.

[11] 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 for Educational Assessment (Instituto National de Evaluación Educativa, INEVAL).

Although the new policy required permanent teacher vacancies to be filled through the merit-based process described above, local education authorities were still allowed to hire teachers who had not undergone the new recruitment process on fixed-term contracts until tenured teachers permanently filled the positions. Local authorities still had significant discretion in selecting contract teachers. However, the Ministry requested that schools and Provincial Directorates give preference to eligible teachers who had already passed mandatory tests for contracts.

Alongside the new recruitment regulation, the Ecuadorian government opened approximately 34,000 new permanent teacher positions between 2007 and 2012 (Ministerio de Educación del Ecuador 2012). It also introduced strong economic incentives to attract highly competitive teacher candidates to the public education system. The nominal monthly starting salary for a new tenured teacher steadily increased from US$291 in 2006 to US$396 in 2010 and then rose sharply to US$775 in 2011, when Ecuador's new Intercultural Education Law brought the teacher pay scale into line with the civil service pay scale (Ministerio de Educación del Ecuador 2012; Schneider, Cevallos Estarellas, and Bruns 2019).[12] By contrast, contract teachers' salaries stagnated at around US$300 per month during the same period. Competing for a tenured position was thus very attractive to contract teachers already 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.[13]

The first years of the new Ecuadorian competitive teacher recruitment process were highly competitive. Between 2007 and 2012, 320,000 candidates registered for eligibility tests,[14] 21,200 eligible candidates passed entrance tests, and 18,820 successful candidates were granted permanent teaching positions (Ministerio de Educación del Ecuador 2012).

III. Experimental Design and Data

A. Background on the "Closing Gaps" Project

The Inter-American Development Bank (IDB), in cooperation with Ecuador's Ministry of Education, launched the "Closing Gaps" project in 2011. This longitudinal experimental study evaluated different dimensions of teacher quality in public schools, starting with the first year of general basic education, which is kindergarten.

[12] Teaching wage increases were not the product of high inflation rates. Ecuador's average inflation rate between 2006 and 2011 was 4.37 percent (Instituto Ecuatoriano de Estadísticas y Censos (INEC) 2020).

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

[14] Teachers were allowed to register for more than one subject-specific eligibility test because teacher training programs in Ecuador allow for more than one specialization. Also, teachers were allowed to re-register for an exam if they had not achieved the minimum passing score. Therefore, the actual number of teacher candidates was lower than the 320,000 registrations for eligibility tests recorded in this period.

The "Closing Gaps" Project randomly chose a sample of 204 public schools from the coastal region of Ecuador.[15] Although it drew its study sample from the coastal region, the children and households in the study were generally similar to national samples (Araujo et al. 2016). The sample was limited to schools with 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 classrooms and teachers. Children enrolled for kindergarten at a given school were ordered by their last and first names and then assigned to kindergarten classrooms by 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 they had not been assigned to (Araujo et al. 2016). [16]

Araujo et al. (2016) were the first to use the "Closing Gaps" data to examine how teacher quality impacts learning outcomes in kindergarten. Their study finds teacher effects of 9 percent of a standard deviation in language and math learning. The results also suggest that children assigned to teachers with higher classroom practice (CLASS) scores have significantly higher achievement. A teacher score that is one standard deviation higher is associated with end-of-year student test scores that are between 6 and 8 percent of a standard deviation higher. By contrast, children assigned to inexperienced teachers had test scores that were 17 percent of a standard deviation 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. Our sample comprises 202 schools and all their kindergarten classrooms.[17]

B. Data on Children and Families

At the beginning of the 2012–2013 school year, the "Closing Gaps" project collected baseline data on children's age, gender, and early childhood education attendance. Children also took the *Test de Vocabulario en Imagenes Peabody (TVIP)* to measure their past learning.[18] The project also ran a household survey with questions on parents' education and living conditions.

---

[15] Ecuador has four natural regions: Coastal (Costa), Andean (Sierra), Amazon (Amazonía), and Island (Galapagos Islands). Due to the unique weather conditions of each region, their school year starts in different months. The 2012–2013 school year started in April 2012 and ended in February 2013 in the Coastal and Island regions. The same school year in the Andean and Amazon Regions started in September 2012 and ended in June 2013.

[16] In our analysis, non-compliers are assigned to the classrooms to which they were originally randomly assigned. This means that we 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.

[17] We lost two schools from the original sample of 204 schools because we did not have complete information on teacher characteristics for these schools.

[18] This is the Spanish version of the Peabody Picture Vocabulary Test (PPVT) (Dunn et al. 1986). The "Closing Gaps" project provided TVIP test scores standardized based 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 observations were given the lowest or highest possible value if the child's raw score had no correspondence to the standardized score in the reference population.

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.[19]

To measure student learning outcomes at the end of the school year, four language and early literacy tests and four math tests were applied to children individually.[20] The language and early literacy tests covered child vocabulary, oral comprehension, and sound, letter, and word recognition.[21] The math tests covered number recognition, sequencing, applied math problems, and identifying basic geometric figures.[22] Test aggregates for language and math were normalized to have a zero mean and a unit standard deviation.

As indicated, 14,930 children were enrolled in kindergarten in the participating schools, but 740 children did not start the academic year at 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 we take additional information on teacher characteristics into account in further analyses.

Table 2 reports summary statistics for child and family characteristics for the full and analysis samples. Children in the analysis sample were approximately five years old at the beginning of the school year. Girls account for almost half of the sample, and about 61 percent of the children attended early childhood education. The average TVIP score is approximately 83, one standard deviation lower than the Mexican and Puerto Rican reference populations used to norm the test. Parents had completed almost nine years of education on average. In addition, column (5) of Table 2 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. The sole exception is the proportion of children who attended early childhood education, which is about five percentage points higher in the analysis sample.

---

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

[20] The “Closing Gaps” project also tested children’s inhibitory control, working memory, capacity to pay attention, and cognitive flexibility. These processes are jointly known as executive function. Our study found no effects on children’s executive function.

[21] The “Closing Gaps” project applied the TVIP to evaluate child vocabulary. Oral comprehension, and sound, letter and word recognition tests were taken from the Spanish 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)

[22] All of the math tests were taken from the Spanish 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 2. SUMMARY STATISTICS FOR CHARACTERISTICS OF CHILDREN AND FAMILIES

	Full sample		Analysis sample		Difference (p-value)
	Mean (SD) (1)	Obs. (2)	Mean (SD) (3)	Obs. (4)	(5)
Children:					
Proportion female	0.49	14930	0.49	12632	0.00
	(0.50)		(0.50)		(1.00)
Proportion who attended early childhood 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)

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

C. Data on Teachers

At the beginning of the school year, the "Closing Gaps" project also collected data on conventional teacher characteristics such as gender, experience, and education. It collected additional rich, non-conventional data on teachers' cognitive ability and personality by the end of the school year. The project applied the Spanish version of the Wechsler Adult Intelligence Scale (WAIS-III) to measure teacher cognitive skills. It also used the NEO PI-R psychometric instrument (Costa Jr. and McCrae 2008) to assess teacher personality by measuring the so-called Big Five personality traits of neuroticism, extraversion, openness, agreeableness, and conscientiousness.[23] On 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 and the Big Five personality trait scores to have a zero mean and a unit standard deviation for our regression analyses.

Finally, to retrieve information on teacher behaviors and classroom practices, the "Closing Gaps" project applied the CLASS instrument in the middle of the school year. CLASS measures teacher behaviors in three domains: emotional support, classroom organization, and

[23] 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. Research in psychology and economics has firmly established that standardized tests of cognitive ability and personality traits predict a variety of job performance outcomes across professions (Almlund et al. 2011; Heineck and Anger 2010; Kuncel, Ones, and Sackett 2010).

instructional support (Hamre, La Paro, and Pianta 2007). Each domain's score can be low (1–2), 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 2011–2012 and 2012–2013 school years on a day not known to the teachers in advance. The videos were coded by experts according to the CLASS protocol to calculate scores for each domain. 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 from 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. In particular, and in response to our request, the National Directorate of Educational Research of Ecuador's Ministry of Education merged data from the "Closing Gaps" project with administrative records from teacher recruitment processes that took place between 2007 and 2013.[24] In our sample, about 13 percent of teachers had passed national entry tests and won a merit-based competition organized by the Ministry of Education between 2007 and 2013, which granted them tenure.

Table 3 reports summary statistics for teacher characteristics. Virtually all teachers in the analysis sample are female,[25] and 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).[26] 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. The average CLASS 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 although teachers maintained positive relations with their students and ran moderately well-organized classrooms, they engaged in very few interactions that support learning. We have full information on 430 teachers for the 2012–2013 school year.[27] For the 2011–2012 school year, we have only 326 CLASS score observations because of changes in the allocation of teacher staff within schools and teachers leaving the participating schools.

[24] These records are stored on the SIME (*Sistema de Información del Ministerio de Educación*) platform. Staff from the National Directorate of Educational Research used the teacher's personal ID number as a unique identifier in the merging process.

[25] In Ecuador, teachers in the lower grades are traditionally women. Nonetheless, as a robustness check, we ran all our analyses for the female teacher sample only. The results were consistent and are available upon request.

[26] Teachers' average cognitive skill score is not a source of concern for this study. Vast neuropsychological literature has shown cultural differences in cognitive test performance. The performance of people from different cultures on the same cognitive test may vary according to the importance of the specific cognitive ability in one's own culture (Fasfous et al. 2013; Bakos et al. 2010).

[27] We start our teacher sample with 450 observations, but we only have 430 observations for teachers' cognitive skills and personality trait scores. Table A1 in Appendix A shows that the differences in observable characteristics between the teacher analysis sample and the teacher dropout sample are not statistically significant, with the sole exception of tenure status at the 10 percent significance level. This evidence helps us rule out a problem of teacher selectivity.

TABLE 3. SUMMARY STATISTICS FOR TEACHER CHARACTERISTICS

	Analysis Sample Mean (SD)	Test-screened tenured teacher Mean (SD)		Diff (p-value)	Conventionally tenured Mean (SD)	Test-screened contract Mean (SD)	Contract Mean (SD)	Difference to test-screened tenured teachers (p-value)		
		YES	NO					Conventionally tenured	Test-screened contract	Contract
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Proportion female	0.99	0.96	0.99	-0.03	0.99	1.00	0.99	0.03	0.04	0.03
	(0.10)	(0.19)	(0.07)	(0.23)	(0.07)	(0.00)	(0.09)	(0.23)	(0.16)	(0.28)
Proportion tenured	0.58	1.00	0.52	0.48***	1.00	0.00	0.00	0.00	-1.00	-1.00
	(0.49)	(0.00)	(0.50)	(0.00)	(0.00)	(0.00)	(0.00)	(.)	(.)	(.)
Years of experience	14.74	12.84	15.01	-2.17**	20.19	9.48	9.34	7.35***	-3.36***	-3.51***
	(8.60)	(7.08)	(8.78)	(0.04)	(8.19)	(4.62)	(5.41)	(0.00)	(0.00)	(0.00)
Years of education	17.15	17.46	17.11	0.36	17.36	16.94	16.79	-0.11	-0.52	-0.67**
	(1.92)	(1.90)	(1.92)	(0.20)	(2.13)	(2.10)	(1.41)	(0.73)	(0.19)	(0.02)
University degree	0.99	1.00	0.99	0.01**	0.98	0.98	0.99	-0.02*	-0.02	-0.01
	(0.11)	(0.00)	(0.11)	(0.03)	(0.12)	(0.14)	(0.09)	(0.08)	(0.32)	(0.32)
Cognitive skills	86.46	89.93	85.96	3.96***	87.73	88.86	82.18	-2.20	-1.07	-7.74***
	(9.53)	(9.73)	(9.40)	(0.01)	(9.32)	(8.78)	(8.61)	(0.14)	(0.56)	(0.00)
Neuroticism	43.85	43.22	43.94	-0.73	43.16	44.89	44.76	-0.06	1.67	1.54
	(6.72)	(6.80)	(6.72)	(0.47)	(6.70)	(7.88)	(6.14)	(0.96)	(0.25)	(0.15)
Extraversion	45.65	48.55	45.24	3.31***	44.42	47.12	45.74	-4.12***	-1.43	-2.80***
	(6.83)	(6.46)	(6.79)	(0.00)	(6.91)	(6.49)	(6.58)	(0.00)	(0.26)	(0.01)
Openness	50.82	52.97	50.51	2.46**	51.00	50.91	49.60	-1.96*	-2.06	-3.37***
	(6.75)	(6.52)	(6.73)	(0.01)	(6.41)	(6.90)	(7.09)	(0.05)	(0.12)	(0.00)
Agreeableness	48.22	50.63	47.87	2.76***	48.25	49.07	46.84	-2.38**	-1.57	-3.79***
	(7.56)	(7.02)	(7.58)	(0.01)	(7.60)	(6.10)	(7.99)	(0.03)	(0.23)	(0.00)
Conscientiousness	57.55	58.64	57.39	1.25	57.26	58.36	57.22	-1.38	-0.28	-1.42
	(8.15)	(6.99)	(8.30)	(0.24)	(8.31)	(8.08)	(8.41)	(0.22)	(0.85)	(0.24)
CLASS average	3.48	3.50	3.48	0.02	3.58	3.41	3.36	0.08*	-0.09*	-0.14***
	(0.28)	(0.27)	(0.28)	(0.62)	(0.27)	(0.26)	(0.26)	(0.06)	(0.08)	(0.00)
CLASS 2011–2012	3.62	3.80	3.61	0.19**	3.71	3.61	3.42	-0.09	-0.19*	-0.38***
	(0.37)	(0.35)	(0.37)	(0.03)	(0.35)	(0.31)	(0.36)	(0.27)	(0.07)	(0.00)
Socio emotional support	4.30	4.48	4.29	0.19*	4.38	4.25	4.12	-0.10	-0.22*	-0.36***
	(0.40)	(0.44)	(0.39)	(0.07)	(0.38)	(0.34)	(0.38)	(0.33)	(0.07)	(0.00)
Classroom management	4.99	5.26	4.97	0.29**	5.14	4.97	4.65	-0.12	-0.29*	-0.61***
	(0.63)	(0.55)	(0.63)	(0.03)	(0.56)	(0.52)	(0.67)	(0.34)	(0.08)	(0.00)
Instructional support	1.36	1.44	1.36	0.08	1.39	1.40	1.27	-0.04	-0.04	-0.17**
	(0.27)	(0.26)	(0.27)	(0.18)	(0.27)	(0.25)	(0.25)	(0.48)	(0.65)	(0.01)
CLASS 2012–2013	3.41	3.45	3.41	0.04	3.47	3.36	3.32	0.02	-0.09*	-0.13***
	(0.28)	(0.25)	(0.29)	(0.24)	(0.29)	(0.25)	(0.27)	(0.55)	(0.06)	(0.00)
Socio emotional support	4.07	4.11	4.07	0.05	4.13	4.01	3.98	0.02	-0.10*	-0.13***
	(0.33)	(0.25)	(0.34)	(0.23)	(0.35)	(0.29)	(0.31)	(0.59)	(0.06)	(0.00)
Classroom management	4.79	4.87	4.78	0.09	4.87	4.71	4.68	0.00	-0.16*	-0.19**
	(0.47)	(0.46)	(0.47)	(0.21)	(0.46)	(0.42)	(0.48)	(0.95)	(0.07)	(0.01)
Instructional support	1.15	1.15	1.15	-0.00	1.19	1.13	1.09	0.04*	-0.02	-0.06**
	(0.18)	(0.15)	(0.18)	(0.96)	(0.21)	(0.16)	(0.12)	(0.09)	(0.54)	(0.02)
N	430	54	376	430	196	50	130			

Source: Authors' calculations. *Notes:* This table reports means and standard deviations (in parenthesis) of teacher characteristics. 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. Columns (8)–(10) display differences between test-screened tenured teachers, and the groups of conventionally tenured teachers, test-screened contract teachers, and contract teachers, as well as 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 (1–2), medium (3–5) or high (6–7). CLASS 2011–2012 statistics are based on 326 observations. * Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level.

