IDB WORKING PAPER SERIES Nº IDB-WP-01199

Teacher Hiring Instruments and Teacher Value Added: Evidence from Peru

Eleonora Bertoni
Gregory Elacqua
Carolina Méndez
Humberto Santos

Inter-American Development Bank
Social Sector- Education Division

December 2020

Teacher Hiring Instruments and Teacher Value Added: Evidence from Peru

Eleonora Bertoni
Gregory Elacqua
Carolina Méndez
Humberto Santos

Inter-American Development Bank
Social Sector-Education Division

December 2020

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Teacher hiring instruments and teacher value added: evidence from Peru / Eleonora Bertoni, Gregory Elacqua, Carolina Méndez, Humberto Santos.
p. cm. — (IDB Working Paper Series; 1199).
Includes bibliographic references.
1. Teachers-Recruiting-Peru. 2. Teachers-Rating of-Peru. 3. Teacher effectiveness-Peru-Evaluation. I. Bertoni, Eleonora. II. Elacqua, Gregory M., 1972- III. Méndez, Carolina. IV. Santos, Humberto. V. Inter-American Development Bank. Education Division. VI. Series.
IDB-WP-1199

Teacher Hiring Instruments and Teacher Value Added: Evidence from Peru

Eleonora Bertoni**, Gregory Elacqua*, Carolina Méndez* & Humberto Santos**

December 2020

Abstract

In this article, we explore whether the evaluation instruments used to recruit teachers in the national teacher hiring process in Peru are good predictors of teacher effectiveness. To this end, we estimate teacher value-added (TVA) measures for public primary school teachers in 2018 and test for their correlation with the results of the 2015 and 2017 national evaluations. Our findings indicate that among the three sub-tests that comprise the first, centralized stage of the process, the curricular and pedagogical knowledge component has the strongest (and significant) correlation with the TVA measure, while the weakest correlation is found with the reading comprehension component. At the second, decentralized stage, we find no significant correlation with our measures of TVA for math, as well as non-robust correlations for the professional experience and classroom observation evaluation instruments. A positive and significant correlation is found between the classroom observation component and TVA for reading. Moreover, we find correlations between our measure of TVA and several teacher characteristics: TVA is higher for female teachers and for those at higher salary levels while it is lower for teachers with temporary contracts (compared to those with permanent positions).

*Inter-American Development Bank; ** Consultant

JEL Classification: I24, I28, J45

Keywords: Teacher Evaluation Instruments, Teacher Effectiveness, Value-Added Models, Latin America

1. INTRODUCTION

Three years after the adoption of the 2012 Teacher Reform Law (*Ley de Reforma Magisterial*, LRM), which instituted a single labor regime for all instructors in the public sector, Peru radically changed the way in which teachers are hired. The government introduced a selection process (*Concurso de Nombramiento*) based on several evaluation instruments, including competency tests and classroom observation. Prior to the reform, the hiring process in Peru lacked transparency and regional and local level officials often had significant discretion in teacher hiring and allocation decisions (Elacqua et al., 2018). With the new teacher evaluation approach, the Ministry of Education (MINEDU) sought to promote meritocratic recruitment while also signaling that being successfully appointed as a public school teacher is a challenging process.[1] The changes introduced were driven by an educational need to identify and select the most competent teachers (i.e., those who obtain the best results with their students)—to be achieved through the adoption of effective teacher evaluation instruments (Cruz-Aguayo et al., 2020).

In this article, we measure the extent to which the centralized and decentralized teacher evaluation instruments used to select public school teachers in Peru since 2015 are good predictors of teacher efficacy. To do this, we first estimate teacher value-added (TVA) measures for public primary school teachers in Peru using National Student Evaluation data (*Evaluación Censal de Estudiantes*, ECE) and the Education Information Management System database (*Sistema de Información de Apoyo a la Gestión de la Institución Educativa*, SIAGIE). The former dataset allows us to examine test score results for a panel of 2nd and 4th grade primary students. The latter is the only data source in Peru that provides classroom information, allowing us to link students

[1] Retrieved November 5, 2020 from https://andina.pe/agencia/noticia-en-enero-del-2015-habra-nuevo-concurso-publico-nombramiento-docente-501882.aspx.

with their teachers. Second, we study the relationship between the standardized evaluation instruments of the Peruvian teacher hiring process and the TVA measure. We also examine the correlations between more traditional evaluation instruments that have been explored in the literature (e.g. professional experience, years of education, type of training, etc.) and the effectiveness of teachers.

We find that among the three sub-tests that comprise the first, centralized stage of the teacher evaluation process (*Prueba Única Nacional*, PUN), the curricular and pedagogical knowledge component has the strongest (and significant) correlation with our TVA measure, while the weakest correlation is found with the reading comprehension component. This result is robust when correcting for bias due to non-random selection into teaching. Moreover, we find that the aggregate PUN score has a higher correlation with estimated TVA than the specific sub-tests, suggesting that the weighted combination of the different instruments possibly increases the ability to predict teacher effectiveness. At the second, decentralized stage, we find no significant correlation with our measures of TVA for math, as well as non-robust correlations for the professional experience and classroom observation instruments. We do, however, find a positive and significant correlation between the classroom observation component and the estimated TVA for reading. Moreover, we find correlations between our measure of TVA and several teacher characteristics: TVA is higher for female teachers and for those at higher salary levels while it is lower for teachers with temporary contracts (compared to those with permanent positions). We find no correlation between our measure of TVA and teacher experience (when greater than 3 years). These results support the central government decision to devote a higher weight to the PUN (67 percent) in the teacher evaluation total score. Our findings also point to a need to examine the

scope of the decentralized stage and, in the short run, to strengthen the monitoring of its implementation.

Our paper relates to the literature on teacher recruitment and teacher effectiveness. Recruiting (and retaining) competent teachers should be the goal of every educational institution. Yet, there is very little research that combines recruitment and retention, with a focus on teacher quality. Indeed, such studies have been hindered by both the difficulty of establishing an agreed-upon definition of teacher quality, and the scarcity of data that allows to identify effective teachers and examine the factors that promote their recruitment and retention (Guarino et al., 2004). This study aims to bridge this gap by combining unique information on teacher recruitment and teacher effectiveness in Peru.

There is extensive debate over the quality of the instruments used to select new teachers. Traditionally, teacher hiring choices for permanent positions are based on applicants' professional and educational backgrounds. Although these characteristics are easy to observe and evaluate, the evidence suggests that neither traditional academic certificates (Aaronson et al., 2007; Clotfelter et al., 2007; Clotfelter et al., 2010; Goldhaber and Brewer, 1999; Hanushek and Rivkin, 2010; Leigh, 2009; Rivkin et al., 2005; Slater, Davies and Burgess, 2012) nor years of experience - after the initial years - (Araujo et al., 2016, Clotfelter et al., 2010; Harris and Sass, 2011; Rivkin et al., 2005) are good predictors of teaching effectiveness.

Given the above, several educational systems have begun to consider additional instruments for selecting teachers, such as standardized tests of basic knowledge (generally mathematics and reading) and/or specific curricular and pedagogical knowledge, practical evaluations such as classroom observation, and interviews with the school principal or other officials. Various studies show that knowledge tests and standardized classroom observations are correlated with greater

teacher effectiveness (Bruno and Strunk, 2019; Kane et al., 2011; Kane and Staiger, 2012), as are interviews with the school principal or other officials (Harris and Sass, 2014; Jacob and Lefgren, 2008). Moreover, a number of scholars find that teachers' scores on knowledge tests are directly associated with higher student learning (Bietenbeck et al., 2017; Clotfelter et al., 2006; Clotfelter et al., 2007; Hanushek et al., 2017; Metzler and Woessmann, 2012), particularly in the subject they teach. Furthermore, aside from having the advantage of being less controversial than standardized tests (Cruz-Aguayo et al., 2020), empirical evidence indicates that classroom observation is a good predictor of teaching quality, particularly for tenured teachers (Araujo et al., 2016; Jacob et al., 2016; Kane et al., 2011; Kane and Staiger, 2012; Milanowski, 2004; Taut et al., 2016; Tyler et al., 2010). That said, Cruz-Aguayo et al. (2017) find no relationship between a teacher's score on the classroom observation instrument and student learning in Ecuador. Such mixed findings suggest the need to better understand how demonstration classes are implemented and evaluated. Finally, although little research has been conducted on the interview instrument, Jacob et al. (2016) find that, in Washington D.C., structured interviews conducted by staff and teachers in the school district during the hiring process are related to greater teaching effectiveness.

In the Latin American context, various combinations of assessment instruments are used to select candidates. Bertoni et al. (2020) analyze the hiring systems of 12 countries in the Latin America and the Caribbean region (LAC), including Peru. They observe that in most of these countries, teachers are required to have a teaching degree or a professional degree in a specific field. In addition to their educational background, teachers' experience is also considered. In Peru, Brazil, Colombia, Ecuador, and Honduras new teachers must pass a standardized assessment. Classroom observation is not, however, a widespread practice in the process of selecting teachers,

used only in Peru, Rio de Janeiro, and Ecuador. Finally, in addition to Peru, Barbados, Chile, Colombia, Honduras, and Jamaica also include interviews as part of their teacher hiring processes.[2]

Our study also contributes to the growing body of work on the predictive power of teacher selection instruments. There is an extensive literature that examines whether teacher screening instruments are correlated with a measure of teacher effectiveness estimated through value-added models (VAM).[3] This paper is the first to estimate TVA in Peru through the use of administrative data sources (i.e., SIAGIE and ECE). In LAC school systems generally, student assessments are usually not designed to be comparable over time, making such studies challenging to conduct. This article also represents the first assessment of whether the teacher evaluation instruments currently used in Peru to select instructors predict teacher effectiveness. While the use of VA measures for accountability purposes is somewhat controversial (e.g. Rothstein 2010), the objective here is to evaluate how well the teacher evaluation instruments in Peru identify higher performing teachers. Finally, we shed light on the differing predictive power of evaluation instruments implemented at the centralized vs. the decentralized level. Peru is one of the only systems in the world that includes both a centralized and a decentralized stage in its process of selecting and assigning teachers to schools (Bertoni et al., 2020; OECD, 2014).[4]

[2] A similar study on 25 of the 35 OECD countries shows that, for the most part, graduates of teacher education programs can begin working directly at the primary, lower secondary, and upper secondary levels. Only France, Korea, Mexico, Spain, and Turkey require candidates to pass a competitive examination before they can begin their teaching careers. In Japan and Greece, candidates must both pass an exam and acquire a license. In Luxembourg (pre-primary and primary levels), candidates need to pass an examination as well as a standardized reading test in the three official languages. Finally, in Australia and Austria (academic secondary school, lower level, and upper secondary level), applicants have to acquire a license to start teaching (OECD, 2014).

[3] Although there has been much discussion over the best way to measure teacher effectiveness, researchers have often used value-added modeling as a tool to disentangle teachers' individual contributions. The intuitive idea behind this approach is that prior achievement can be used as a control for the history of previous inputs and, in some models, the ability endowment (Koedel, Mihaly and Rockoff, 2015).

[4] Most other systems that also use classroom observation instruments as a means of selecting teachers (e.g., Colombia and Ecuador) do so at the national, rather than at the school level.

The remainder of the paper proceeds as follows. Section 2 provides background information on the teacher selection process in the Peruvian public school system. Section 3 describes the empirical strategy employed while Section 4 introduces the data and discusses possible biases. Section 5 presents the main results and Section 6 concludes.

2. TEACHER HIRING PROCESS IN THE PERUVIAN PUBLIC SCHOOL SYSTEM

In 2015, the Peruvian government instituted a new evaluation system for aspirant instructors seeking a tenured teaching position in the public school system. To be eligible to participate, candidates must hold a bachelor's degree in education. The evaluation consists of two stages: the first is centralized and carried out at the national level, and the second is decentralized, at the school level. The centralized stage is carried out by the Ministry of Education (MINEDU) and includes a standardized written test (*Prueba Única Nacional*, PUN) that is divided into three sub-tests, which carry different weights in the aggregated score: logical reasoning (25 percent), reading comprehension (25 percent), and curricular and pedagogical knowledge (50 percent). To pass the centralized stage, candidates need to answer at least 60 percent of the questions correctly on each sub-test. Applicants are also evaluated in a specific area of specialization relative to the education level (pre-primary/primary/secondary) and subject (e.g., secondary science) they plan to teach.

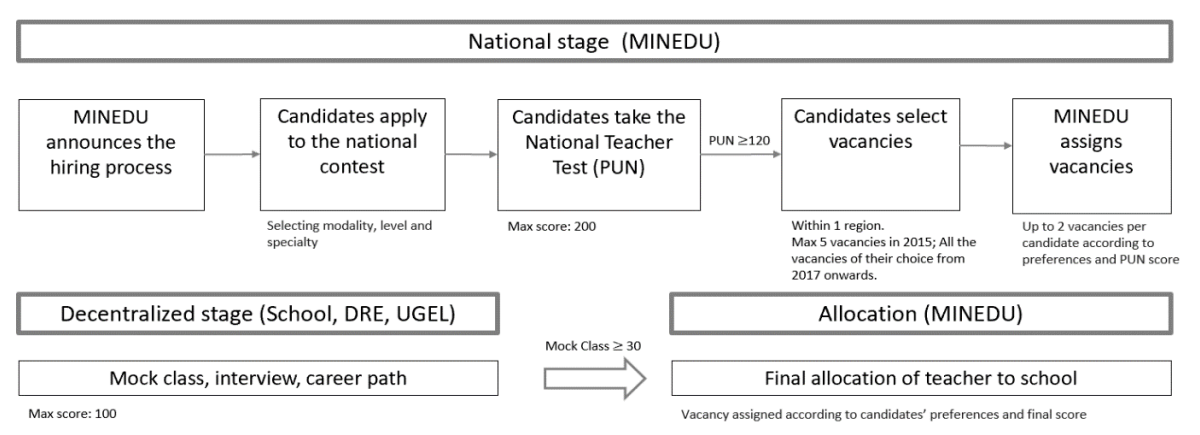
Only candidates that score above the required threshold at the centralized stage can then rank their school preferences, chosen within their area of specialization and in one of the 26 regions of Peru. In the 2015 evaluation process, candidates could list a maximum of 5 school preferences, which became an unlimited number from 2017 onwards. Once the preferences are established, the Ministry of Education assigns each candidate to a maximum of 2 (in 2015) or of 3 (in 2017) of their preferred schools, based on their PUN score and their preference ranking. In 2015, each vacancy could have up to 20 candidates; this was reduced to 10 in 2017.

Once candidates have been assigned to their preferred schools, they begin the decentralized stage, which is carried out by each school or by the local education administrative units (*Unidades de Gestión Educativa Local*, UGEL) in the case of single-teacher institutions. This stage consists of three instruments that have different weights in the aggregated score: an evaluation of a candidate's resume (25 percent), a personal interview (25 percent), and a classroom observation (50 percent). To pass the decentralized stage, candidates must score a minimum of 30 (out of 50) points on the classroom observation component.