In addition, columns (2) and (3) of Table 3 display summary statistics for the analysis sample according to whether the teacher passed the national entry tests and won a competition for

tenure. Column (4) shows mean differences between the groups. There are some statistical differences in characteristics 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. On average, they have higher cognitive skills 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 also have slightly more years of education on average and are more conscientious and less neurotic, although these differences are not statistically significant in our sample.

We first compared our sample of test-screened tenured teachers to all of their kindergarten teacher peers. However, our comparison group is rather heterogeneous since it includes teachers who were tenured by local authorities before 2007 (which we refer to as conventionally tenured teachers), contract teachers who had passed national entrance exams but had not yet won a competition for tenure (test-screened contract teachers), and contract teachers who had not passed national entrance exams or any standardized screening process.[28] Table 3 presents descriptive statistics for conventionally tenured teachers (45 percent of the full sample) in column (5), test-screened contract teachers (12 percent of the full sample) in column (6), and contract teachers (30 percent of the full sample) in column (7), as well as differences between these groups and the sample of test-screened tenured teachers in columns (8) to (10). We find that conventionally tenured teachers have significantly more years of experience and slightly higher average CLASS scores than test-screened tenured teachers. By contrast, conventionally tenured teachers have lower cognitive skills and personality trait scores, with statistically significant differences for extraversion, openness, and agreeableness. Test-screened contract teachers are similar to test-screened tenured teachers in cognitive skills and personality traits but have significantly fewer years of experience and lower CLASS scores.

Finally, contract teachers have the lowest averages among all teachers for most qualifications and characteristics, including years of experience, cognitive skills, personality traits, and CLASS scores. These values are remarkably lower in magnitude and statistical significance than the averages for the test-screened tenured teachers. The average cognitive skills score for contract teachers is especially low, at almost one standard deviation below the score of test-screened tenured teachers.

At first glance, the descriptive statistics show that the competitive teacher recruitment processes implemented by Ecuador's Ministry of Education did successfully screen the candidates with the highest skills and qualifications from the pool of non-tenured teachers. [29]

---

[28] From the Closing-Gaps' teacher survey data, we deduce that 63 percent of the contract teachers who had not passed national entry exams, actually took these exams but failed them (they did not achieve the minimum eligibility score). The remaining 37 percent of these contract teachers had not taken the exams by the time the survey was applied (2012-2013 school year).

[29] Because our estimation strategy relies on within-school variation, we also present within-school mean differences in the characteristics and qualifications of test-screened tenured teachers relative to their peers in Table A2. Most schools in the sample have only two kindergarten teachers, so finding statistically significant differences at the teacher level within schools is problematic. However, we do observe patterns that are similar to those in the overall descriptive statistics. We still observe that test-screened tenured teachers have higher CLASS scores than test-screened contract teachers and contract teachers. Overall, we also find that contract teachers exhibit lower qualifications and skills than test-screened tenured teachers (panel D).

D. Validity of the Experimental Design

For the causal inferences we aim to establish to be valid, students must be successfully randomized into classrooms and to teachers within schools. To evaluate whether random assignment to test-screened tenured teachers was successful, we test for balance in predetermined variables across classrooms. For this purpose, we regress teacher test-screened tenured status on predetermined student and family variables. We condition our estimations on school fixed effects because students were randomly assigned to teachers within schools.

TABLE 4. RANDOMIZATION TEST

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

Source: Authors' calculations. *Notes:* OLS model estimated with standard errors (in parentheses) clustered 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). The total number of observations is restricted to the student sample with full information on student, parent, classroom, and teacher characteristics. * Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level

Table 4 provides the results of our randomization test. As shown, none of the student or family characteristics predict the likelihood of a child being assigned to a test-screened tenured teacher, except for parents' years of schooling. However, this variable's regression coefficient is close to zero. Moreover, an F-test for the joint significance of all predetermined demographic variables is statistically insignificant (p=0.372). Consequently, we conclude that random assignment to test-screened tenured teachers was successful and that there is no threat to our identification strategy.

In addition, Table A3 in Appendix A presents the results of our randomization test, taking into account the different types of teachers in the comparison group. Our estimations show that, within schools, student and family background characteristics do not significantly explain whether a child is likely to be assigned to a test-screened tenured teacher relative to a test-screened contract teacher or contract teacher. In the comparison of assignment to a test-screened tenured teacher and assignment to a conventionally tenured teacher, there are differences in parents' years of schooling and attendance of early childhood education, but only at the 10 percent significance level and with regression coefficients of close to zero. Again, we are confident that random assignment to test-screened tenured teachers was successful.

We also look at 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 children who were enrolled later (5 percent). We evaluate whether being a no-show, dropout, or late enrollment correlates to 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. We present our results in Table A4 and show that correlations between no-shows, dropouts, or late enrollments and being assigned to a test-screened tenured teacher are statistically insignificant and virtually equal to zero. Additionally, in Table A5 we show that being a no-show, dropout, or late enrollment was not affected by random assignment to a conventionally tenured teacher, a test-screened contract teacher, or to a contract teacher, relative to a test-screened tenured teacher.

IV. Estimation Strategy and Results

A. Estimated Effects of Test-Screened Tenured Teachers

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:

$$Y_{ics} = \alpha_s + test\_tenured_{cs}\beta_1 + X_{ics}\beta_2 + \bar{X}_{cs}\,\beta_3 + C_{cs}\beta_4 + T_{cs}\beta_5 + P_{cs}\beta_6 + CLASS_{cs}\beta_7 + u_{ics}, \quad (1)$$

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 students were randomly assigned to teachers 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 make the estimates more precise. We progressively add the following controls: a vector of observable student and parent characteristics (X_{ics}), which includes a child baseline TVIP score; a vector of classroom averages for student and parent characteristics ($\bar{X}_{cs}$); and a class size indicator (C_{cs}).[30] Finally, we include a vector of teacher observable characteristics (T_{cs}), a vector of teacher cognitive ability and personality traits (P_{cs}), and an indicator of teacher classroom practices ($CLASS_{cs}$) to observe whether the test-screened tenured teacher effect holds or lessens.

The fact that teacher cognitive skills and personality measurements were collected by the end of the school year could raise concern about reverse causality, if one assumes that these characteristics could be affected by the classroom environment. However, it should be noted that we are not trying to find the causal effect of teacher cognitive skills and personality on

[30] Student and parent characteristics included: children's baseline TVIP score, age, gender, and early childhood education attendance, as well as parents' average years of education and living standard indicator.

student learning. Rather, we add these covariates at the end of our estimations to observe whether they affect our experimental results. In addition, despite the evidence that people's personalities can change over their life cycle, most research identifies the source of these changes as a combination of factors—with genetics playing a key role—rather than single environmental events (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 average together the CLASS scores from 2011–2012 with those form 2012–2013 to deal with this problem, as suggested by Araujo et al. (2016).[31] Furthermore, we add our CLASS covariate at the end of our estimations to check whether our results 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 gives a more conservative estimate of standard errors that takes into account cross-classroom correlations in errors within schools (Chetty et al. 2011).[32]

We report regression results for language learning outcomes in Table 5.[33] Column (1) presents the effect of test-screened tenured teachers on language learning outcomes estimated without additional covariates and considering school fixed effects. We find a positive causal effect of test-screened tenured teachers on language learning at the 5 percent significance level. Children randomly assigned to test-screened tenured teachers scored 10.5 percent of a standard deviation higher on the end-of-year language tests. We present regression results that incorporate controls for student, family, and classroom characteristics in column (2). Here, the effect remains significant, with a slight decrease in size. Subsequently, we explore whether the test-screened tenured teacher effect is still present once other teacher characteristics and qualifications are considered. In columns (3) and (4), we include the teacher background characteristics of gender, experience,[34] education, and tenure status,[35] as observed at the beginning of the school year. Here, the significance or size of the results does not change substantially. Finally, in column (5) we add controls for teacher cognitive ability, personality, and teacher classroom practices (CLASS score). We still find a significant, positive effect of test-screened tenured teachers, with end-of-year language test scores that are around 10 percent of

[31] Our findings do not change if we run the entire analysis with the lagged 2011–2012 CLASS score instead of the average CLASS score. Results for math are even stronger in significance level when we use the lagged CLASS score 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.

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

[33] 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. 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.

[34] In an alternative estimation, we model experience as nonlinear by including the square of experience in equation (1); however, the results do not change. In addition, we mimic the estimation of Araujo et al. (2016) and model experience as a dummy variable that takes a value of one for rooky teachers with three or fewer years of experience, and zero for experienced teachers. The results do not change, and experience remains uncorrelated with student learning outcomes. All results are available upon request.

[35] We include tenure status because there are permanent and contract teachers in the control group.

a standard deviation higher. Interestingly, any explanatory power of experience and tenure status on language learning outcomes disappears once we take into account teacher classroom practices (CLASS).

We also report our estimates for math learning outcomes in Table 5. As shown in column (6), we observe a positive effect of test-screened tenured teachers, with end-of-year math test scores that are 8.5 percent of a standard deviation higher when no additional controls are included, albeit only at the 10 percent significance level. Once additional student, family, and classroom characteristics are taken into account in column (7), this effect becomes consistently statistically insignificant but remains positive and close in size to the original estimation.

TABLE 5. ESTIMATED EFFECTS OF TEST-SCREENED TENURED TEACHERS ON LANGUAGE AND MATH

Teacher	Language					Math				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Test-screened tenured	0.105**	0.089**	0.115***	0.080**	0.101***	0.085*	0.068	0.093**	0.059	0.076
	(0.043)	(0.039)	(0.037)	(0.038)	(0.038)	(0.048)	(0.046)	(0.046)	(0.050)	(0.050)
Female teacher			0.245**	0.248**	0.171*			0.097	0.100	0.037
			(0.113)	(0.106)	(0.096)			(0.236)	(0.227)	(0.227)
Years of experience			0.005**	0.002	0.002			0.005***	0.003	0.002
			(0.002)	(0.002)	(0.002)			(0.002)	(0.002)	(0.002)
Years of education			0.001	-0.000	-0.003			-0.004	-0.005	-0.006
			(0.007)	(0.007)	(0.007)			(0.007)	(0.007)	(0.007)
Tenure status				0.071**	0.048				0.070**	0.046
				(0.032)	(0.033)				(0.035)	(0.035)
Cognitive skills					0.035**					0.010
					(0.014)					(0.016)
Neuroticism					-0.001					0.010
					(0.013)					(0.014)
Extraversion					-0.010					0.009
					(0.014)					(0.014)
Openness					0.004					0.011
					(0.015)					(0.018)
Agreeableness					-0.008					-0.002
					(0.017)					(0.016)
Conscientiousness					-0.003					-0.016
					(0.015)					(0.018)
CLASS average					0.042***					0.056***
					(0.016)					(0.019)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES
Parent controls	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES
Classroom controls	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES
Observations	12632	12632	12632	12632	12632	12632	12632	12632	12632	12632
R^2	0.149	0.440	0.441	0.441	0.442	0.123	0.309	0.310	0.310	0.311

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

The fact that we find stronger effects of test-screened tenured teachers for language than for math might be puzzling. However, a review of the Ecuadorian curriculum for kindergarten reveals that instruction focuses on developing socioemotional and language skills rather than mathematical skills (UNESCO International Bureau of Education 2007; Ministerio de Educación del Ecuador 2019). In line with this, teacher training, national entrance tests, and in-service tests

for this grade have emphasized the development of children's socioemotional and language competencies.[36]

The estimated effects are substantial. By comparison, estimations of the effects of certified or test-screened teachers in the US typically have ranged between one and seven percent of a standard deviation for language and math test scores (Clotfelter, Ladd, and Vigdor 2007; Goldhaber 2007; Goldhaber and Anthony 2007; Goldhaber, Gratz, and Theobald 2017; Harris and Sass 2009).

Aside from the effect of test-screened tenured teachers, the only teacher characteristic that correlates with higher end-of-year scores of both language and math is teacher classroom practices (CLASS). This is consistent with the conclusions of Araujo et al. (2016). For language learning outcomes only, we also find a significant, positive association with teachers' cognitive skills. Notably, tenure status does not show a significant correlation with learning outcomes once we account for teacher cognitive skills, personality traits, and classroom practices, which is also in line with the findings of Araujo et al. (2016)

In Appendix B, we show that our results are robust to using the subsample of schools with at least one treated kindergarten classroom, in other words, a class taught by a test-screened tenured teacher. Among the 202 schools with all their kindergarten classrooms included in the sample, 160 are purely control schools with 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. Accordingly, 36 schools in our sample have a combination of treated and control kindergarten classrooms. In this subsample, the student sample size substantially decreases from 12,632 to 2,497 kindergarten children. Nevertheless, we still find that students randomly assigned to test-screened tenured teachers scored at least 10.2 percent of a standard deviation higher on end-of-year language tests, and 8.9 percent of a standard deviation higher end-of-year math tests.

Finally, in Appendix C, we estimate the effect of test-screened tenured teachers on learning outcomes, accounting for the competition rounds described in Table 1. Among the test-screened tenured teachers in our sample, about 25 percent won first-round competitions regulated by the Ministerial Resolution of December 2007 (AM No. 438-07), only about 4 percent won second-round competitions regulated by the Ministerial Resolution of January 2010 (AM No. 018-10), and 71 percent won third-round competitions regulated by the Ministerial Resolution of November 2011 (AM No. 379-11). Our results suggest that first-round competitions recruited more effective teachers. Interestingly, these teachers did not have substantial economic incentives, since the largest salary increase for newly tenured teachers was implemented at the beginning of the third round of competitions.

[36] The guidelines for the subject knowledge test for in-service kindergarten teachers, published by Ecuador's Ministry of Education and in effect between 2008 and 2012, state that "The test for the first level of general basic education contains 60 questions grouped into the following seven blocks or topics: (a) Developmental and Educational Psychology, (b) Children's Literature, (c) Childcare, (d) Plastic Arts (e) Aesthetic Culture, (f) Psychomotricity, and (g) Corporal Expression" (Ministerio de Educación del Ecuador 2011b).