Finally, the Ministry of Education uses the weighted sum of the scores obtained in the centralized and decentralized stages (the centralized stage has a weight of 67 percent on the final score) to assign the vacancies in order based on merit and on the candidate's preferences.[5] Figure 1 summarizes the teacher hiring process in Peru.

Once the assignment process has been completed, the candidates who did not manage to obtain a permanent teaching position are able to apply for a temporary position. At this point, the candidates are evaluated solely according to their final score at the centralized stage, where no minimum passing score is required. They are hired through a public tender that takes place in each UGEL. Specifically, candidates select one UGEL of their preference that has vacancies in their area of specialization on the PUN. They are then ranked by "merit" in descending order according to their score at the centralized stage. Those with the highest score are the first to choose among the available vacancies of that UGEL.

[5] In case of a tie in the final score for the same vacancy, the Ministry of Education applies the following criteria in order of priority to identify a single winner for each vacancy: (1) higher score on the classroom observation; (2) higher score on the curricular and pedagogical knowledge sub-test; (3) higher score on the resume's academic and professional training; (4) higher score on the resume's professional experience; (5) higher score on the resume's merits. If the same applicant wins for more than one vacancy, the Ministry of Education automatically assigns the vacancy with the highest priority level, according to the preferences of the applicant.



Source: Bertoni et al. (2019)
Note: Ministry of Education - *Ministerio de Educación* (MINEDU). National Teacher Test - *Prueba Única Nacional* (PUN*).* Regional Education Directorates - *Dirección Regional de Educación* –(DRE). Local Education Management Units - Unidad de Gestión Educativa Local (UGEL).

Figure 1. Teacher Hiring Process in Peru.

2.1. Teacher Evaluation Instruments

This section presents a more detailed description of the evaluation instruments used in the Peruvian teacher hiring process. Table 1 reports the thresholds of the centralized and decentralized stages of the hiring process, the implications of the results obtained, and the authority responsible for evaluating the respective sub-tests.

The centralized stage (PUN) includes 3 sub-tests consisting of multiple-choice questionnaires (three possible answers for each question): (1) reading comprehension, which comprises 5 texts with 5 questions for each (25 questions in total, 2 points for each correct answer); (2) logical reasoning, made up of 25 basic mathematics and logic problems (2 points for each correct answer); and (3) curricular and pedagogical knowledge, consisting of 40 questions about the process of learning reading and mathematics, where the candidate is asked to evaluate hypothetical classroom situations in order to analyze the pedagogical techniques involved (2.5 points for each correct answer). Applicants have a maximum of 4 hours and 30 minutes to complete the test. Incorrect or blank answers are given a score of zero (0).

The decentralized stage is administered by each school's Evaluation Committee, which oversees the implementation of a second set of instruments: (1) evaluation of the candidate's resume; (2) interview; and (3) classroom observation. The Committee is composed of: (i) the director, owner, or manager of the school; (ii) the assistant director or another professor named at the same level or type of school as the evaluated one; and a teacher named at the same level or type of school as the evaluated one.[6] Following each evaluation, the Committee enters the candidate's score on the online platform managed by the MINEDU, which automatically compiles the applicant's final score.

With specific regard to the evaluation of the candidate's resume, the Evaluation Committee reviews, in the presence of the applicant, the documents presented to verify the professional experience declared. This process takes place together with the interview. The candidate is given a score according to three aspects: (i) academic and professional training; (ii) merits (publications, participation in conferences, etc.); and (iii) professional experience. The maximum score for this instrument is 25 points and no minimum score is required.

The interview seeks to assess the suitability of the applicant in terms of the school's educational project and the candidate's teaching vocation. To ensure equal assessment, it is recommended that all applicants for a given position are interviewed by the same members of the Committee. The interviews are rated based on criteria established by each Committee, according to the guidelines provided by the MINEDU, which grants up to 15 points for "fit with the school's educational project" and up to 10 points for "teaching vocation."

[6] The Evaluation Committee in single-teacher or multi-grade multi-teacher school is made up of: i) the head of the Pedagogical Management Area or her representative or an Education Specialist from the UGEL of the same level or type of school as the one evaluated, ii) the Director of the Educational Network or an Education Specialist from the UGEL of the same level or type of school as the one evaluated; and iii) a teacher named at the same level or type of school as the one evaluated.

The classroom observation is also organized by the Evaluation Committee and implemented according to the "Classroom Observation Rubrics for EBR applicants (Annex N ° 1 of the *Manual del Comité de Evaluación*)" guidelines. It takes place during a learning session that lasts between 45 and 90 minutes. Depending on the number of applicants, one or two Committee members conduct the observation. The candidate is evaluated relative to 5 dimensions: (i) Actively involves students in the learning process; (ii) Promotes reasoning, creativity and/or critical thinking; (iii) Evaluates learning progress so as to provide feedback to students and adapt teaching; (iv) Fosters an environment of respect and proximity and (v) Positively regulates student behavior. Each of the five dimensions is assigned one of 4 ascending scores: Level I (very poor, 1 point), Level II (in progress, 2 points), Level III (sufficient, 3 points) and Level IV (notable, 4 points).[7] The Committee observers closely follow the development of the session without intervening, in the meantime taking notes using the "Applicant Evaluation Protocol" provided by the MINEDU.

[7] The EBR score on this instrument is obtained by taking the sum of the rating given on each of the 5 dimensions multiplied by a factor of 2.5.

Table 1. Teacher evaluation instruments and corresponding thresholds

	Min/Max score	Use of result	Responsibility
Centralized stage *(67%)*	**120/200**	Eliminatory and ranking	MINEDU
Reading comprehension (25%)	30/50		
Logical reasoning (25%)	30/50		
Curricular and pedagogical knowledge (50%)	60/100		
Decentralized stage *(33%)*	**30/100**		School's Evaluation Committee
Classroom observation (50%)*	30/50	Eliminatory and ranking	
Actively involves students in the learning process	1/4		
Promotes reasoning, creativity, and/or critical thinking	1/4		
Evaluates learning progress to provide feedback	1/4		
Fosters an environment of respect and proximity	1/4		
Positively regulates student behavior	1/4		
Interview (25%)	-/25	Ranking	
Suitability for school's educational project (60%)	-/15		
Teaching vocation (40%)	-/10		
Resume (25%)	-/25	Ranking	
Academic and professional training (40%)	-		
Merits (20%)	-		
Professional experience (40%)	-		
Total	**150/300**		

* Scores are multiplied by a factor of 2.5.

Source: MINEDU Evaluation Committee Manual, 2017[8]

In order to assess the interplay between the two stages of the teacher hiring process, Table 2 reports correlation coefficients among the different teacher evaluation instruments for candidates that participated in both evaluation stages. Overall, we observe a very low correlation between the centralized and the decentralized stages, although it is positive and significant. Moreover, when examining the disaggregated components of the decentralized stage, we observe a non-significant correlation between the centralized stage and the classroom observation instrument, despite the latter carrying the greatest weight among the decentralized stage instruments. This may be related to the different skills evaluated in the two stages, but could also be explained by the higher

[8] http://evaluaciondocente.perueduca.pe/media/11566233590Manual-del-Comit%C3%A9-de-Evaluaci%C3%B3n-2019.pdf

discretion and heterogeneity involved in implementing the evaluation instruments at the school level. The highest correlation between the disaggregated evaluation instruments is found between the classroom observation and the interview, probably because they both entail direct personal interaction with the candidate and, in taking place on the same day, the composition of the Evaluation Committee is unlikely to change.