B. Estimates of Test-Screened Tenured, Conventionally Tenured, Test-Screened Contract, and Contract Teachers

In the previous section, we compared the outcomes for kindergarten children randomly assigned to test-screened tenured teachers with the ones for children randomly assigned to their peer teachers within schools. However, our comparison group contains conventionally tenured teachers, test-screened contract teachers, and contract teachers. These teachers differ in some of their characteristics and qualifications, as shown in Table 3. In addition, tenured teachers and contract teachers have different incentives in terms of wages and job security.

In this context, it is important to explore possible differences within the original comparison group. We estimate the effect on learning outcomes of test-screened tenured teachers, conventionally tenured teachers, and test-screened contract teachers, as compared with contract teachers, using a specification similar to equation (1):

$$Y_{ics} = \alpha_s + test\_tenured_{cs}\beta_1 + conventionally\_tenured_{cs}\beta_2 + test\_contract_{cs}\beta_3 + X_{ics}\beta_4 + \bar{X}_{cs}\beta_5 + C_{cs}\beta_6 + T_{cs}\beta_7 + P_{cs}\beta_8 + CLASS_{cs}\beta_9 + u_{ics}, \quad (2)$$

Table 6 presents regression results for this specification. Column (1) shows a positive and significant effect of test-screened tenured teachers on language learning when no controls other than school fixed effects are taken into account. Kindergarten students taught by test-screened tenured teachers score 16.9 percent of a standard deviation higher on end-of-year language tests than students assigned to contract teachers. Children assigned to conventionally tenured teachers also have significantly higher end-of-year language test scores than those assigned to contract teachers. However, the size of this effect is about half of the test-screened tenured teacher effect. The impact of test-screened tenured and conventionally tenured teachers on student language learning persists after controlling for additional child, family, and classroom covariates (column 2). The conventionally tenured teacher effect decreases and becomes statistically insignificant when teacher cognitive skills, personality traits, and classroom practices are considered in column (4). By contrast, the effect of a test-screened tenured teacher remains positive and statistically significant and decreases only slightly to 14.8 percent of a standard deviation. An F-test to evaluate whether the coefficients of conventionally tenured and test-screened tenured teachers are statistically different from one another rejects the null hypothesis at the 10 percent significance level for the estimation in column (1) and at the 1 percent significance level in column (4). Overall, the evidence suggests that the effect of test-screened tenured teachers is larger in size and statistical significance.

For math learning outcomes, we find a strong and positive effect of test-screened tenured teachers. Compared with a contract teacher, the average effect of a test-screened tenured teacher is end-of-year math test scores that are 15.5 percent of a standard deviation higher when no controls are included aside from school fixed effects, as shown in column (5). Conventionally tenured teachers also show a significant positive effect of 10.5 percent of a standard deviation. These effects hold even when we include additional student, family, and classroom characteristics in column (6). When we include a full set of teacher covariates in column (8), the effect of a test-screened tenured teacher is still positive and statistically significant, although its size decreases to 12.9 percent of a standard deviation. By contrast, the

effect of conventionally tenured teachers not only decreases in size to 5.4 percent of a standard deviation, but it also becomes statistically insignificant. Still, the results of an F-test show that overall, the effect of test-screened tenured and conventionally tenured teachers are statistically similar.

TABLE 6. ESTIMATES OF EFFECTS OF TEST-SCREENED TENURED, CONVENTIONALLY TENURED, AND TEST-SCREENED CONTRACT TEACHERS VS. CONTRACT TEACHERS ON LANGUAGE AND MATH

Teacher	Language				Math			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test-screened tenured	0.169***	0.155***	0.157***	0.148***	0.155***	0.136**	0.137**	0.129**
	(0.050)	(0.045)	(0.044)	(0.043)	(0.057)	(0.052)	(0.053)	(0.051)
Conventionally tenured	0.097**	0.098***	0.078**	0.047	0.105**	0.100***	0.080**	0.054
	(0.040)	(0.035)	(0.035)	(0.035)	(0.040)	(0.036)	(0.038)	(0.038)
Test-screened contract	-0.012	0.032	0.030	-0.005	0.008	0.043	0.042	0.030
	(0.061)	(0.056)	(0.055)	(0.061)	(0.061)	(0.055)	(0.054)	(0.059)
Female teacher			0.248**	0.171*			0.101	0.037
			(0.103)	(0.097)			(0.222)	(0.225)
Years of experience			0.002	0.002			0.003	0.002
			(0.002)	(0.002)			(0.002)	(0.002)
Years of education			-0.000	-0.002			-0.005	-0.006
			(0.007)	(0.007)			(0.007)	(0.007)
Cognitive skills				0.036**				0.007
				(0.014)				(0.016)
Neuroticism				-0.001				0.010
				(0.013)				(0.014)
Extraversion				-0.010				0.008
				(0.015)				(0.014)
Openness				0.004				0.011
				(0.015)				(0.017)
Agreeableness				-0.008				-0.002
				(0.017)				(0.016)
Conscientiousness				-0.003				-0.016
				(0.015)				(0.018)
CLASS average				0.042***				0.057***
				(0.016)				(0.019)
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	12632	12632	12632	12632	12632	12632	12632	12632
R^2	0.150	0.441	0.441	0.442	0.124	0.310	0.310	0.312
p value (Test-screened tenured =Conventionally tenured*)*	0.087	0.152	0.040	0.009	0.319	0.480	0.255	0.130

Source: Authors' calculations. *Notes:* Each column reports coefficients from OLS regressions estimated with standard errors (in parentheses) clustered at the school level and school fixed effects. Columns (2)–(7) control for the student characteristics of baseline TVIP score, age, gender, and preschool attendance; the parent characteristics of years of education and living standard conditions; and the classroom characteristics of class size and classroom averages of student and parent characteristics.* Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level

Our estimates confirm the positive and significant effects of teachers who have passed national entry exams and won competitions for tenured positions. In particular, this specification shows that the current test-screened tenured teachers are not only better qualified in their observable characteristics than the pool of non-tenured teachers, but are also more effective. However, our results also show that test-screened contract teachers who have not won a merit-based competition do not have the same effect as test-screened tenured teachers. These results suggest that performance on the screening assessments alongside other teacher

candidates who passed entrance exams was relevant. We discuss these arguments in more detail in section VI.

In addition, while the results may appear to show a positive effect of teacher tenure status on learning outcomes, it is important to recognize that once cognitive skills and classroom practices are included in the regressions in Table 6, the effects of conventionally tenured teachers on language and math learning outcomes become statistically insignificant. In contrast, the effect of test-screened tenured teachers remains significant. This finding is consistent with the results presented in Table 5, where teacher tenure status becomes statistically insignificant after controlling for cognitive skills and classroom practices. Furthermore, it is consistent with the findings of Araujo et al. (2016), who found no significant association between tenure status and student learning outcomes.

V. Heterogeneous Effects

There is ongoing debate about the extent to which highly qualified teachers can close learning gaps between children of different abilities and socioeconomic backgrounds (Borman and Kimball 2005; Boyd et al. 2008; James and Wyckoff 2020; Phillips 2010; Hanushek et al. 2022). Addressing this issue is particularly important in Ecuador and other Latin American countries due to the large disparities in cognitive development between children from high and low socioeconomic backgrounds, which persist throughout their schooling (Schady et al. 2015). Accordingly, in this section we examine potential heterogeneous effects between different groups of students.

We first examine the heterogeneous effects of test-screened tenured teachers on children who entered kindergarten with different levels of language development, as measured by the TVIP. Children who started the school year with lower TVIP scores tended to come from families with lower living standard scores and parental education.[37] We use the baseline TVIP score to sort the student sample into quintiles, from lowest to highest. We then apply the regression model formalized in equation (1) to each TVIP quintile. Table 7 presents the regression results for language and math learning outcomes. Column (1) shows that the impact of test-screened tenured teachers on language is significantly stronger for children in the lowest TVIP quintile when no additional covariates are considered. When we include additional controls in Column (2), we find that test-screened tenured teachers help close the learning gap relative to the highest quintile for all students from the first to the fourth TVIP quintile. In addition, test-screened tenured teachers effectively closed math learning gaps relative to the highest quintile for children in the second through fourth TVIP quintiles, as shown in columns (3) and (4). However, they are not as effective for children in the first TVIP quintile.

[37] There is a 0.36 correlation between TVIP scores and the living standard indicator and a 0.31 correlation between TVIP and parents' years of education. In addition, our data shows that the household living standard indicator and years of parental education have statistically significant power to predict children's TVIP scores. Results are available upon request.

TABLE 7. ESTIMATES OF EFFECTS OF TEST-SCREENED TENURED TEACHERS ON LANGUAGE AND MATH BY GENDER, TVIP QUINTILES, AND EXECUTIVE FUNCTION QUINTILES (HETEROGENEITY)

	Test-screened tenured				Observations
	Language		Math		
	(1)	(2)	(3)	(4)	
TVIP Q1	0.240***	0.164**	0.065	0.044	2544
	(0.078)	(0.077)	(0.056)	(0.061)	
TVIP Q2	0.076	0.124*	0.112	0.142*	2793
	(0.083)	(0.074)	(0.082)	(0.083)	
TVIP Q3	0.100	0.147*	0.179**	0.254**	2318
	(0.067)	(0.079)	(0.086)	(0.104)	
TVIP Q4	0.109	0.165**	0.164**	0.157*	2550
	(0.068)	(0.080)	(0.080)	(0.089)	
TVIP Q5	-0.048	-0.043	-0.052	-0.079	2427
	(0.086)	(0.095)	(0.103)	(0.114)	
Executive Function Q1	0.171**	0.165**	0.045	0.005	2457
	(0.074)	(0.080)	(0.070)	(0.081)	
Executive Function Q2	0.173**	0.157**	0.024	-0.011	2557
	(0.072)	(0.072)	(0.069)	(0.081)	
Executive Function Q3	0.101	0.119	0.145*	0.123	2494
	(0.093)	(0.080)	(0.078)	(0.078)	
Executive Function Q4	-0.006	0.012	0.144	0.173	2558
	(0.069)	(0.074)	(0.095)	(0.110)	
Executive Function Q5	0.092	0.088	0.133	0.129*	2566
	(0.069)	(0.069)	(0.083)	(0.071)	
Male	0.117**	0.151***	0.076	0.102*	6451
	(0.053)	(0.050)	(0.070)	(0.059)	
Female	0.101	0.059	0.096	0.050	6181
	(0.068)	(0.055)	(0.068)	(0.077)	
School fixed effects	YES	YES	YES	YES	
Student controls	NO	YES	NO	YES	
Parent controls	NO	YES	NO	YES	
Classroom controls	NO	YES	NO	YES	
Teacher characteristics	NO	YES	NO	YES	

Source: Authors' calculations. *Notes* Each cell in columns (1), (2), (3) and (4) is the result of a different OLS regression estimated with standard errors (in parentheses) clustered at the school level and school fixed effects. The regressions in columns (2) and (4) control for the student characteristics of age, gender (though not in regressions by gender), TVIP (though not in regressions by TVIP quintiles), and preschool attendance; the parent characteristics of years of education and living standard conditions; the classroom characteristics of class size and classroom averages of student and parent characteristics; and the teacher characteristics of gender, years of experience, years of education, tenure status, cognitive skills, personality traits, and CLASS average score. TVIP stands for Test de Vocabulario en Imágenes Peabody, the Spanish version of the Peabody Picture Vocabulary Test (PPVT). CLASS stands for Classroom Assessment Scoring System. * Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level

We then analyze potential heterogeneous effects among children with different executive function (EF) scores. These scores were assessed by the "Closing Gaps" project and the Ecuadorian Ministry of Education through four tests of inhibitory control, working memory, capacity to pay attention, and cognitive flexibility (Araujo et al. 2016).[38] The EF score attempts to measure a child's ability to regulate thoughts, actions, and emotions, which can facilitate learning (Beisly, Kwon, and Jeon 2020; Ahmed et al. 2019; Best, Miller, and Naglieri 2011).[39]

[38] Given that these tests were administered at the end of the school year and the findings of Araujo et al. (2016), children's EF were likely influenced by teachers; however, they are still a valid measure of variability in children's self-regulatory ability.

[39] In our data, higher EF scores are statistically significantly associated with more years of parental education, but not with the living standard indicator. Regression results are available upon request.

For our heterogeneity analysis, we use the EF score to sort the student sample into quintiles from lowest to highest. We then estimate our regression model for each EF quintile. Table 7 presents our results, showing that the effect of test-screened tenured teachers on language is significantly stronger for children with lower EF scores. This means that these teachers were more effective for children with lower self-control, attention, and, consequently, ability to learn. Although these results are remarkable, they should not be surprising given that test-screen tenured teachers have better pedagogical practices (CLASS scores), which Araujo et al. (2016) found to positively affect children's EF and learning outcomes in Ecuadorian kindergartens. We find no heterogeneous effects for math.

Finally, we examine possible differences by student gender and present the results in Table 7. We find that the positive effects of test-screened tenured teachers on language are mainly driven by male students. This result is likely related to the fact that male students in our sample have significantly lower average EF scores, and test-screened tenured teachers seem to have the greatest effect on students that score low in this area.

Overall, our analysis suggests that the impact of test-screened tenured teachers on learning outcomes is stronger for children who started the school year with lower levels of language development or who have lower EF scores. Our results therefore suggest that test-screened tenured teachers have contributed to closing learning gaps for disadvantaged students in Ecuador.

VI. Potential Mechanisms

So far, we have shown that teachers who passed national entrance tests and were tenured under Ecuador's new recruitment policy were more effective than their peers at improving kindergarten learning outcomes, especially compared to non-tenured teachers. We now explore the possible channels through which these improvements were achieved.

A. Teacher experience, cognitive skills, and classroom practice

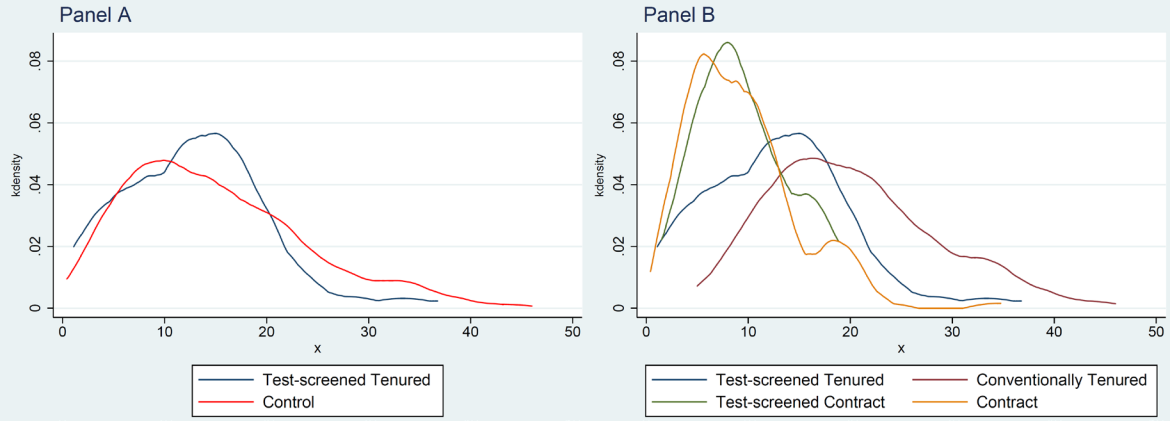
Descriptive statistics on teachers from the "Closing Gaps" data show that test-screened tenured teachers have higher skills and qualifications than non-tenured teachers. This is likely the result of the competitive selection process, which directly or indirectly assessed some of these characteristics. Specifically, the competitions scored and rated measures of teacher experience, cognitive skills, and classroom practices.