Table 2. Correlations of teacher evaluation instruments

	Total PUN score	Reading comprehension	Logical reasoning	Curriculum knowledge	Total decentralized Stage	Classroom observation	Interview
Centralized stage							
Total PUN score	1.000						
Reading comprehension	0.620***	1.000					
Logical reasoning	0.668***	0.256***	1.000				
Curriculum knowledge	0.812***	0.257***	0.254***	1.000			
Decentralized stage							
Total decentralized stage	0.073***	-0.024	0.062**	0.090***	1.000		
Classroom observation	0.044	-0.015	0.059**	0.040	0.870***	1.000	
Interview	0.049*	-0.004	0.077**	0.030	0.780***	0.606***	1.000
Professional experience	0.076***	-0.032	-0.002	0.143***	0.513***	0.126***	0.157***

***p<0.01; **p<0.05; *p<0.1.

Figure 2 plots, by year, the probability of being employed the year after the teacher hiring process and candidate performance on the three different tests of the centralized evaluation stage. The sample refers to all candidates who participated in the 2017 selection.[9] Note that the probability of being employed is different from zero below the cutoffs given that, as explained above, applicants who do not pass the centralized stage can still be hired as temporary teachers at the end of the hiring process. In 2015, 30 percent of teachers in the public sector held a temporary contract (Nexus, 2015). Moreover, 61 percent of the candidates that selected vacancies for a permanent position were already temporary teachers in a public school, having on average 4 years of experience in the public sector and 2 in the private sector.

[9] Figures for the 2015 teacher hiring process are similar.

We observe that the probability of being employed is positively correlated with the instruments of the centralized evaluation stage, and this increases passing the threshold of each test. Additionally, there is a flat relationship below the threshold for all but the curriculum knowledge component, where the probability of being employed starts to rise at around 20 points below the cutoff in both years. This could be explained by the fact that, by construction, this is the component that carries the greatest weight in the centralized stage. Moreover, the Ministry of Education relies on this score, among others, as a priority criterion to break ties among candidates (See footnote 5).

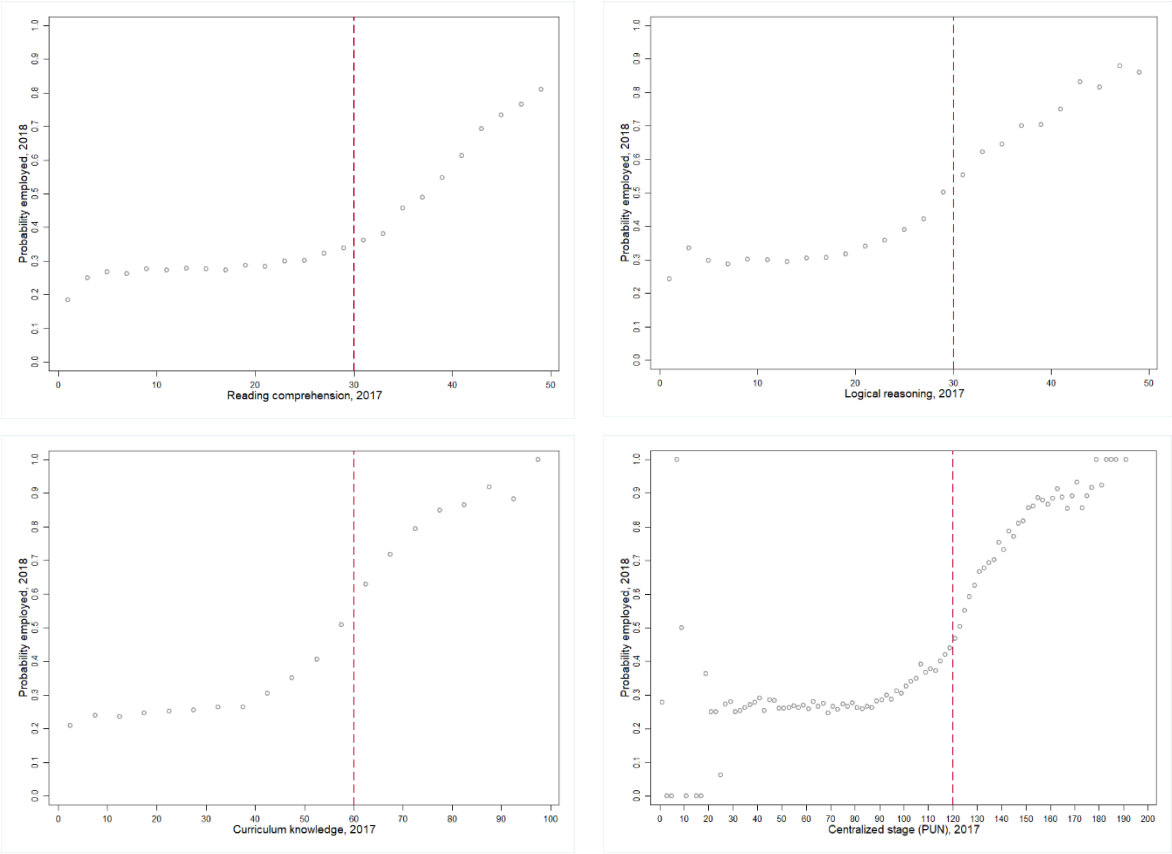


Figure 2. Distribution of teacher evaluation scores and the probability of being employed, 2017

3. METHODOLOGY

3.1. Teacher Value-Added (TVA)

A wide variety of value-added models (VAMs) have been estimated in the literature. Borrowing from the linear VAM of Koedel et al. 2015, we first estimate a basic TVA model of the form:

$$Y_{isjt} = \beta_0 + Y_{isjt-2}\,\beta_1 + X_{it}\,\beta_2 + S_{st}\,\beta_3 + \theta_j + \varepsilon_{isjt} \quad (1)$$

Where Y_{isjt} is the 2018 ECE test score in math or reading for 4th grade primary student *i* at school *s* that is taught by teacher *j* ; Y_{isjt-2} is the student's 2nd grade ECE score in either math or reading in 2016; X_{it} is a vector of student characteristics such as socioeconomic status (a dummy equal to 1 if the student's mother completed a secondary education) and a set of dummy variables that reflects the student's grades in the respective subjects in second grade[10]; S_{st} is a vector of school characteristics such as a dummy equal to 1 if the school is rural and a continuous variable for the school's total enrollment in 2018; θ_j is a vector of teacher indicator variables; and ε_{isjt} is the idiosyncratic error term. The parameters that capture teacher value-added are contained in the vector θ_j and are obtained by estimating the coefficient associated with the fixed effect of the teacher linked to the student. Specifically, the estimator of θ is defined as follows[11] (McCaffrey et al., 2012):

$$\hat{\theta}_j = (\bar{y}_j - \bar{x}'_j\hat{\beta}) - (\tilde{y}_{\cdot} - \tilde{x}'_{\cdot}\hat{\beta})$$

[10] In 2016, the grades were expressed using the following categories: AD (outstanding achievement), A (expected achievement), B (in progress) and C (initial).

[11] To estimate $\hat{\theta}_j$ we use the *areg* Stata command described in McCaffrey et al. (2012). Given that models with fixed effects for units are overparameterized, because the means for the individual units cannot be estimated separately from the mean of the individual persons, Stata commands have different approaches to solving the indeterminacy: (i) estimation of unit means that combine unit means with person means; (ii) estimation of contrasts between each of the unit means and the mean of a "holdout" unit; and (iii) estimation of contrasts between each of the unit means and the average of the unit means. All three lead to the same rank ordering of units by estimated unit fixed effects. However, they do not provide estimates of the same quantities and are not all equally appropriate for all purposes.

$$j = 1, \ldots, J$$

where $\bar{y}_j$ is the average 2018 4th grade ECE score of the students taught by teacher $j = 1, \ldots, J$ and $\tilde{y}_.$ is the average 2018 4th grade ECE score of all students in the sample; $\bar{x}_j$ is the average students' characteristics of teacher j; and $\tilde{x}_.$ is the average value of students' covariates in the sample. Finally, $\hat{\beta}$ is the vector of estimated coefficients of the control variables included in equation (1).

Equation (1) is written as a "lagged-score" VAM.[12] Several value-added studies have incorporated student and/or school fixed effects into variants of equation (1) due to concerns about bias from nonrandom student-teacher sorting.[13] By including prior attainment, we control for some of the school effect, the impact of previous teachers' inputs, as well as for student's own ability and prior effort (Slater, Davies, and Burgess, 2012).

A challenge in the case of Peru is that the ECE student assessment is not consecutive. That is, between the two tests (2nd and 4th grade) a student may have been taught by different teachers. The literature provides several options to address this problem. We follow Hock and Isenberg (2017) and estimate single effects for each teacher by applying weighted least squares (WLS) to equation (1), with the weights equal to "teacher dosages"[14] where the dosages are the percentage

[12] An alternative, restricted version of the model where the coefficient on the lagged test score is set to unity (β1=1) is referred to as a "gain score" VAM. The "gain-score" terminology comes from the fact that the lagged-score term with the restricted coefficient can be moved to the left-hand side of the equation and the model becomes one of test score gains.

[13] These concerns are based in part on empirical evidence showing that students are not randomly assigned to teachers, even within schools (e.g., Kalogrides and Loeb, 2013; Treviño et al, 2018).

[14] For example, if Teacher A individually taught 15 students and shared another 10 students equally with Teacher B, the regression would produce a Teacher A-alone effect and a Team-AB effect. Assuming both teachers receive half credit for each shared student, the total share of students from this group for Teacher A would be 0.5 × 10 = 5. Thus, in the example above, the Full Roster Method would produce a regression coefficient (i.e., teacher's overall effectiveness measure) for Teacher A that is precisely equal to 15/20 × [Teacher-A-alone estimate] + 5/20 × [Team AB estimate].

of instructional time that each teacher spends with the student. Following Hock and Isenberg (2017) and other studies (for example, Slater, Davies and Burgess, 2012), we assume that the time is divided equally between teachers (for example, if a student had one math teacher in third grade and a different one in fourth grade, each is assigned a weight of 0.5).

3.1.1. Extensions to Basic VAM

We extend this basic model in three ways. First, we introduce the score of the student in the other subject (e.g., in the estimation of value-added in math, the student's ECE language score in second grade). Second, we add non-linearities in both subjects (i.e., squared of math and second grade ECE language scores) and the interaction between them. There is evidence that adding this set of variables reduces bias in the estimation of the value-added measures, compared with a model using only the lagged outcome in the same subject (e.g. Ehlert et al., 2014; Lockwood and McCaffrey, 2014). Finally, we also estimate all of the value-added measures employing a two-step VAM, or "average residuals" VAM (Koedel et al., 2015). The two-step VAM uses the same information as equation (1) while performing the TVA estimation in two stages. In the first stage, we estimate the model in equation (1) without teacher fixed effects. This allows us to account for students´ characteristics aggregated at the classroom level, which could not be added in our one-step model estimated over one cohort of students. Specifically, we estimate a model of the form:

$$Y_{isjt} = \beta_0 + Y_{isjt-2}\beta_1 + X_{it}\beta_2 + S_{st}\beta_3 + C_{st}\beta_3 + \eta_{isjt} \qquad (2)$$

where all variables are defined as in equation (1), and C_{st} is a set of student variables aggregated at the classroom level (percentage of high-SES students, average ECE math and reading scores in 2nd grade). In the second stage, estimated residuals from equation (2) are used as a dependent variable in a regression against teacher fixed effects:

$$\eta_{isjt} = \theta_j + \varepsilon_{isjt} \quad (3)$$

Where the vector θ_j contains TVA estimates. The key feature that distinguishes the one-step from the two-step model is that the latter partials out the variation in Y_{isjt} attributable to lagged test scores and to other controls before estimating the teacher effects. In this way, any differences in teacher performance correlated with the covariates, at either the individual or unit-of-analysis level, will be attributed to the covariates. In other words, the two-step model essentially equalizes competing units based on observable student characteristics prior to comparing value-added between units. Note that the one-step model potentially conflates the unit effects with other factors (e.g., class composition). Meanwhile, the two-step procedure has the potential to "over correct" for observable differences between units. The direction and magnitude of the bias in each model cannot, however, be determined with certainty as the underlying true values of the unit effects are unknown. That said, the potential bias in the one-step model likely favors advantaged schools, while the potential bias in the two-step model likely favors disadvantaged schools.[15]

3.1.2. Teacher Characteristics and TVA

Following Bau and Das (2020), we estimate the association between estimated TVA and observed teacher characteristics for public school instructors using the following specification:

$$TVA_j = \beta_0 + X_j\beta_1 + \alpha_s + \varepsilon_j \quad (4)$$

where TVAj is a teacher j's average value-added over math and reading; Xj includes teacher characteristics such as sex, age, age squared, an indicator variable for whether a teacher

[15] Ehlert et al. (2016) argue that the two-step model is better suited for achieving key policy objectives in teacher evaluation systems, including the establishment of an incentive structure that maximizes teacher effort, as it overcorrects by context.

has a temporary contract, dummy variables for salary scales, and, where available, an indicator variable for having three or more years of experience in 2018.[16] α_s is a fixed effect for schools.

3.2. Correlations Between TVA and Teacher Evaluation Instruments

After estimating the VAM, we run pairwise correlations between the estimated TVA and the different respective instruments of the centralized (i.e., reading comprehension, logical reasoning, curricular and pedagogical knowledge, and aggregated PUN score) and decentralized (i.e. classroom observation, interview, and professional experience) stages of the teacher evaluation. If a positive and significant correlation exists, we can conclude that the evaluation instruments used in Peru are adequately identifying and selecting the most effective teachers.

3.2.1 Accounting for selection into teaching

Following Jacob et al. (2018) we seek to estimate the true correlation between teacher evaluation instruments and TVA, purged of the bias due to non-random selection into teaching (i.e., we can only estimate TVA for the instructors who are actually teaching in a public school after the hiring process). To this end, we use the sharp discontinuity around the cut-offs of the centralized stage to predict the probability of being hired, as an exclusion restriction in a parametric selection correction. Using the data of all applicants in the 2015 and 2017 teacher recruitment years, we first estimate candidates' predicted probabilities of being hired using the three PUN cut-offs as instruments through a linear probability model of the form:

$$T_i = \beta_0 + X_i \beta_1 + S_i \beta_2 + C_i \beta_3 + \varepsilon_i \quad (5)$$

[16] Information on years of experience is only available for the set of teachers that participated in the teacher hiring process in either 2015 or 2017.

where T_i takes the value of 1 when teacher i is teaching in a public school the year after the hiring process (2016 or 2018); X_i is a vector of teacher characteristics such as sex and age, a dummy equal to one when the teacher has some prior experience (public and/or private), and indicator variables for teacher education program (institute, university, or both). S_i is the vector of scores on the three PUN instruments (reading comprehension, logical reasoning, and curricular and pedagogical knowledge), and C_i is a set of indicator variables taking the value of one if the candidate surpassed the minimum required score on each test (30/50 points, 30/50 points, and 60/100 points respectively). Results of this first-stage estimation are reported in Table A1 in the Appendix. All else equal, women have between an 8 and 13pp lower probability of being employed, while having some previous teaching experience increases the probability by around 21pp. Meanwhile, high variation in the quality of the teacher education program seems to penalize the probability of being employed. Consistent with that observed in Figure 2, the curriculum knowledge component is positively and strongly correlated with the probability of being employed by around 0.4pp.