We begin our analysis of teacher experience, cognitive skills, and classroom practices as potential mechanisms behind the observed test-screened tenured teacher effect by estimating the probability of being tenured through a new merit-based competition as a function of observable teacher characteristics. In this analysis, we consider only non-tenured teachers as the comparison group because conventionally tenured teachers were not part of the competition's screening processes. We present our results in Table A6, showing that the probability of being a test-screened tenured teacher is significantly correlated with years of experience, cognitive skills, and classroom practices as measured by the CLASS instrument. We observe that an additional year of experience is associated with about a one percentage point increase in the probability of being a test-screened tenured teacher, an increase of one standard deviation in cognitive skills is associated with about a six-percentage-point rise in probability, and an increase of one standard deviation in the CLASS score is associated with a

seven-percentage-point increase in probability, after controlling for other characteristics such as gender, education, and personality traits. Gender is the only other characteristic associated with the probability of being tenured through a merit-based competition, but only at the 10 percent significance level.

We continue our analysis of the potential mechanisms by comparing the distributions of teacher experience, cognitive skills, and classroom practices across teacher types.[40] We first compare test-screened tenured teachers to all of their kindergarten teacher peers. We then compare test-screened tenured teachers, conventionally tenured teachers, test-screened contract teachers, and contract teachers. Figure 1 shows the distributions of teachers' years of experience. In panel A, we see that the comparison group has a higher concentration of less experienced teachers. In panel B, however, we see that the distributions of test-screened contract teachers and contract teachers are right-skewed, with the least experienced teachers concentrated in these groups. In addition, we implement the procedure introduced by Goldman and Kaplan (2018) to statistically compare the distributions. We are able to reject the null hypothesis that the cumulative distribution functions (CDF) of test-screened tenured teachers and test-screened contract teachers are identical at the 1 percent level of significance. However, the null hypothesis of equality cannot be rejected for test-screened contract teachers and contract teachers, which confirms our observations.

FIGURE 1 DISTRIBUTIONS OF TEACHERS' YEARS OF EXPERIENCE



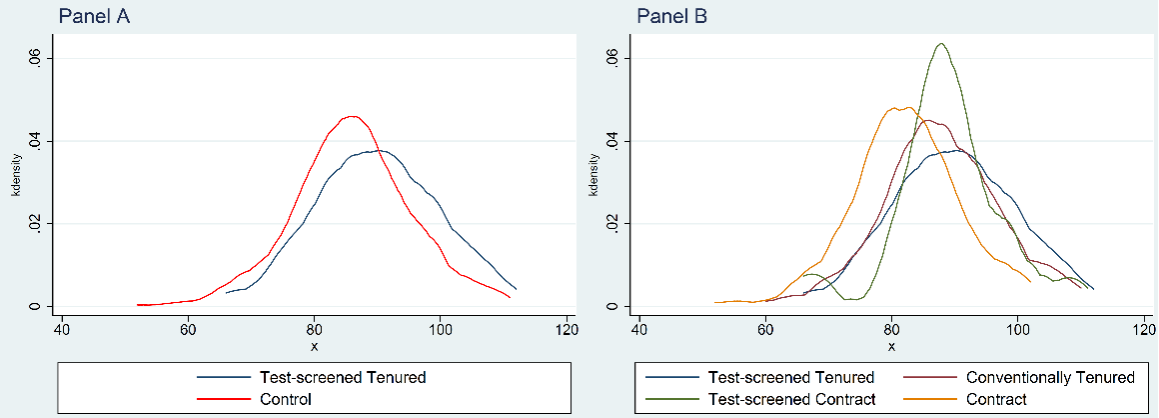
Source: Authors' calculations. *Notes:* This figure reports kernel density distributions of teacher years of experience.

Figure 2 shows the distributions of teachers' cognitive skills. Panel A shows that the distribution of cognitive skills of test-screened tenured teachers falls to the right, indicating a positive selection. By contrast, panel B shows that the distribution of cognitive skills of contract teachers falls to the left, with lower overall scores. The distributions of test-screened tenured teachers, conventionally tenured teachers, and test-screened contract teachers virtually overlap in panel B, but the test-screened tenured teachers' distribution shows a concentration of high-score observations in its right tail. When we statistically test our observations, we fail to reject the null hypothesis that the CDF of test-screened tenured teachers and test-screened contract teachers are identical at all levels of significance. However, we reject the null hypothesis at the

[40] We estimated the distributions with kernel density functions.

1 percent level of significance when comparing test-screened tenured teachers and contract teachers, or test-screened contract teachers and contract teachers. This information adds to the evidence that the first stage of the competitions screened out candidates with lower cognitive skills.

FIGURE 2 DISTRIBUTIONS OF TEACHERS' COGNITIVE SKILLS



Source: Authors' calculations. *Notes*: This figure reports kernel density distributions of cognitive skills, which are 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.

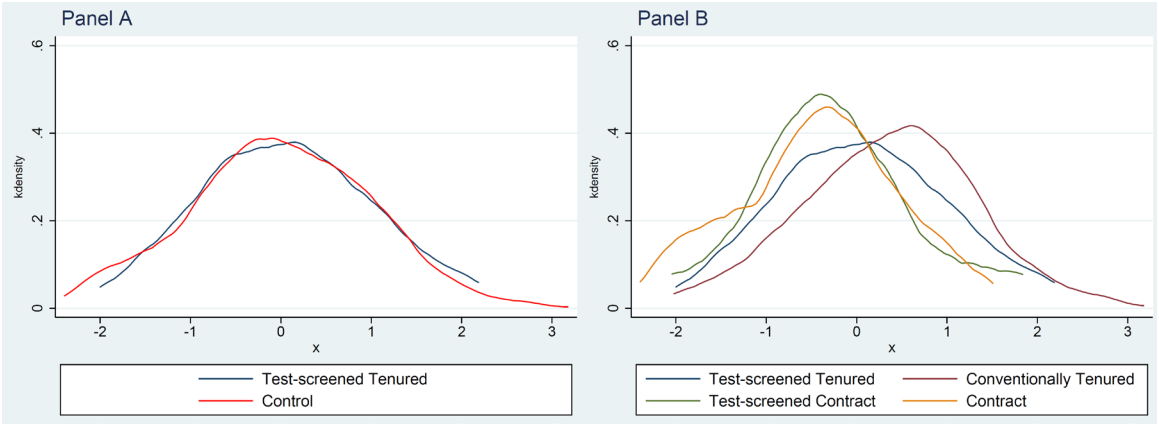
Finally, figure 3 presents the distribution of CLASS scores. At first glance, panel A shows that the distribution of CLASS scores for test-screened tenured teachers overlaps that of the comparison group. However, panel B shows that the distribution of test-screened tenured teachers is centered, and the distribution of conventionally tenured teachers falls to the right. Meanwhile, the distributions of test-screened contract teachers and contract teachers almost overlap and fall to the left, with a high concentration of lower scores for contract teachers. We are able to reject the null hypothesis that the CDF of test-screened tenured teachers and contract teachers are identical at the 10 percent significance level. In contrast, we fail to reject the null hypothesis when comparing test-screened contract teachers and contract teachers at all levels of significance. Although test-screened tenured teachers have significantly higher mean CLASS scores than test-screened contract teachers, the statistical power is insufficient to reject the CDF equality test.

Overall, the distributions and descriptive statistics show that test-screened contract teachers are similar to contract teachers in characteristics such as experience and classroom practices. In contrast, test-screened contract teachers are closer to test-screened tenured teachers in the distribution of their cognitive skills.

Subsequently, we explore whether teacher experience, cognitive skills, or classroom practices (CLASS) mediate the effects of test-screened tenured teachers on learning outcomes by introducing an interaction term between each of these variables and the treatment, as an extension of equation (1). For this analysis, we consider only non-tenured teachers as the comparison group because conventionally tenured teachers were not part of the competitions' screening processes.

FIGURE 3 DISTRIBUTIONS OF TEACHERS' CLASS SCORES

(STANDARDIZED AVERAGE 2011–2012 SCORE)



Source: Authors' calculations. *Notes:* This figure reports kernel density distributions of CLASS scores (2011–2012 average). CLASS stands for Classroom Assessment Scoring System.

Table 8 presents our results and shows that only the classroom practice score (CLASS) significantly interacts with the test-screened tenured teacher treatment, as shown in columns (5) and (6) for language, and (11) and (12) for math. This interaction virtually offsets the association between CLASS and end-of-year language and math test scores. In contrast, the interactions terms of years of experience and cognitive skills are equivalent to zero in statistical significance and size.[41]

We also explore the individual domains of CLASS as mediators of the effect of test-screened tenured teachers on learning outcomes, focusing on language. We present our results in Table A8 in the appendix and show that the domain of socio-emotional support significantly interacts with the test-screened tenured teacher effect. The classroom management domain also significantly interacts with the effect, but it becomes statistically insignificant for language when we use a full set of controls. By contrast, the instructional support domain does not interact with the effect.

In addition, Table A9 of the appendix presents the results of a specification that compares test-screened tenured teachers to test-screened contract teachers. We also find that only the classroom practice score (CLASS) interacts positively and significantly with the test-screened tenured teacher effect. However, these results should be taken with caution because our teacher and student samples decrease, and there are very few schools with both a classroom taught by a test-screened tenured teacher and a classroom taught by a test-screened contract teacher.

Given that the test-screened tenured teacher effect is mediated by the quality of classroom practice, we infer that the competitions' demonstration class instrument should have led to the selection of more effective teachers.

[41] Table A7 in the Appendix presents the same mediator analysis with the original comparison group that includes conventionally tenured teachers. The results are confirmatory and show a significant and strong interaction between the test-screened tenured teacher effect and CLASS for language and math.

TABLE 8. TEACHER EXPERIENCE, COGNITIVE SKILLS, AND CLASSROOM PRACTICE AS POTENTIAL MECHANISMS OF TEST-SCREENED TENURED TEACHER EFFECT ON LANGUAGE AND MATH, COMPARED TO NON-TENURED TEACHERS

	Language						Math					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Test-screened tenured	0.066	0.142	0.222***	0.229***	0.217***	0.230***	-0.046	-0.030	-0.002	-0.044	-0.004	-0.041
	(0.119)	(0.101)	(0.067)	(0.053)	(0.063)	(0.056)	(0.083)	(0.088)	(0.058)	(0.048)	(0.052)	(0.049)
Test-screened tenured x Experience	0.013*	0.007					0.003	-0.001				
	(0.007)	(0.006)					(0.006)	(0.006)				
Experience	-0.003	0.002					0.000	0.003				
	(0.005)	(0.005)					(0.006)	(0.005)				
Test-screened tenured x Cognitive skills			0.074	0.068					0.049	0.048		
			(0.063)	(0.046)					(0.053)	(0.041)		
Cognitive skills			-0.003	0.039					-0.003	0.022		
			(0.028)	(0.026)					(0.028)	(0.023)		
Test-screened tenured x CLASS average					0.116**	0.081*					0.147***	0.159***
					(0.052)	(0.046)					(0.037)	(0.047)
CLASS average					0.017	0.019					0.004	-0.023
					(0.034)	(0.027)					(0.025)	(0.026)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Parent controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Classroom controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Teacher characteristics	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Observations	7123	7123	7123	7123	7123	7123	7123	7123	7123	7123	7123	7123
R^2	0.141	0.439	0.141	0.439	0.142	0.439	0.124	0.306	0.124	0.306	0.126	0.307

Source: Authors' calculations. *Notes:* Each column reports coefficients from OLS regressions estimated with standard errors (in parentheses) clustered at the school level and school fixed effects. This table excludes conventionally tenured teachers and their students. Columns (2), (4), (6), (8), (10), (12), (14) and (16) control for the student characteristics of baseline TVIP score, age, gender, preschool attendance; the parent characteristics of years of education and living standard conditions; the classroom characteristics of class size and classroom averages of student and parent characteristics; and the teacher characteristics of gender years of experience, years of education, tenure status, cognitive skills, personality traits, and CLASS average score. TVIP stands for Test de Vocabulario en Imágenes Peabody, the Spanish version of the Peabody Picture Vocabulary Test (PPVT). CLASS stands for Classroom Assessment Scoring System. * Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level

B. Teacher test scores

Teachers' pedagogical knowledge and subject-specific knowledge could also be a mechanism driving the test-screened tenured teacher effect. To test this hypothesis, we obtained from the administrative records of Ecuador's Ministry of Education the total test score from the first stage of the competitions for all teachers who achieved the minimum score required to be eligible for tenured positions.[42] In other words, in our data we observe the total test scores of test-screened tenured teachers and test-screened contract teachers.[43] This total test score is the sum of the logical-verbal reasoning test, pedagogical knowledge test, and subject-specific knowledge test scores, weighted according to the third-round competition rule (AM No. 379-11). Our analysis relies on the total test score because the administrative records did not retain detailed information on scores on each specific test for test-screened contract teachers, nor did they have any information on demonstration class scores for this group.

We observe each teacher's first total test score, as well as their subsequent total scores.[44] We use the first total test score in our analysis because it is more likely to indicate skills and knowledge rather than training to pass the tests. We normalize this score to have a zero mean and a unit standard deviation for our regression analyses.

We also check the correlation between the first total test score that we obtained from administrative records and the cognitive skills score from the Wechsler Adult Intelligence Scale applied by the "Closing Gaps" project. The correlation is about 0.32, indicating that these tests are positively correlated but do not measure exactly the same skills. On the one hand, this correlation probably shows an overlap between the abilities assessed by the intelligence scale and those assessed by the logical-verbal reasoning test, which is part of the total test score. On the other hand, the knowledge assessed by the pedagogical and subject-specific tests is likely beyond the scope of what can be captured by the intelligence scale. Since the first total test score includes a test of pedagogical knowledge, we also explore its correlation with the CLASS score and find a rather low value of 0.12.

We introduce the first total test score in equation (1) to evaluate whether it is associated with student learning outcomes. Columns (1) through (7) of Table 9 display the results of our analysis for language learning outcomes. They show that teachers' first total test scores are positively and significantly associated with end-of-year language test scores, whether or not we include a full set of controls. A one standard deviation increase in teachers' total test score is associated with an increase of between 7 and 13 percent of a standard deviation in students' end-of-year language test scores. When we compare columns (3) and (4), we observe that adding test scores to the equation reduces the size of the effect of test-screened tenured teachers. In addition, the estimates in columns (3) to (5) reveal that other

[42] The administrative records did not retain information on the test scores of teachers who failed the first-stage exams. Therefore, we cannot observe the test scores of contract teachers who took but did not pass these exams. The teacher survey from the "Closing Gaps" data suggests that 63 percent of these contract teachers took the first-stage exams but did not pass.

[43] We observe the total test score of 53 test-screened tenured teachers and 50 test-screened contract teachers. With respect to our original sample, we lose one test-screened tenured teacher and her students because her total test score was not available in the administrative records.