In a second stage, we include these predicted probabilities as a control function in the correlation model between evaluation instruments and TVA to account for any potential selection associated with the probability of being hired.[17]

4. DATA

We employ several different data sources. First, teacher and student information come from the Education Information Management System (*Sistema de Información de Apoyo a la Gestión de la Institución Educativa*, SIAGIE) databases for the period 2016-2018. This is the only data

[17] This is an extension of the traditional approach based on the Inverse Mills Ratio, with the identifying assumption that the instruments (PUN cut-offs) are associated with the likelihood of being hired, but do not directly influence the outcome (Jacob et al., 2018).

source in Peru that provides classroom level information, allowing us to link students with the instructors that teach a particular subject in a given classroom.

Second, using the National Student Evaluation (*Evaluación Censal de Estudiantes*, ECE) database, we compute individual results for primary student performance on the standardized math and reading tests. The census nature of these data allowed us to link the scores in second (ECE 2016) and fourth (ECE 2018) grade of primary school for the same student for each subject. Moreover, we use information collected through the ECE Parent Survey[18] to construct socioeconomic status (SES) indicators at the student level.

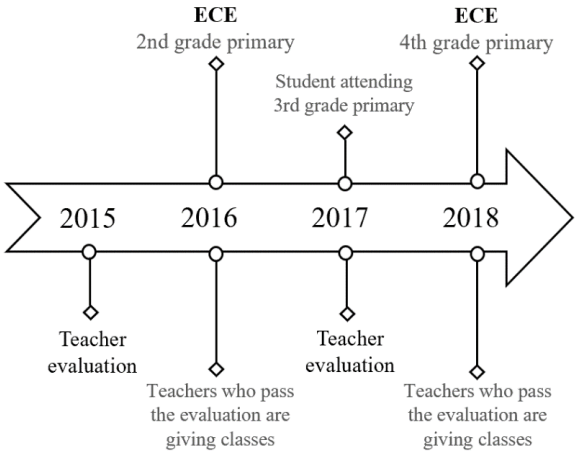
A third source of data consists of the results of the 2015 and 2017 teacher selections. This database includes individual-level applicant characteristics and detailed evaluation results for all participants in the hiring process. We retrieved teachers' scores on the centralized (PUN) and those received at the decentralized evaluation stage and estimate their correlation with our VA measures.

Fourth, we use school-level data from the 2017 and 2018 National Educational Census (*Censo Educativo*) database. This database includes school characteristics such as: area (urban/rural), type of school (i.e., single-teacher, multi-grade, or multi-teacher), an indicator of whether the school is bilingual, total enrollment, and access to basic services, among others.

Finally, we complement the teacher-level information with the 2018 Vacancy Management and Control System database (*Sistema de Administración y Control de Plazas*, Nexus), which provides information on gender, age, workday, salary level, and type of contract for all teachers in the system.

[18] This is a background questionnaire that complements the ECE. For primary students, parents fill out the survey, which includes questions on the socioeconomic status of the student's household.

Figure 3 summarizes the timeline of the processes explored in this analysis. For the calculation of teachers' individual VA, we identify the instructors who, in 2017 and 2018, taught the 2018 cohort of fourth grade public primary students in Peru. Thanks to the structure of the ECE, it is possible to trace the scores of the same cohort of students back in 2016. Moreover, we take advantage of the 2015 and 2017 teacher evaluation processes to estimate the correlation between the evaluation score and the VA measure of the instructors who were teaching in 2018.



Source: Authors' own elaboration.

Figure 3. Timeline of teacher and student evaluations

4.1. Data Limitations and Sample Restrictions

In this section, we discuss the data limitations and the sample restrictions that apply to this study. First, the information reported in the SIAGIE is not directly collected by the MINEDU but is generally compiled by the school principals. This could lead to inaccuracies, given that carrying out this task possibly represents an additional burden to principals' already numerous responsibilities. Indeed, when compared to the magnitudes reported in the National Educational Census collected by the MINEDU, teacher and student data in the SIAGIE appear to be underreported. This is especially true for teacher data. For example, as Table 3 shows, in 2016 the

National Educational Census reported a total of 143,538 public primary teachers while the SIAGIE reported less than half that number. Contrary to the SIAGIE, the Nexus database is administered by the UGEL and, although slightly incomplete, reports teacher data magnitudes that more closely resemble those presented in the Census.

Table 3. Teacher and student data coverage comparison (primary, public)

Source of data:	**Census**		**SIAGIE**		**Nexus**	
	2016	**2018**	**2016**	**2018**	**2016**	**2018**
Managed by:	MINEDU		School Principal		UGEL	
N. Teachers	143,538	147,369	70,737	99,068	130,667	132,614
N. Students, 2nd grade	471,785	467,889	468,298	455,486	-	-
N. Students, 4th grade	427,104	454,512	424,666	455,486	-	-
N. Schools[1]	29,565	29,741	8,989	12,375	29,593	29,683

Source: MINEDU 2016, 2018; SIAGIE 2016, 2018; Nexus 2016, 2018.
Note: [1]The reported number of schools for the SIAGIE dataset refers to the SIAGIE teacher dataset. The number of schools in the SIAGIE student dataset resembles that of the Census data.

Given these limitations, Table 4 presents an overview of the sample restrictions. The population of potential students includes the 454,512 Peruvian 4th grade public primary school students who were evaluated in the ECE in 2018. This is the maximum number of students that could be included in VAMs if the data were perfect (e.g., availability of 2nd grade primary scores, links to teachers and classrooms, etc.). Similarly, the teacher population of interest includes around 26,800 teachers[19] who taught 3rd grade math or reading in 2017 and/or 4th grade math or reading in 2018.

First, use of the SIAGIE data restricts the population of 4th grade primary teachers by around 33 percent, while the population of 4th grade primary students is quite similar (the upward difference of around 1 percent could be due to grade miscoding). Second, the teacher-student link restricts the sample of students by 22 percent and that of teachers by an additional 1 percent. Third, given the

[19] The MINEDU does not collect data of teachers by grade at the national level. To get an estimate of this number we are assuming a proportion of 4th grade primary teachers over the total number of primary teachers similar to the one of the SIAGIE data, that is, around 18%.

restrictions imposed by use of the SIAGIE data, a panel dataset of 2nd and 4th grade student test-scores (necessary to estimate the lagged-score TVA) is ultimately available for 42 percent of 4th grade primary students. This could be due to reasons such as movement of students across schools, repetition and/or dropout rates, as well as inconsistencies in the student ID variable between the SIAGIE and the ECE dataset. The availability of the student panel restricts, in turn, the sample of 4th grade primary teachers by an additional 28 percent. Last, we look at sample restrictions imposed by use of the 2015 and 2017 teacher evaluation data. As Table 2 shows, around 23 percent of the instructors teaching in 2018 participated in the teacher evaluation in 2015 and/or 2017 and 6 percent of those made it to the second, decentralized stage. Once restrictions (1)-(3) are applied to our data, out of the sample of 10,415 4th grade primary teachers for whom we can estimate TVA, it is possible to recover centralized and decentralized teacher evaluation data for 23 percent (2,434) and 8 percent (880) of those, respectively.