[44] Teachers who had already passed an eligibility test were allowed to retake it to improve their total test score and increase their chances of earning a tenured position. In our data, 68 percent of teachers have only one total test score, 30 percent have two total test scores, and 2 percent have three total test scores.

teacher characteristics, such as cognitive skills and classroom practices (CLASS), lose their explanatory power and become statistically insignificant after introducing the teacher test score. However, when we introduce an interaction term between the test-screened tenured teacher effect and teachers' total test scores in columns (6) and (7), we find no significant correlation. This suggests that although the eligibility test score is significantly correlated with student learning outcomes, it is not the mediator of the test-screened tenured teacher effect. These results are consistent with our findings on the lack of mediation of teachers' cognitive skills; however, they should be taken with caution because there are very few schools in our sample that have both a classroom taught by a test-screened tenured teacher and a classroom taught by a test-screened contract teacher.

We report our estimates for math in columns (8) through (14) of Table 9. We find no consistent significant association between teachers' first total test score and students' math learning outcomes, which is in line with our previous findings.

Overall, our results provide compelling evidence of a positive and statistically significant association between student language learning outcomes and teacher test scores from Ecuador's competitive recruitment policy. These results suggest that the mandatory logical-verbal reasoning, pedagogical knowledge, and subject content knowledge tests helped identify teachers who more effectively developed their students' language competencies. However, teacher test scores do not seem to mediate test-screened tenured teacher effectiveness. Our findings differ from those of Cruz-Aguayo, Ibarrarán, and Schady (2017), most likely because our data account for potential biases from matching students to teachers. In addition, their study implements a value-added model with a very brief time span of three months between students' baseline and final tests in the middle of the school year, which was likely too short to detect improvements in learning outcomes.

VII. Conclusions

In this paper, we assess the effectiveness of Ecuador's new teacher recruitment policy, which from 2007 onwards required teacher candidates to pass mandatory skills and content knowledge assessments before they were allowed to participate in merit-based competitions for tenured positions in public schools. For our identification strategy, we combine administrative information on teacher recruitment from Ecuador's Ministry of Education with data from the "Closing Gaps" project, which randomly assigned almost 13,000 kindergarten children to classrooms and teachers in the 2012–2013 school year.

Our study shows that the competitive teacher recruitment process successfully screened the candidates with the highest skills and qualifications in the pool of non-tenured teachers. Furthermore, we find that teachers who passed national entrance tests and won competitions for tenure in Ecuador positively affect language learning in kindergarten. The evidence from our study also finds weaker effects of test-screened tenured teachers on math learning in kindergarten. In addition, our analysis suggests that the mechanism driving the effect of test-screened tenured teachers is teachers' classroom practices, which were evaluated in recruitment competitions. We also find that the teachers' total test scores, which are the sum of their scores on tests of logical-verbal reasoning, pedagogical knowledge, and subject content knowledge, positively correlate with students' language learning outcomes.

TABLE 9. TEACHER TEST SCORES AS MECHANISM OF TEST-SCREENED TENURED TEACHER EFFECT ON LANGUAGE AND MATH

	Language							Math						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Test-screened Tenured			0.175***	0.193**	0.179***	0.322**	0.369***			0.162	0.160	0.160	0.080*	0.074
			(0.064)	(0.079)	(0.063)	(0.147)	(0.114)			(0.107)	(0.105)	(0.104)	(0.041)	(0.086)
Years of experience			0.000	-0.006	-0.000					-0.014	-0.013	-0.014		
			(0.005)	(0.007)	(0.005)					(0.009)	(0.010)	(0.010)		
Cognitive skills				0.080*	0.024						-0.013	-0.012		
				(0.042)	(0.044)						(0.066)	(0.073)		
First total test score	0.070***	0.121***	0.130***		0.118***	0.096***	0.169***	0.043	0.098**	-0.008		-0.003	0.057	0.122*
	(0.025)	(0.044)	(0.023)		(0.026)	(0.028)	(0.064)	(0.040)	(0.039)	(0.035)		(0.036)	(0.053)	(0.071)
Test-screened tenured x First eligibility test score						-0.049	-0.042						-0.038	-0.041
						(0.067)	(0.086)						(0.093)	(0.105)
CLASS average			0.038	0.083**	0.035					0.194***	0.194***	0.195***		
			(0.026)	(0.032)	(0.027)					(0.049)	(0.053)	(0.052)		
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	YES	YES	YES	NO	YES	NO	YES	YES	YES	YES	NO	YES
Parent controls	NO	YES	YES	YES	YES	NO	YES	NO	YES	YES	YES	YES	NO	YES
Classroom controls	NO	YES	YES	YES	YES	NO	YES	NO	YES	YES	YES	YES	NO	YES
Teacher controls	NO	NO	YES	YES	YES	NO	NO	NO	NO	YES	YES	YES	NO	NO
Observations	3092	3092	3092	3123	3092	3092	3092	3092	3092	3092	3123	3092	3092	3092
R^2	0.148	0.431	0.437	0.437	0.437	0.149	0.432	0.138	0.307	0.310	0.311	0.310	0.138	0.307

Source: Authors' calculations. *Notes:* Each column reports coefficients from OLS regressions estimated with standard errors (in parentheses) clustered at the school level and school fixed effects. Teacher tests scores are the sum of their scores on the logical-verbal reasoning, pedagogical knowledge, and subject-specific knowledge tests. When indicated, regressions control for the student characteristics of baseline TVIP score, age, gender, and preschool attendance; the parent characteristics of years of education and living standard conditions; and the classroom characteristics of class size and classroom averages of student and parent characteristics; the teacher characteristics of gender, years of experience, years of education, and personality traits. TVIP stands for Test de Vocabulario en Imágenes Peabody, which is the Spanish version of the Peabody Picture Vocabulary Test (PPVT). CLASS stands for Classroom Assessment Scoring System.* Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level.

We do not find effects for teachers who have passed entrance tests but have not yet won competitions for tenured positions and are therefore working under fixed-term contracts. Although these teachers have similar cognitive skills to test-screened tenured teachers, they have statistically significantly lower scores for classroom practice, which appears to be the mechanism driving the effect on learning outcomes. Thus, our study suggests that the demonstration class, which is the last stage in the competitive recruitment process, helped identify the most effective teachers among the candidates who had already passed standardized tests. Moreover, the fact that contract teachers perform poorly and have the lowest experience, cognitive skills, and classroom practice scores suggests that weaker teacher candidates were screened out or discouraged by the competitive recruitment process.

We conclude by drawing policy implications from our analysis. Our results show that between 2007 and 2012, the Ecuadorian teacher reform succeeded in recruiting better-qualified and more effective teachers. However, by 2012, the annual cost of each of these newly tenured teachers was more than double that of a contract teacher, which may raise questions about whether the policy was worthwhile. In Appendix Table A10, following Levin et al (2017), we present a simple cost-effectiveness analysis that compares the annual cost required to gain one additional standard deviation of language achievement through a test-screened tenured teacher, a conventionally tenured teacher, and a contract teacher. Our analysis shows that in the case of Ecuador, the cost of one additional standard deviation of student achievement is vastly lower for a test-screened tenured teacher than it is for a contract teacher. Moreover, newly tenured teachers are also more cost-effective than conventional ones. In this respect, our results also show that simply adopting the experimental evidence on the effectiveness of contract teachers from a different context (Duflo, Dupas, and Kremer 2015; Muralidharan and Sundararaman 2013) would have been a costly policy miscalculation in Ecuador.

The experimental results from Ecuador's competitive teacher recruitment process, together with current non-experimental evidence from Mexico (Estrada 2019; Bedoya, De Hoyos, and Estrada 2023), Colombia (Brutti and Sánchez Torres 2022; Busso et al. 2024), and Peru (Bertoni et al. 2023), can certainly help guide policy reforms in other contextually similar Latin American countries. In Ecuador, implementing this policy has shown that a very small fraction of education spending can have an immediate impact on learning outcomes, which can then lead to long-term economic benefits for students (Chetty, Friedman, and Rockoff 2014; Chetty et al. 2011). While Ecuador spent approximately 87 percent of its education budget on personnel salaries between 2008 and 2013, less than 0.1 percent was used to actually implement the new teacher recruitment policy (see figure A1). Evidence on effective teacher screening and hiring is particularly relevant for Latin American governments because they spend most of their education budgets on permanent and non-permanent personnel, and, consequently, rely almost exclusively on teacher quality policies to improve student learning outcomes.

REFERENCES

Ahmed, Sammy F., Sandra Tang, Nicholas E. Waters, and Pamela Davis-Kean. 2019. 'Executive Function and Academic Achievement: Longitudinal Relations from Early Childhood to Adolescence.' *Journal of Educational Psychology* 111 (3): 446–58. https://doi.org/10.1037/edu0000296.

Alkire, Sabina, Adriana Conconi, Gisela Robles, José Manuel Roche, María Emma Santos, Suman Seth, and Ana Vaz. 2016. 'The Global Multidimensional Poverty Index (MPI): 5-Year Methodological Note'. 37. *OPHI Briefing*. OPHI Briefing. Oxford. https://www.ophi.org.uk/wp-content/uploads/MPI_Methodology_2010-2015_Jan2016.pdf.

Almlund, Mathilde, Angela Lee Duckworth, James J. Heckman, and Tim Kautz. 2011. 'Personality Psychology and Economics'. In *Handbook of The Economics of Education*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, Volume 4, 1–181. Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-444-53444-6.00001-8.

Angrist, Joshua D., and Jonathan Guryan. 2008. 'Does Teacher Testing Raise Teacher Quality? Evidence from State Certification Requirements'. *Economics of Education Review* 27:483–503. https://doi.org/10.1016/j.econedurev.2007.03.002.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. 'Teacher Quality and Learning Outcomes in Kindergarten'. *The Quarterly Journal of Economics* 131 (3): 1415–53. https://doi.org/10.1093/qje/qjw016.

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

Araujo P., M. Daniela, and Daniela Bramwell. 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*, 34. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000232430.

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

Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. 2019. 'An Experimental Evaluation of Three Teacher Quality Measures: Value-Added, Classroom Observations, and Student Surveys'. *Economics of Education Review* 73 (December):101919. https://doi.org/10.1016/j.econedurev.2019.101919.

Bakos, Daniela Schneider, Natalie Denburg, Rochele Paz Fonseca, and Maria Alice de Mattos Pimenta Parente. 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): 101–7. https://doi.org/10.3922/j.psns.2010.1.013.

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

Bedoya, Juan, Rafael De Hoyos, and Ricardo Estrada. 2023. 'Rule-Based Civil Service Evidence from a Nationwide Teacher Reform in Mexico'. 10540. Policy Research Working Paper. Washington, D. C. http://www.worldbank.org/prwp.

Beisly, Amber, Kyong-Ah Kwon, and Shinyoung Jeon. 2020. 'Executive Function and Learning Behaviors: Associations with Academic Skills among Preschoolers'. *Early Child Development and Care* 190 (15): 2469–83. https://doi.org/10.1080/03004430.2019.1585347.

Bertoni, Eleonora, Gregory Elacqua, Carolina Mendez, and Humberto Santos. 2023. 'Teacher Selection Instruments and Teacher Value-Added: Evidence From Peru'. *Educational Evaluation and Policy Analysis*, February, 01623737221149417. https://doi.org/10.3102/01623737221149417.

Best, John R., Patricia H. Miller, and Jack A. Naglieri. 2011. 'Relations between Executive Function and Academic Achievement from Ages 5 to 17 in a Large, Representative National Sample'. *Learning and Individual Differences* 21 (4): 327–36. https://doi.org/10.1016/j.lindif.2011.01.007.

Bietenbeck, Jan, Marc Piopiunik, and Simon Wiederhold. 2018. 'Africa's Skill Tragedy: Does Teachers' Lack of Knowledge Lead to Low Student Performance?' *Journal of Human Resources* 53 (3): 553–78. https://doi.org/10.3368/jhr.53.3.0616-8002R1.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a C, and Justin Sandefur. 2018. 'Experimental Evidence on Scaling up Education Reforms in Kenya'. *Journal of Public Economics* 168:1–20. https://doi.org/10.1016/j.jpubeco.2018.08.007.

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

Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah Rockoff, and James Wyckoff. 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): 793–818. https://doi.org/10.1002/pam.20377.

Bruno, Paul, and Katharine O. Strunk. 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): 426–60. https://doi.org/10.3102/0162373719865561.

Bruns, Barbara, and Javier Luque. 2015. 'Great Teachers: How to Raise Student Learning in Latin American and the Caribbean'. Washington D.C.: The World Bank. https://doi.org/10.1596/978-1-4648-0151-8.

Brutti, Zelda, and Fabio Sánchez Torres. 2022. 'Turning around Teacher Quality in Latin America: Renewed Confidence and Lessons from Colombia'. *Economic Analysis and Policy* 73 (March):62–93. https://doi.org/10.1016/J.EAP.2021.10.008.

Busso, Matias, Sebastián Montaño, Juan Muñoz-Morales, and Nolan G. Pope. 2024. 'The Unintended Consequences of Merit-Based Teacher Selection: Evidence from a Large-Scale Reform in Colombia'. *Journal of Public Economics* 239 (November):105238. https://doi.org/10.1016/J.JPUBECO.2024.105238.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. 'How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star'. *The Quarterly Journal of Economics* 126 (4): 1593–1660. https://doi.org/10.1093/qje/qjr041.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. 'Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood'. *American Economic Review* 104 (9): 2633–79. https://doi.org/10.1257/aer.104.9.2633.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. 'Teacher-Student Matching and the Assessment of Teacher Effectiveness'. *Journal of Human Resources* 41 (4): 778–820. https://doi.org/10.2307/40057291.

———. 2007. 'Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects'. *Economics of Education Review* 26 (6): 673–82. https://doi.org/10.1016/j.econedurev.2007.10.002.

Costa Jr., Paul T, and Robert R McCrae. 2008. 'The Revised NEO Personality Inventory (NEO-PI-R).' In *The SAGE Handbook of Personality Theory and Assessment, Vol 2: Personality Measurement and Testing.*, 179–98. Thousand Oaks, CA, US: Sage Publications, Inc. https://doi.org/10.4135/9781849200479.n9.

Cruz-Aguayo, Yyannú, Diana Hincapié, and Catherine Rodríguez. 2020. *Profesores a Prueba: Claves Para Una Evaluación Docente Exitosa*. *Profesores a Prueba: Claves Para Una Evaluación Docente Exitosa*. Inter-American Development Bank. https://doi.org/10.18235/0002149.

Cruz-Aguayo, Yyannú, Pablo Ibarrarán, and Norbert Schady. 2017. 'Do Tests Applied to Teachers Predict Their Effectiveness?' *Economics Letters* 159 (October):108–11. https://doi.org/10.1016/J.ECONLET.2017.06.035.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. 'School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools'. *Journal of Public Economics* 123:92–110. https://doi.org/10.1016/j.jpubeco.2014.11.008.

Dunn, Lloyd, Eligio Padilla, Delia Lugo, and Leota Dunn. 1986. *Test de Vocabulario En Imagenes Peabody*. Circle Pines, MN: American Guidance Service.

Elacqua, Gregory, Diana Hincapié, Emiliana Vegas, and Mariana. Alfonso. 2017. 'Profesión: Profesor En América Latina. ¿Por Qué Se Perdió El Prestigio Docente y Cómo Recuperarlo?' Washington, D.C.: Inter-American Development Bank. https://doi.org/10.18235/0000901.

Estrada, Ricardo. 2019. 'Rules versus Discretion in Public Service: Teacher Hiring in Mexico'. *Journal of Labor Economics* 37 (2): 545–79. https://doi.org/10.1086/700192.

Fasfous, Ahmed F, Natalia Hidalgo-Ruzzante, Raquel Vilar-López, Andrés Catena-Martínez, and Miguel Pérez-García. 2013. 'Cultural Differences in Neuropsychological Abilities Required to Perform Intelligence Tasks'. *Archives of Clinical Neuropsychology* 28 (8): 784–90. https://doi.org/10.1093/arclin/act074.

Glewwe, Paul, Rongjia Shen, Bixuan Sun, and Suzanne Wisniewski. 2020. 'Teachers in Developing Countries'. In *The Economics of Education*, edited by Steve Bradley and Colin B T Green, Second, 371–89. Elsevier. https://doi.org/10.1016/b978-0-12-815391-8.00027-6.

Goldhaber, Dan. 2007. 'Everyone's Doing It, but What Does Teacher Testing Tell Us about Teacher Effectiveness?' *The Journal of Human Resources* 42 (4): 765–94. https://doi.org/10.2307/40057329.

Goldhaber, Dan, and Emily Anthony. 2007. 'Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching'. *The Review of Economics and Statistics* 89 (1): 134–50. https://doi.org/10.1162/rest.89.1.134.

Goldhaber, Dan, Trevor Gratz, and Roddy Theobald. 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 (December):112–29. https://doi.org/10.1016/j.econedurev.2017.09.002.

Goldhaber, Dan, Cyrus Grout, and Nick Huntington-Klein. 2017. 'Screen Twice, Cut Once: Assessing the Predictive Validity of Applicant Selection Tools'. *Education Finance and Policy* 12 (2): 197–223. https://doi.org/10.1162/EDFP_a_00200.

Goldman, Matt, and David M Kaplan. 2018. 'Comparing Distributions by Multiple Testing across Quantiles or CDF Values'. *Journal of Econometrics* 206 (1): 143–66. https://doi.org/https://doi.org/10.1016/j.jeconom.2018.04.003.

Hamre, Bridget K., Karen M. La Paro, and Robert C. Pianta. 2007. *Classroom Assessment Scoring System (CLASS) Manual*. Baltimore: Paul H. Brookes Publishing Co., Inc.

Hanushek, Eric A. 2021. 'Addressing Cross-National Generalizability in Educational Impact Evaluation'. *International Journal of Educational Development* 80 (January):102318. https://doi.org/10.1016/J.IJEDUDEV.2020.102318.

Hanushek, Eric A, Jacob D Light, Paul E Peterson, Laura M Talpey, and Ludger Woessmann. 2022. 'Long-Run Trends in the U.S. SES-Achievement Gap'. *Education Finance and Policy* 17 (4): 608–40. https://doi.org/10.1162/edfp_a_00383.

Hanushek, Eric A., Marc Piopiunik, and Simon Wiederhold. 2019. 'The Value of Smarter Teachers'. *Journal of Human Resources* 54 (4): 857–99. https://doi.org/10.3368/jhr.54.4.0317.8619R1.

Harris, Douglas N., and Tim R. Sass. 2009. 'The Effects of NBPTS- Certified Teachers on Student'. *Journal of Policy Analysis and Management* 28 (1): 55–80. https://doi.org/10.2307/29738986.

Heineck, Guido, and Silke Anger. 2010. 'The Returns to Cognitive Abilities and Personality Traits in Germany'. *Labour Economics* 17 (3): 535–46. https://doi.org/10.1016/J.LABECO.2009.06.001.

Instituto Ecuatoriano de Estadísticas y Censos (INEC). 2020. 'Serie Histórica IPC'. Ecuador En Cifras. 2020. https://www.ecuadorencifras.gob.ec/historicos-ipc/.

Jacob, Brian A., Jonah E. Rockoff, Eric S. Taylor, Benjamin Lindy, and Rachel Rosen. 2018. 'Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools'. *Journal of Public Economics* 166 (October):81–97. https://doi.org/10.1016/j.jpubeco.2018.08.011.

James, Jessalynn, and James Wyckoff. 2020. 'Teacher Labor Markets: An Overview'. In *The Economics of Education*, edited by Steve Bradley and Colin B T Green, 355–70. Elsevier. https://doi.org/10.1016/b978-0-12-815391-8.00026-4.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. 'What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City'. *Economics of Education Review* 27:615–31. https://doi.org/10.1016/j.econedurev.2007.05.005.

Kane, Thomas J., and Douglas O. Staiger. 2012. 'Gathering Feedback for Teaching Combining High-Quality Observations with Student Surveys and Achievement Gains'. *MET Project Reseach Paper*. Bill & Melinda Gates Foundation.

Kuncel, Nathan R., Deniz S. Ones, and Paul R. Sackett. 2010. 'Individual Differences as Predictors of Work, Educational, and Broad Life Outcomes'. *Personality and Individual Differences*. https://doi.org/10.1016/j.paid.2010.03.042.

Levin, Henry M., Patrick J. McEwan, Clive Belfield, A. Brooks Bowden, and Robert Shand. 2017. *Economic Evaluation in Education: Cost-Effectiveness and Benefit-Cost Analysis*. Third. Los Angeles, London, New Delhi, Singapore, Washington DC, Melbourne.: SAGE Publications, Inc.

Metzler, Johannes, and Ludger Woessmann. 2012. 'The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from within-Teacher within-Student Variation'. *Journal of Development Economics* 99 (2): 486–96. https://doi.org/10.1016/J.JDEVECO.2012.06.002.

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).

———. 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).

———. 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).

———. 2011a. *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).

———. 2011b. 'Instructivo de La Prueba Para Docentes de Primer Año de Educación Básica'. Quito, Ecuador: Ministerio de Educación del Ecuador. http://web.educacion.gob.ec/_upload/pruebadeprimero.pdf.

———. 2012. 'Ecuador 2007-2012 a Transformation in Education. Presentation at the Inter-American Development Bank'. Washington D.C.: Ministerio de Educación del Ecuador (MinEduc).

———. 2019. 'Currículo de Los Niveles de Educación Obligatoria. Subnivel PREPARATORIO'. Quito: Ministerio de Educación del Ecuador.

Muñoz-Sandoval, Ana F., Richard W. Woodcock, Kevin S. McGrew, and N. Mather. 2005. *Batería III Woodcock-Muñoz*. Itasca, IL: Riverside Publishing.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2013. 'Contract Teachers: Experimental Evidence from India'. 19440. *NBER Working Papers*. NBER Working Papers. https://doi.org/10.3386/w19440.

OECD. 2023. *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. PISA. Paris : OECD Publishing. https://doi.org/10.1787/53f23881-en.

Phillips, Kristie 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): 464–93. https://doi.org/10.1086/651192.

Pritchett, Lant, and Justin Sandefur. 2015. 'Learning from Experiments When Context Matters'. *American Economic Review* 105 (5): 471–75. https://doi.org/10.1257/AER.P20151016.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2011. 'Can You Recognize an Effective Teacher When You Recruit One?' *Education Finance and Policy* 6 (1): 43–74. https://doi.org/10.1162/EDFP_a_00022.

Rothstein, Jesse. 2010. 'Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement'. *The Quarterly Journal of Economics* 125 (1): 175–214. https://doi.org/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. https://www.globalpartnership.org/content/early-grade-mathematics-assessment-egma-conceptual-framework-based-mathematics-skills.

———. 2009b. 'Early Grade Reading Assessment Toolkit'. Research Triangle Park, NC: RTI International. https://globalreadingnetwork.net/eddata/early-grade-reading-assessment-toolkit-2009.

Schady, Norbert, Jere R. Behrman, M. Caridad Araujo, Rodrigo Azuero, Raquel Bernal, David Bravo, Florencia Lopez-Boo, et al. 2015. 'Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries'. *Journal of Human Resources* 50 (2): 446–63. https://doi.org/10.3368/jhr.50.2.446.

Schneider, Ben Ross, Pablo Cevallos Estarellas, and Barbara Bruns. 2019. 'The Politics of Transforming Education in Ecuador: Confrontation and Continuity, 2006–17'. *Comparative Education Review* 63 (2): 259–80. https://doi.org/10.1086/702609.

Steinberg, Matthew P., and Morgaen L. Donaldson. 2016. 'The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era'. *Education Finance and Policy* 11 (3): 340–59. https://doi.org/10.1162/EDFP_a_00186.

Todd, Petra E., and Kenneth I. Wolpin. 2003. 'On the Specification and Estimation of the Production Function for Cognitive Achievement'. *The Economic Journal* 113 (485): F3–33. https://doi.org/10.1111/1468-0297.00097.

UNESCO International Bureau of Education. 2007. 'Ecuador Early Childhood Care and Education (ECCE) Programmes. Country Profile Commissioned for the EFA Global Monitoring Report 2007'. Geneva.

VandenBos, Gary R. (Ed.). 2007. *APA Dictionary of Psychology*. Washington, DC, US: American Psychological Association.

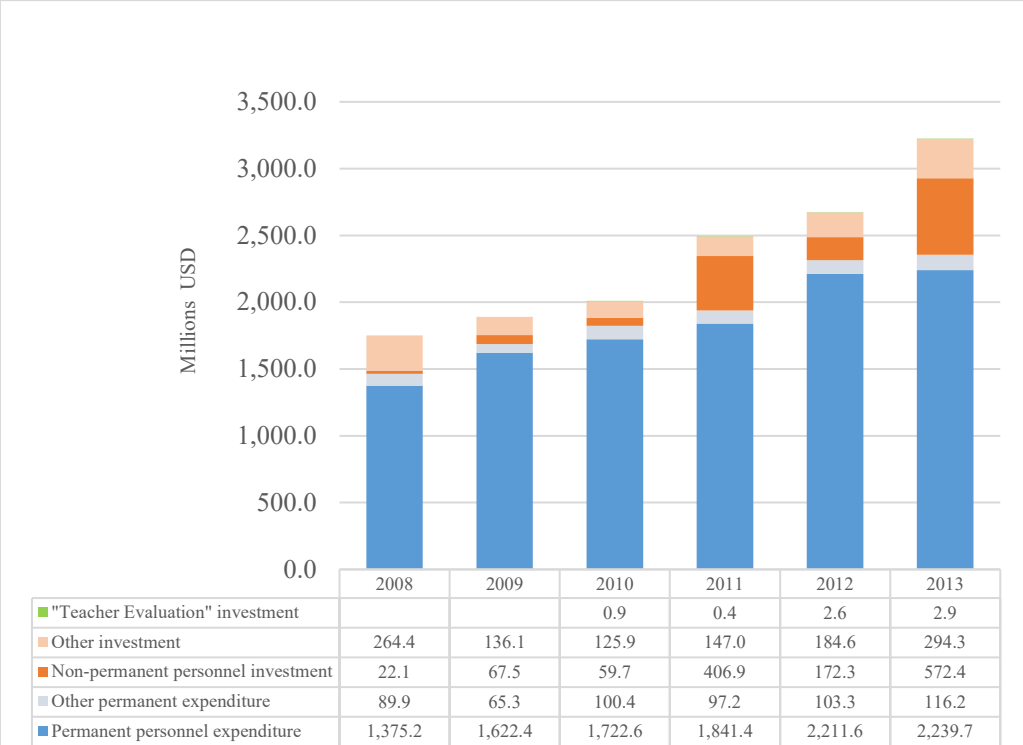
Wechsler, David, and Psychological Corporation. 1997. *WAIS-III: Administration and Scoring Manual: Wechsler Adult Intelligence Scale*. Third. San Antonio, TX: Psychological Corporation.

Williams, Martin J. 2020. 'External Validity and Policy Adaptation: From Impact Evaluation to Policy Design'. *The World Bank Research Observer* 35 (2): 158–91. https://doi.org/10.1093/WBRO/LKY010.

World Bank. 2019. 'Education Statistics (EdStats)'. The World Bank Group. 2019. http://datatopics.worldbank.org/education/country/ecuador.

APPENDIX A: Figures and Tables

FIGURE A1 EXECUTED BUDGET OF ECUADOR'S MINISTRY OF EDUCATION (MILLIONS USD)



	2008	2009	2010	2011	2012	2013
"Teacher Evaluation" investment			0.9	0.4	2.6	2.9
Other investment	264.4	136.1	125.9	147.0	184.6	294.3
Non-permanent personnel investment	22.1	67.5	59.7	406.9	172.3	572.4
Other permanent expenditure	89.9	65.3	100.4	97.2	103.3	116.2
Permanent personnel expenditure	1,375.2	1,622.4	1,722.6	1,841.4	2,211.6	2,239.7

Source: Prepared by the authors based on Ecuador's General Budget (*Presupuesto General del Estado*-PGE) published by the Ministry of Finance at https://www.finanzas.gob.ec/ejecucion-presupuestaria/, and Ecuador's Annual Investment Plan (*Plan Annual de Inversiones*-PAI) published by the Planning Secretariat at https://sni.gob.ec/inicio. *Notes:* The executed education budget, in millions of US current dollars, includes national spending on early education, general basic education, and high school. It does not include national expenditures for universities or other higher education. "Teacher evaluation" investment is the budget of the investment program for designing and applying mandatory entrance exams for teachers. In 2008 and 2009, "Teacher Evaluation" investment was not disaggregated from other investment (data unavailable).

TABLE A1 SUMMARY STATISTICS FOR TEACHERS, FULL SAMPLE AND ANALYSIS SAMPLE

Teacher characteristics	Teacher Full Sample Mean (SD)	Teacher Analysis Sample Mean (SD)	Teacher Dropout Sample Mean (SD)	Difference (p-value)
	(1)	(2)	(3)	(4)
Proportion female	0.99	0.99	0.95	-0.04
	(0.10)	(0.10)	(0.22)	(0.43)
Proportion tenured	0.57	0.58	0.35	-0.23*
	(0.50)	(0.49)	(0.49)	(0.05)
Age	42.23	42.13	44.45	2.32
	(9.58)	(9.42)	(12.62)	(0.43)
Years of experience	14.91	14.74	18.63	3.88
	(8.88)	(8.60)	(13.37)	(0.21)
Years of education	17.14	17.15	16.90	-0.25
	(1.93)	(1.92)	(2.20)	(0.62)
University degree	0.99	0.99	0.95	-0.04
	(0.11)	(0.11)	(0.22)	(0.45)
Cognitive skills	86.46	86.46	.	
	(9.53)	(9.53)	(.)	
Neuroticism	43.85	43.85	.	
	(6.72)	(6.72)	(.)	
Extraversion	45.65	45.65	.	
	(6.83)	(6.83)	(.)	
Openness	50.82	50.82	.	
	(6.75)	(6.75)	(.)	
Agreeableness	48.22	48.22	.	
	(7.56)	(7.56)	(.)	
Conscientiousness	57.55	57.55	.	
	(8.15)	(8.15)	(.)	
CLASS average	3.48	3.48	3.43	-0.05
	(0.29)	(0.28)	(0.34)	(0.49)
CLASS 2011–2012	3.63	3.62	3.69	0.06
	(0.37)	(0.37)	(0.42)	(0.57)
CLASS 2012–2013	3.41	3.41	3.32	-0.10
	(0.28)	(0.28)	(0.28)	(0.15)
N	450	430	20	

Source: Authors' calculations. *Notes:* This table reports means and standard deviations (in parenthesis) of teacher characteristics for the full and analysis samples. Column (4) displays differences between the full sample and the analysis sample, as well as 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 (1–2), medium (3–5), or high (6–7). CLASS statistics for 2011–2012 are based on 326 observations. * Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level.