Table 4. Sample restrictions.

Populations of interest (PoI) - Primary, public	**2018**	**%PoI**
▪N. Students, 4th grade	454,512	100%
▫N. Teachers	147,369	100%
▫▫N. Teachers, 4th grade	26,812	100%
▫N. Teachers evaluated in 2015/2017 - Centralized stage	33,524	23%
▫N. Teachers evaluated in 2015/2017 - Decentralized stage	8,838	6%
Sample restrictions		
(1) SIAGIE data		
▪N. Students, 4th grade	455,486	100%
▫N. Teachers	99,068	67%
▫▫N. Teachers, 4th grade	18,024	67%
*(2) Teacher-Student SIAGIE link**		
▪N. Students, 4th grade	354,426	78%
▫▫N. Teachers, 4th grade	17,635	66%
(3) Student test-score data (ECE)		
▪N. Students, 4th grade	325,772	72%
VAM sample		
▪N. Students, 4th grade with 2nd grade score (Panel)	192,226	42%
▫▫N. Teachers, 4th grade (Panel)	10,415	39%
(4) Teacher evaluation data - Centralized stage		
N. Teachers, 4th grade (Panel)	2,434	
(5) Teacher evaluation data - Decentralized stage		
N. Teachers, 4th grade (Panel)	880	

* The two datasets were merged through a combination of variables: school ID, school annexed, shift, grade, and classroom.

Note: The different symbols allow to simplify the association of the figures presented and their population of interest.

Table 5 presents descriptive statistics of the 192,226 students in the panel for whom we estimate the VAM. The first subsample includes the 71 percent of students in the ECE Panel who had the same teacher (for math or reading) in 3rd and 4th grade of primary and who remained in the same school (Group 1). The second sample (henceforth, "full sample") includes the full sample of students in the ECE panel which, in addition to Group 1, also includes the 22 percent of students that had a different teacher between 3rd and 4th grade and/or changed schools between the two

years (Group 2 and Group 3).[20] The latter case justifies the inclusion of school characteristics in the TVA measure to control for changes in students' test scores influenced by the move to a different school.[21]

Table 5. VA estimation models - Student sample

Students - 4th grade primary (2018)		
	N.	**%**
Dosage=1		
Group 1 - Same teacher in 3rd/4th grade and same school	136,340	70.9
Dosage=.5		
Group 2 - Different teacher in 3rd/4th grade and same school	36,546	19.01
Group 3 - Different teacher in 3rd/4th grade and different school	4,342	2.68
Students for whom teacher ID is missing	14,998	7.80
N.	192,226	100

4.2. Sample Bias

Given the data limitations and the sample restrictions described above, we check for sample selection issues to avoid biased TVA estimates.

First, we compare the characteristics of the teachers included in the SIAGIE with those of the teachers in the more complete Nexus database. Table A3 reports the full exercise. We see that, when compared to the full set of teachers, the SIAGIE overrepresents female teachers and teachers with full-time contracts. Teachers in urban areas and in multi-teacher and larger schools are also overrepresented, suggesting that the rate of reporting in the SIAGIE is largely related to the administrative and technical capacity of the school. To avoid overestimating the correlation

[20] We could not recover the teachers' ID for 8% of the students; they are thus excluded from the estimation.

[21] Table A2 presents a test of the differences between the two samples of students. Overall, the differences are considerably small between the two groups. Interestingly, the fact that the socioeconomic level of the students is not statistically different between the two groups suggests that the decision to change schools (or classrooms) is not related to the student's socioeconomic status.

between TVA and the results of the teacher evaluation, these features must be considered when interpreting our results.

Second, we explore whether the characteristics of the students for whom we are able to recover both 2nd and 4th grade ECE test scores differ from those for whom we have only 2nd grade scores. We find that the students who can be tracked up to 4th grade are generally of higher SES (Table A4).

Third, we study differences between the characteristics of teachers who participated in the teacher evaluation in either 2015 or 2017 and those in the system in 2018 that did not participate in these evaluations.[22] Teachers reporting PUN scores are on average 14 years younger, mainly employed as temporary teachers in 2018, and are more likely to work in smaller and rural schools (Table A5).[23] Finally, we turn to the decentralized stage of the teacher evaluation to compare the characteristics of those candidates that passed the PUN in either 2015 or 2017 and thus progressed to this stage, to those who did not. Compared to the teachers who did not pass the PUN, applicants at the decentralized stage tend to be younger males, who studied at a university rather than a vocational training institute, and who performed better on the curricular and pedagogical knowledge segment of the centralized evaluation (Table A6).

4.3. Descriptive Statistics

Tables 6 and 7 report descriptive statistics of the variables used in the VA estimation models for math and reading, respectively. In the full sample of students, the average 4th grade student scores 504 in math and 502 in reading (corresponding to the "In progress" ECE

[22] Peru implemented its new teacher recruitment for the first time in 2015. Teachers who were already in the system by then were not required to go through the same evaluation process.
[23] This analysis does not include teachers that took the PUN in 2015 or 2017 but were not teaching in 2018.

achievement level[24]), and above 600 points (corresponding to the "Satisfactory" level) in 2nd grade in both subjects. Of the 24 percent of students categorized as high-SES (i.e., those whose mother completed a secondary education), 73 percent performed at the "Expected achievement" level in terms of their 2nd grade scores on math and reading. Meanwhile, 4 percent of students attend school in rural areas and the average number of students per school is around 600 students.

[24] In 2016, 2nd grade ECE achievement levels for math (reading) were: "Initial" <512 (<458); "In progress" 512-638 (458-583); and "Satisfactory" >638 (>583). In 2018, 4th grade ECE achievement levels for math (reading) were: "Less than initial" <352 (<357); "Initial" 352-422 (357-445); "In progress" 422-526 (445-522); and "Satisfactory" ≥ 526 (≥ 522). (MINEDU 2016 and 2018, Office for measuring the quality of learning - *Oficina de Medición de la Calidad de los Aprendizajes*, UMC).

Table 6. Descriptive statistics of VA model – Math

	Same teacher in 3rd and 4th grade					Full sample				
	N.	Mean	Std. Dev.	Min	Max	N.	Mean	Std. Dev.	Min	Max
Dependent variables										
4th grade ECE score - Math	103,055	504	92	11	787	161,940	501	92	11	787
Student's characteristics										
High-SES (student's mother completed secondary school) (%)	103,055	0.24	0.428	0	1	161,940	0.24	0.428	0	1
2nd grade ECE score - Math	103,055	644	116	145	896	161,940	642	116	143	896
2nd grade ECE score - Reading	103,055	606	73	189	826	161,940	605	73	189	826
2nd grade score in the subject (%)										
A - Expected	103,055	0.73	0.442	0	1	161,940	0.74	0.438	0	1
AD - Outstanding	103,055	0.19	0.391	0	1	161,940	0.18	0.384	0	1
B - In progress	103,055	0.06	0.246	0	1	161,940	0.07	0.249	0	1
C - Initial	103,055	0.01	0.117	0	1	161,940	0.01	0.111	0	1
Classroom aggregates										
High-SES students (%)	103,055	0.24	0.189	0	1	161,940	0.24	0.201	0	1
Mean 2nd grade ECE score - Math	103,055	642	71	276	892	161,940	640	73	207	894
School's characteristics										
Rural school (%)						161,940	0.04	0.204	0	1
Total enrollment						161,940	593	356	0	1942

Table 7. Descriptive statistics of VA model – Reading.