TABLE A2 MEAN DIFFERENCES IN TEACHER CHARACTERISTICS, WITHIN SCHOOLS

	Teacher Characteristics												
	Female	Years of experience	Years of education	University degree	Cognitive skills	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness	CLASS average	CLASS 2011–2012	CLASS 2012–2013
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Panel A													
Test-screened tenured	-0.025	-4.485*	0.108	-0.000***	0.380	-0.952	3.570	1.989	2.125	0.799	-0.054	0.095	-0.061
	(0.034)	(2.562)	(0.581)	(0.000)	(2.781)	(1.587)	(2.186)	(2.064)	(1.878)	(2.441)	(0.080)	(0.212)	(0.067)
Observations	430	430	430	430	430	430	430	430	430	430	430	326	430
R^2	0.456	0.643	0.523	0.494	0.517	0.548	0.489	0.479	0.536	0.494	0.619	0.751	0.559
Panel B													
Conventionally tenured	0.000**	8.497**	0.177	0.000**	1.108	0.048	-2.863	-0.146	-1.577	0.260	0.159	-0.056	0.125
	(0.000)	(3.545)	(0.821)	(0.000)	(3.950)	(2.298)	(3.404)	(2.911)	(2.723)	(3.858)	(0.111)	(0.292)	(0.090)
Observations	250	250	250	250	250	250	250	250	250	250	250	207	250
R^2	0.775	0.805	0.669	0.831	0.702	0.697	0.637	0.604	0.706	0.660	0.734	0.769	0.708
Panel C													
Test-screened contract	0.000***	-8.458	1.000	-0.000	13.000	7.385	1.650	2.650	3.010	0.235	-0.016	-0.213***	-0.025
	(0.000)	(14.561)	(1.540)	(0.000)	(29.251)	(9.645)	(3.125)	(2.509)	(2.386)	(18.436)	(0.178)	(0.000)	(0.038)
Observations	104	104	104	104	104	104	104	104	104	104	104	43	104
R^2	0.660	0.782	0.903	1.000	0.707	0.791	0.758	0.870	0.792	0.727	0.769	0.839	0.824
Panel D													
Contract	0.074	-3.466	-0.963	-0.000***	-5.185	1.972	-3.825	-6.258	-3.876	-3.872	-0.093	-0.325***	-0.007
	(0.135)	(3.440)	(1.464)	(0.000)	(4.932)	(3.634)	(4.685)	(4.548)	(4.660)	(4.698)	(0.169)	(0.000)	(0.176)
Observations	184	184	184	184	184	184	184	184	184	184	184	118	184
R^2	0.617	0.796	0.715	1.000	0.763	0.728	0.743	0.702	0.745	0.726	0.757	0.883	0.769

Source: Authors' calculations. *Notes:* This table reports mean differences in teacher characteristics within schools, comparing test-screened tenured teachers to: the full sample of peer teachers (Panel A), the subsample of conventionally tenured teachers (Panel B), the subsample of test-screened contract teachers (Panel C), and subsample of contract teachers (Panel D). We estimate average differences within schools with individual OLS regressions with school fixed effects and standard errors (in parentheses) clustered at the school level. Teacher 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 (1–2), medium (3–5), or high (6–7). CLASS statistics for 2011–2012 are based on 326 observations. * Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level.

TABLE A3 RANDOMIZATION TEST FOR TEST-SCREENED TENURED VS. CONVENTIONALLY TENURED, TEST-SCREENED CONTRACT, AND CONTRACT TEACHERS

	Test-screened tenured vs. full sample	Test-screened tenured vs.		
		Conventionally tenured	Test-screened contract	Contract
	(1)	(2)	(3)	(4)
Children:				
Age (months)	-0.000	-0.001	0.000	-0.000
	(0.000)	(0.001)	(0.000)	(0.000)
Proportion female	0.001	0.002	0.002	-0.000
	(0.003)	(0.005)	(0.004)	(0.004)
TVIP	0.000	-0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Proportion who attended early	-0.005	-0.012*	0.001	-0.002
childhood education	(0.005)	(0.007)	(0.002)	(0.008)
Family:				
Parents' years of schooling	0.001**	0.002*	0.000	0.001
	(0.001)	(0.001)	(0.000)	(0.001)
Living standard indicator	-0.000	-0.001	0.003	-0.000
	(0.003)	(0.004)	(0.002)	(0.005)
School fixed effects	YES	YES	YES	YES
Observations	12632	7104	3123	5595
R^2	0.580	0.669	0.956	0.840
F	1.09	0.97	0.35	0.64
p	0.372	0.449	0.910	0.695

Source: Authors' calculations. *Note:* OLS models estimated with standard errors (in parentheses) clustered 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). The total number of observations is restricted to the student sample with full information on student, parent, classroom, and teacher characteristics. The estimation in column (2) restricts the sample to students randomly assigned *to* test-screened tenured or conventionally tenured teachers. The estimation in column (3) restricts the sample to students randomly assigned to test-screened tenured or test-screened contract teachers. The estimation in column (4) restricts the sample to students randomly assigned to test-screened tenured and contract teachers. * Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level

TABLE A4 NO-SHOW, ATTRITION, AND LATE ENROLLMENT TESTS

Teacher Characteristics	No-shows		Dropout		Late enrollments	
	(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)
School fixed effects	YES	YES	YES	YES	YES	YES
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

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

TABLE A5 NO-SHOW, DROPOUT, AND LATE ENROLLMENT TESTS FOR TEST-SCREENED TENURED TEACHER VS. CONVENTIONALLY TENURED, TEST-SCREENED CONTRACT, AND CONTRACT TEACHERS

Teacher Characteristics	No-shows		Dropouts		Late enrollments	
	(1)	(2)	(3)	(4)	(5)	(6)
Conventionally tenured	0.007	0.012	0.005	0.003	0.005	-0.003
	(0.007)	(0.009)	(0.007)	(0.007)	(0.013)	(0.015)
Test-screened contract	0.002	0.002	0.005	0.002	-0.003	-0.002
	(0.011)	(0.012)	(0.011)	(0.011)	(0.017)	(0.017)
Contract	0.000	0.002	0.003	0.004	0.002	0.003
	(0.008)	(0.009)	(0.008)	(0.008)	(0.013)	(0.013)
Female		-0.019		0.006		-0.019
		(0.019)		(0.019)		(0.022)
Years of experience		0.000		-0.000		0.001**
		(0.000)		(0.000)		(0.000)
Years of education		0.000		-0.001		-0.001
		(0.002)		(0.001)		(0.001)
Cognitive skills		-0.002		-0.000		-0.002
		(0.003)		(0.002)		(0.003)
Neuroticism		-0.003		-0.001		0.001
		(0.003)		(0.002)		(0.003)
Extraversion		0.003		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		0.000
		(0.003)		(0.002)		(0.003)
CLASS mean 2011–12		-0.002		0.007**		-0.000
		(0.003)		(0.003)		(0.004)
School fixed effects	YES	YES	YES	YES	YES	YES
Observations	14909	14235	14170	13543	14495	13855
R^2	0.045	0.044	0.033	0.034	0.045	0.048
F	0.53	1.43	0.17	1.28	0.27	1.06
p	0.664	0.148	0.918	0.228	0.846	0.393

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

TABLE A6. PROBABILITY OF BEING TENURED BY A MERIT-BASED COMPETITION

Teacher Characteristics	Test-screened Tenured Teacher				
	(1)	(2)	(3)	(4)	(5)
Years of experience	0.014***	0.013***	0.013***	0.013***	0.013***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Cognitive skills	0.084***	0.063**	0.064**	0.064**	0.065**
	(0.026)	(0.029)	(0.030)	(0.029)	(0.030)
CLASS average	0.076***	0.070**			
	(0.028)	(0.029)			
Socio emotional support			0.061**		
			(0.030)		
Classroom management				0.068**	
				(0.027)	
Instructional support					0.031
					(0.031)
Female		-0.452*	-0.446*	-0.431*	-0.438*
		(0.234)	(0.235)	(0.233)	(0.238)
Years of education		0.015	0.013	0.016	0.015
		(0.016)	(0.016)	(0.016)	(0.016)
Neuroticism		-0.013	-0.011	-0.016	-0.019
		(0.028)	(0.028)	(0.027)	(0.028)
Extraversion		0.024	0.028	0.023	0.034
		(0.031)	(0.031)	(0.031)	(0.031)
Openness		0.027	0.030	0.026	0.028
		(0.030)	(0.030)	(0.030)	(0.030)
Agreeableness		0.020	0.017	0.024	0.025
		(0.032)	(0.032)	(0.032)	(0.032)
Conscientiousness		-0.029	-0.031	-0.031	-0.037
		(0.032)	(0.032)	(0.032)	(0.032)
Constant	0.121**	0.316	0.330	0.275	0.299
	(0.055)	(0.368)	(0.370)	(0.367)	(0.372)
Observations	234	234	234	234	234
R^2	0.134	0.167	0.161	0.169	0.149
F	11.902	4.471	4.276	4.534	3.899
p	0.000	0.000	0.000	0.000	0.000

Source: Authors' calculations. *Notes:* CLASS stands for Classroom Assessment Scoring System. * Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level

TABLE A7 TEACHER EXPERIENCE, COGNITIVE SKILLS, AND CLASSROOM PRACTICE AS POTENTIAL MECHANISMS OF TEST-SCREENED TENURED TEACHER EFFECT ON LANGUAGE AND MATH

	Language						Math					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Test-screened tenured	0.113	0.014	0.107***	0.098**	0.113***	0.089**	0.142*	0.032	0.082*	0.067	0.095**	0.064
	(0.080)	(0.072)	(0.040)	(0.040)	(0.036)	(0.040)	(0.082)	(0.089)	(0.046)	(0.050)	(0.041)	(0.050)
Test-screened tenured x Experience	0.001	0.006					-0.002	0.003				
	(0.004)	(0.004)					(0.005)	(0.005)				
Experience	0.005**	0.000					0.006***	0.002				
	(0.002)	(0.002)					(0.002)	(0.002)				
Test-screened tenured x Cognitive skills			-0.006	0.009					0.015	0.035		
			(0.046)	(0.031)					(0.042)	(0.029)		
Cognitive skills			0.041**	0.034**					0.012	0.004		
			(0.018)	(0.016)					(0.018)	(0.017)		
Test-screened tenured x CLASS average					0.057*	0.078***					0.074**	0.085**
					(0.031)	(0.029)					(0.032)	(0.040)
CLASS average					0.048**	0.028					0.059***	0.040*
					(0.019)	(0.018)					(0.020)	(0.022)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Parent controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Classroom controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Teacher characteristics	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Observations	12632	12632	12632	12632	12632	12632	12632	12632	12632	12632	12632	12632
R^2	0.150	0.443	0.150	0.443	0.150	0.443	0.124	0.312	0.123	0.312	0.125	0.312

Source: Authors' calculations. *Notes:* Each column reports coefficients from OLS regressions estimated with standard errors (in parentheses) clustered at the school level and school fixed effects. Columns (2), (4), (6), (8), (10), and (12) control for the student characteristics of baseline TVIP score, age, gender, and preschool attendance; the parent characteristics of years of education and living standard conditions; the classroom characteristics of class size and classroom averages of student and parent characteristics; and the teacher characteristics of gender, years of experience, years of education, tenure status, cognitive skills, personality traits, and CLASS average score. TVIP stands for Test de Vocabulario en Imágenes Peabody, the Spanish version of the Peabody Picture Vocabulary Test (PPVT). CLASS stands for Classroom Assessment Scoring System.* Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level

TABLE A8 CLASSROOM PRACTICE (CLASS DOMAINS) AS POTENTIAL MECHANISMS OF TEST-SCREENED TENURED TEACHER EFFECT ON LANGUAGE AND MATH

	Language						Math					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Test-screened tenured	0.226***	0.242***	0.200***	0.217***	0.235***	0.232***	0.006	-0.029	-0.028	-0.067	0.022	-0.009
	(0.061)	(0.053)	(0.068)	(0.058)	(0.079)	(0.057)	(0.057)	(0.050)	(0.053)	(0.050)	(0.057)	(0.043)
Test-screened tenured x CLASS Socio-emotional support	0.112**	0.108**					0.142***	0.167***				
	(0.052)	(0.045)					(0.042)	(0.048)				
CLASS Socio-emotional support	0.027	0.013					0.005	-0.027				
	(0.029)	(0.024)					(0.026)	(0.028)				
Test-screened tenured x CLASS Classroom management			0.119**	0.065					0.149***	0.141***		
			(0.055)	(0.052)					(0.041)	(0.051)		
CLASS Classroom management			-0.002	0.005					-0.005	-0.022		
			(0.037)	(0.033)					(0.027)	(0.032)		
Test-screened tenured x CLASS Instructional Support					0.089	0.052					0.116***	0.137**
					(0.058)	(0.051)					(0.044)	(0.054)
CLASS Instructional Support					0.038	0.050*					0.028	0.008
					(0.035)	(0.030)					(0.028)	(0.022)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Parent controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Classroom controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Teacher characteristics	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Observations	7123	7123	7123	7123	7123	7123	7123	7123	7123	7123	7123	7123
R^2	0.142	0.440	0.141	0.439	0.142	0.440	0.126	0.307	0.126	0.307	0.126	0.307

Source: Authors' calculations. *Notes:* Each column reports coefficients from OLS regressions estimated with standard errors (in parentheses) clustered at the school level and school fixed effects. Columns (2), (4), (6), (8), (10), and (12) control for the student characteristics of baseline TVIP score, age, gender, and preschool attendance; the parent characteristics of years of education and living standard conditions; the classroom characteristics of class size and classroom averages of student and parent characteristics; and the teacher characteristics of gender, years of experience, years of education, tenure status, cognitive skills, and personality traits. TVIP stands for Test de Vocabulario en Imágenes Peabody, the Spanish version of the Peabody Picture Vocabulary Test (PPVT). CLASS stands for Classroom Assessment Scoring System. * Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level

TABLE A9 TEACHER EXPERIENCE, COGNITIVE SKILLS, AND CLASSROOM PRACTICE AS POTENTIAL MECHANISMS OF TEST-SCREENED TENURED TEACHER EFFECT, COMPARED TO TEST-SCREENED CONTRACT TEACHERS

	Language							Math						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Test-screened tenured	0.193**	-0.195	0.132	0.377***	0.188*	0.266	0.100	0.160	-0.100	0.452	0.084	0.307***	0.068	0.070
	(0.079)	(0.129)	(0.302)	(0.130)	(0.098)	(0.224)	(0.069)	(0.105)	(0.187)	(0.319)	(0.061)	(0.097)	(0.067)	(0.120)
Test-screened tenured x Experience		0.025***	0.003						0.008	-0.015				
		(0.007)	(0.015)						(0.013)	(0.017)				
Experience	-0.006	-0.009	-0.008					-0.013	-0.003	-0.001				
	(0.007)	(0.006)	(0.016)					(0.010)	(0.015)	(0.019)				
Test-screened tenured x Cognitive skills				-0.047	0.008						-0.011	-0.197**		
				(0.095)	(0.078)						(0.086)	(0.075)		
Cognitive skills	0.080*			0.098	0.076			-0.013			0.029	0.088		
	(0.042)			(0.062)	(0.063)			(0.066)			(0.065)	(0.071)		
Test-screened tenured x CLASS average						0.069	0.231***						0.154*	0.223**
						(0.107)	(0.074)						(0.078)	(0.091)
CLASS average	0.083**					0.044	-0.101	0.194***					0.005	0.016
	(0.032)					(0.095)	(0.065)	(0.053)					(0.070)	(0.089)
School fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Student controls	YES	NO	YES	NO	YES	NO	YES	YES	NO	YES	NO	YES	NO	YES
Parent controls	YES	NO	YES	NO	YES	NO	YES	YES	NO	YES	NO	YES	NO	YES
Classroom controls	YES	NO	YES	NO	YES	NO	YES	YES	NO	YES	NO	YES	NO	YES
Teacher characteristics	YES	NO	YES	NO	YES	NO	YES	YES	NO	YES	NO	YES	NO	YES
Observations	3123	3123	3123	3123	3123	3123	3123	3123	3123	3123	3123	3123	3123	3123
R^2	0.437	0.149	0.437	0.148	0.437	0.149	0.438	0.311	0.137	0.312	0.137	0.312	0.140	0.312

Source: Authors' calculations. *Notes:* Each column reports coefficients from OLS regressions estimated with standard errors (in parentheses) clustered at the school level and school fixed effects. Columns (1), (3), (5), (7), (8), (10), (12) and (14) control for the student characteristics of baseline TVIP score, age, gender, and preschool attendance; the parent characteristics of years of education and living standard conditions; the classroom characteristics of class size and classroom averages of student and parent characteristics; and the teacher characteristics of gender, years of experience, years of education, tenure status, cognitive skills, personality traits, and CLASS average score. TVIP stands for Test de Vocabulario en Imágenes Peabody, the Spanish version of the Peabody Picture Vocabulary Test (PPVT). CLASS stands for Classroom Assessment Scoring System.* Significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level

TABLE A10 COST-EFFECTIVENESS ANALYSIS

	Language		
Teacher	Cost per year 2012	Effect	Cost-Effectiveness Ratio
Upper Bound			
Test-screened tenured	USD 11,438.00	0.169	USD 67,680.47
Conventionally tenured	USD 11,438.00	0.097	USD 117,917.53
Contract	USD 4,900.00	0.001	USD 4,900,000.00
Lower Bound (full set of controls)			
Test-screened tenured	USD 11,438.00	0.148	USD 77,283.78
Conventionally tenured	USD 11,438.00	0.047	USD 243,361.70
Contract	USD 4,900.00	0.001	USD 4,900,000.00

Source: Authors' calculations based on a public report from Ecuador's Ministry of Education (2012). *Notes:* Annual salary estimates for test-screened tenured and conventionally tenured teachers are based on the monthly unified salary for a newly tenured teacher (USD 817) in 2012, multiplied by 14, according to Ecuadorian law. Annual salaries for contract teachers are based on the average monthly unified salary for a contract teacher (USD 350) in 2012, multiplied by 14, according to Ecuadorian law. Costs in nominal US dollars.

APPENDIX B: 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 202 schools and all their kindergarten classrooms. Of these, 160 are purely control schools that do not have a 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. Accordingly, 36 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.

We repeat our original analysis using the subsample of 36 schools. Accordingly, we first evaluate the validity of the experimental design and present our results in Table B.1.

TABLE B.1. RANDOMIZATION TEST, SCHOOL SUBSAMPLE (ROBUSTNESS CHECK)

	Test-screened tenured teachers
Children:	
Age (months)	-0.002
	(0.002)
Proportion female	0.006
	(0.017)
TVIP	0.000
	(0.001)
Proportion who attended preschool	-0.029
	(0.024)
Family:	
Parents' years of schooling	0.006**
	(0.003)
Living standard indicator	0.000
	(0.015)
School fixed effects	YES
Observations	2497
R^2	0.062
F	1.25
p	0.306

Source: Authors' calculations. *Notes:* 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). * Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level

We find that none of the student or family characteristics individually predicts assignment to a test-screened tenured teacher, with the exception of parents' years of schooling. However, the size of this estimate is very small. Moreover, the F test shows that the background characteristics altogether do not predict assignment to treatment or control group.

In addition, Table B.2 shows 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, dropout or late enrollment.

TABLE B.2. NO-SHOW, ATTRITION AND LATE ENROLLMENT TESTS (ROBUSTNESS CHECK)

Teacher	No-shows		Dropouts		Late enrollments	
	(1)	(2)	(3)	(4)	(5)	(6)
Test-screened tenured	-0.005	-0.005	-0.004	-0.004	-0.004	0.001
	(0.007)	(0.007)	(0.007)	(0.006)	(0.013)	(0.015)
Female		-0.074***		0.009		-0.103***
		(0.019)		(0.015)		(0.025)
Years of experience		-0.000		0.000		0.000
		(0.001)		(0.001)		(0.001)
Years of education		-0.001		0.001		-0.001
		(0.003)		(0.002)		(0.003)
Cognitive skills		0.003		0.004		-0.008*
		(0.005)		(0.005)		(0.004)
Neuroticism		-0.001		0.000		-0.005
		(0.004)		(0.003)		(0.005)
Extraversion		0.000		0.006		-0.010
		(0.005)		(0.005)		(0.007)
Openness		0.005		-0.002		-0.008
		(0.005)		(0.005)		(0.009)
Agreeableness		-0.013***		-0.001		-0.004
		(0.004)		(0.005)		(0.012)
Conscientiousness		0.002		0.002		0.015
		(0.006)		(0.005)		(0.009)
CLASS average		0.003		0.002		-0.006
		(0.004)		(0.005)		(0.009)
School fixed effects	YES	YES	YES	YES	YES	YES
Observations	2799	2765	2697	2663	2792	2756
R^2	0.024	0.026	0.045	0.046	0.037	0.043
F	0.54	33.30	0.41	2.08	0.11	116.94
p	0.468	0.000	0.527	0.053	0.747	0.000

Source: Authors' calculations. *Notes:* 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. Dropouts 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.* Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level

Subsequently, we estimate our regression model as described in equation (1) for end-of-year language and math test scores and present them in Table B.3.

The student sample size substantially decreases from 12,632 to 2,497 kindergarten children. Nevertheless, we still find that students randomly assigned to test-screened tenured teachers achieve at least 10.2 percent of a standard deviation higher end-of-year language test scores, as shown in columns (1) to (4). In math, the effect of test-screened tenured teachers is at least 8.9 percent of a standard deviation higher end-of-year test score as shown in columns (5) to (8). To conservatively estimate our standard errors, all regressions are clustered at the school level. When the standard errors are clustered at the classroom level where treatment occurred, the significance level of the test-screened tenured teacher effect increases to at least 5 percent for language and math in all our estimates.

Results for language are virtually identical to those found in our main analysis and have further strengthened our confidence in the significant and unbiased effect of Ecuador's test-screened tenured teachers on kindergarten language learning.

TABLE B.3. ESTIMATED EFFECTS OF TEST-TENURED TEACHERS, SCHOOL SUBSAMPLE (ROBUSTNESS CHECK)

	Language				Math			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test-screened tenured	0.105**	0.104***	0.107**	0.109***	0.085*	0.092*	0.202***	0.228***
	(0.044)	(0.037)	(0.042)	(0.036)	(0.048)	(0.046)	(0.054)	(0.055)
Female			0.183**	0.281**			-0.050	0.056
			(0.088)	(0.115)			(0.085)	(0.146)
Years of experience			0.005*	0.009***			0.004	0.008***
			(0.003)	(0.003)			(0.003)	(0.003)
Years of education			-0.009	-0.005			-0.026**	-0.011
			(0.011)	(0.010)			(0.011)	(0.011)
Tenure status			0.060	0.009			-0.242***	-0.278***
			(0.079)	(0.060)			(0.073)	(0.076)
Cognitive skills				0.008				-0.028
				(0.028)				(0.033)
Neuroticism				0.008				0.055**
				(0.021)				(0.026)
Extraversion				0.054*				0.041
				(0.027)				(0.033)
Openness				0.028				0.022
				(0.019)				(0.033)
Agreeableness				0.039				0.032
				(0.029)				(0.031)
Conscientiousness				-0.045				-0.055
				(0.035)				(0.048)
CLASS average				0.037				0.066*
				(0.033)				(0.038)
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	2497	2497	2497	2497	2497	2497	2497	2497
R^2	0.105	0.426	0.427	0.432	0.045	0.257	0.261	0.266

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

APPENDIX C: Estimates for Test-Screened Tenured Teachers, Accounting for Competition Round

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

The impact of each competition may be different. On the one hand, the weighting of the competition components changed slightly over time (see Table 1). The weighting of tests increased from 45 to 55 percent, while the demonstration class decreased from 20 to 10 percent. On the other hand, the quality of tests may have varied over time and across competitions. Even though the same type of teacher skills and subject knowledge tests were applied between 2007 and 2013, there is no evidence that they were psychometrically comparable.

Table C.1 presents descriptive statistics for test-screened tenured teachers organized by the competition round. Interestingly, the test-screened tenured teachers who won the first round of the competitions scored better than their colleagues who were not screened under the new recruitment policy on all characteristics analyzed. Test-screened tenured teachers who won third-round competitions also have higher scores and qualifications, but significantly fewer years of experience. The sample size of teachers who won second-round competitions is far too small to obtain robust statistics, but we include them in our estimations to preserve the integrity of the analysis and sample size.

We estimate the effect of test-screened tenured teachers on learning outcomes accounting for the competition round. We again use an extended specification of equation (1):

$$Y_{ics} = \alpha_s + test\_tenured_{cs}^{AM\ 438-07}\beta_1 + test\_tenured_{cs}^{AM\ 018-10}\beta_2 + test\_tenured_{cs}^{AM\ 379-11}\beta_3 + X_{ics}\beta_4 + \bar{X}_{ics}\beta_5 + C_{cs}\beta_6 + T_{cs}\beta_7 + P_{cs}\beta_8 + CLASS_{cs}\beta_9 + u_{ics}, \quad (3)$$

Table C.2. presents regression results for language learning. All model specifications show that test-screened tenured teachers who won first-round competitions have a stronger effect than teachers who won third-round competitions. The size of the effect ranges from end-of-year language test scores that are 16.0 percent of a standard deviation higher in column (1) to 16.7 percent of a standard deviation higher in column (4), which takes into account the full set of child, family, classroom, and teacher covariates. In contrast, the effect of test-screened tenured teachers who won third-round competitions remains positive but is statistically insignificant.

Table C.2 also presents estimates for math learning. Once again, we find a stronger positive effect in test-screened tenured teachers who won first-round competitions. Column (5) shows that the children randomly assigned to test-screened tenured teachers who won first-round competitions scored 17.2 percent of a standard deviation higher on end-of-year math tests, which is highly significant. When additional child, family, classroom, and teacher covariates are included in column (8), this effect decreases to 16.6 percent of a standard deviation. As was the case with language learning, test-screened tenured teachers who won

third-round competitions had a positive but not statistically significant effect on math learning.

TABLE C.1 SUMMARY STATISTICS FOR TEST-SCREENED TENURED TEACHERS BY COMPETITION ROUND

	Sample Mean (SD)					Difference from non–test-screened tenured teachers (p-value)		
			Test-screen tenured			Test-screen tenured		
	Full Sample	Non-Test-screened Tenured	AM 438-07	AM 018-10	AM 379-11	AM 438-07	AM 018-10	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			

Source: Authors' calculations. *Notes*: This table reports means and standard deviations (in parenthesis) of characteristics of test-screened tenured teachers, by competition round under each separate ministerial resolution. Columns (6)–(8) display the difference between each group and non–test-screened tenured teachers (Column 2), as well as 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 using 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). * Statistically significant at 0.1 level, ** at 0.05 level, *** at 0.01 level.

Estimates of the effects of test-screened tenured teachers who won second-round competitions are very imprecise and not statistically different from zero for language and math. They should be treated with considerable caution due to the very small sample size.

In summary, our results suggest that first round of competitions regulated by the Ministerial Resolution of December 2007 recruited more effective teachers. Interestingly, these teachers were not offered substantial economic incentives linked to the merit-based competitions, since the largest salary increase for newly tenured teachers was implemented at the beginning of the third round of competitions.

TABLE C.2 ESTIMATES OF EFFECTS OF TEST-SCREENED TENURED TEACHERS ON LANGUAGE AND MATH, ACCOUNTING FOR MINISTERIAL RESOLUTION COMPETITIONS

Teacher	Language				Math			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test-screened tenured *first-round (*AM 438-07)	0.160***	0.185***	0.166**	0.167**	0.172***	0.183***	0.166**	0.166**
	(0.061)	(0.069)	(0.066)	(0.065)	(0.059)	(0.062)	(0.074)	(0.075)
Test-screened tenured *second-round (*AM 018-10)	-0.105	-0.019	-0.077	-0.014	-0.155	-0.093	-0.112	-0.075
	(0.234)	(0.180)	(0.250)	(0.240)	(0.262)	(0.210)	(0.254)	(0.238)
Test-screened tenured *third-round (*AM 379-11)	0.090	0.051	0.047	0.073	0.058	0.025	0.017	0.038
	(0.056)	(0.046)	(0.048)	(0.048)	(0.063)	(0.059)	(0.066)	(0.064)
Female teacher			0.262**	0.183			0.114	0.052
			(0.132)	(0.117)			(0.241)	(0.241)
Years of experience			0.002	0.001			0.002	0.002
			(0.002)	(0.002)			(0.002)	(0.002)
Years of education			-0.000	-0.003			-0.005	-0.006
			(0.007)	(0.006)			(0.007)	(0.007)
Tenure status			0.073**	0.050			0.073**	0.049
			(0.033)	(0.034)			(0.035)	(0.035)
Cognitive skills				0.034**				0.008
				(0.014)				(0.016)
Neuroticism				-0.000				0.011
				(0.013)				(0.014)
Extraversion				-0.010				0.009
				(0.014)				(0.014)
Openness				0.005				0.012
				(0.015)				(0.018)
Agreeableness				-0.006				-0.000
				(0.017)				(0.017)
Conscientiousness				-0.004				-0.017
				(0.015)				(0.018)
CLASS average				0.041**				0.055***
				(0.016)				(0.019)
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	12632	12632	12632	12632	12632	12632	12632	12632
R^2	0.149	0.440	0.442	0.443	0.123	0.309	0.310	0.312

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