	Same teacher in 3rd and 4th grade					Full sample				
	N.	Mean	Std. Dev.	Min	Max	N.	Mean	Std. Dev.	Min	Max
Dependent variables										
4th grade ECE score - Reading	103,090	502	91	-58	869	161,991	499.79	91	-58	869
Student's characteristics										
High-SES (student's mother completed secondary school) (%)	103,090	0.24	0.428	0	1	161,991	0.24	0.428	0	1
2nd grade ECE score - Math	103,090	644	115.7	145	896	161,991	642	116	143	896
2nd grade ECE score - Reading	103,090	606	73	189	826	161,991	605	73	189	826
2nd grade score in the subject (%)										
A - Expected	103,090	0.75	0.43	0	1	161,991	0.76	0.429	0	1
AD - Outstanding	103,090	0.19	0.39	0	1	161,991	0.18	0.382	0	1
B - In process	103,090	0.05	0.22	0	1	161,991	0.05	0.226	0	1
C - Initial	103,090	0.01	0.11	0	1	161,991	0.01	0.106	0	1
Classroom aggregates										
High-SES students (%)	103,090	0.24	0.19	0	1	161,991	0.24	0.201	0	1
Mean 2nd grade ECE score - Reading	103,090	605	44.5	402	792	161,991	604	46	388	826
School's characteristics										
Rural school (%)						161,991	0.04	0.204	0	1
Total enrollment						161,991	593	356	0	1942

Table 8 presents descriptive statistics for the model variables used to estimate equation (4).[25] On average, in 2018, in both samples 75 percent of the teachers included in the estimation are female, while 14 percent (17 percent) have a temporary contract in the restricted (full) sample. More than 60 percent of the teachers are concentrated in the two lowest salary levels.[26]

Table 8. Descriptive statistics of TVA and teacher characteristics model.

	Same teacher in 3rd and 4th grade			**Full sample**		
	N.	Mean	Std. Dev.	N.	Mean	Std. Dev.
Female	8,654	0.75	0.43	12,132	0.76	0.43
Age	8,654	49.92	8.53	12,132	49.56	8.92
Temporary contract	8,654	0.14	0.34	12,132	0.17	0.38
1st salary level (lowest)	8,654	0.45	0.50	12,132	0.49	0.50
2nd salary level	8,654	0.20	0.40	12,132	0.19	0.39
3rd salary level	8,654	0.23	0.42	12,132	0.20	0.40
4th salary level	8,654	0.09	0.29	12,132	0.08	0.27
5th salary level	8,654	0.03	0.18	12,132	0.03	0.17
6th salary level	8,654	0.01	0.10	12,132	0.01	0.09

5. MAIN RESULTS

5.1. Teacher Value-Added Model (TVA)

Tables 9 and 10 show the results of the TVA estimation for math and reading, respectively. Columns 1-4 present results for the subsample of students who had the same teacher (in math or reading) in 3rd and 4th grade of primary and who remained in the same school. As previously discussed, in this scenario, we are sure that we are attributing the test score progress of students between 2nd and 4th grade to the same teacher. Columns 5-8 present results for the full sample estimation, weighted by teacher dosage. Columns 1 and 5 show results for the basic VAM; Columns 2 and 6 include the ECE score in the other subject (e.g., in the estimation of value-added in math, the score of students on the second grade ECE reading test); Columns 3 and 7 include

[25] Descriptive statistics for the restricted sample of teachers for whom we could recover information on years of experience are included in Table A7.

[26] Although in Peru the teacher salary scale is divided into 8 levels, less than 2% of the country's teachers are found in the two highest levels (7th and 8th), and none in our sample.

non-linearities in students' ECE scores; and Columns 4 and 8 present results from the first stage of the two-step VAM estimations (Equation 2) in which we control for students' characteristics at the classroom level and, in the full sample of students, for school-level controls.

Table 9 shows the TVA estimation regression results for math. Students' SES is positively correlated with the 4th grade ECE math score in all specifications, an effect ranging between 6 and 8 points.[27] Students scoring in the highest category in this subject on their 2nd grade evaluation ("Expected" being the reference category) then score between 44 and 52 points higher on their 4th grade ECE, while students in the lowest category score between 46 and 60 points lower. The most important control in this model is the student lagged ECE scores. When we include only the lagged score in the same subject, we find that the 4th grade math score increases by around 0.32 points for each additional point in the 2nd grade score (columns 1 and 5). When the 2nd grade reading score is included (columns 2-4 and columns 6-8) the effect on math is reduced to 0.25 points, and the 2nd grade reading score increases the 4th grade math score by about 0.21 points. This suggests that reading skills positively influence math skills. Additionally, when non-linearities are included, we find that the impact of 2nd grade scores is positive and decreasing; the positive interaction effect implies that a student with a high score in both subjects has an additional positive effect on the 4th grade score. The findings are robust to the two-step VAM estimation (columns 4 and 8) and show positive correlations of classroom characteristics with students' ECE score. Lastly, we include controls at the school level (column 8): attending a rural school has a negative effect on 4th grade scores while the effect of school size does not seem robust to the different specifications.

[27] These correspond to an effect of 0.06 and 0.08 standard deviations. As shown in Tables 6 and 7 the standard deviation for 4th grade score in math and reading is 92 and 91, respectively.

Table 10 shows the same set of results for reading. While the findings are slightly different in terms of magnitude, the interpretation is similar as that for the math results. At the same time, we observe a smaller impact of 2nd grade ECE math scores on 4th grade ECE reading scores (0.13 points) compared to the opposite relation (0.21 points).

In order to check the stability of our TVA measures, in Table 11 we present correlations between the results of the models estimated in Tables 9 and 10. Table 11 shows that the TVA measures we estimate are highly correlated across all of the different specifications, which suggests that the different models have a low significant impact on the ranking of teachers according to the estimated VA measure.

Table 9. TVA estimation regression results – Math.

	4th grade ECE score - Math							
	Same teacher in 3rd and 4th grade				*Full sample*			
	One-step	One-step	One-step	Two-step	One-step	One-step	One-step	Two-step
	Math	Math	Math	Math	Math	Math	Math	Math
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student's characteristics								
High-SES (mother completed secondary school)	7.6090***	5.9703***	6.0759***	6.2284***	7.5817***	5.8715***	5.9562***	6.1581***
	(0.4898)	(0.4839)	(0.4824)	(0.5673)	(0.4427)	(0.4369)	(0.4352)	(0.5034)
2nd grade note in subject (Ref. Expected (A))								
Outstanding (AD)	52.5414***	47.6206***	48.1721***	43.9874***	52.4301***	47.5352***	48.0707***	43.7482***
	(0.5713)	(0.5710)	(0.5719)	(0.6029)	(0.5128)	(0.5122)	(0.5125)	(0.5305)
In progress (B)	-44.0432***	-40.7437***	-39.4841***	-28.4405***	-43.4390***	-39.9876***	-38.8195***	-28.1467***
	(0.8450)	(0.8354)	(0.8344)	(0.9012)	(0.6834)	(0.6843)	(0.6864)	(0.7358)
Initial (C)	-59.2642***	-54.9433***	-52.9304***	-46.3653***	-59.7163***	-55.3122***	-53.4680***	-46.5053***
	(1.6949)	(1.6729)	(1.6704)	(1.8594)	(1.5471)	(1.5485)	(1.5657)	(1.8059)
2nd grade ECE score - Math	0.3253***	0.2554***	0.4362***	0.4108***	0.3286***	0.2573***	0.4222***	0.4111***
	(0.0023)	(0.0026)	(0.0214)	(0.0232)	(0.0021)	(0.0024)	(0.0201)	(0.0216)
2nd grade ECE score - Reading		0.2125***	0.4862***	0.7377***		0.2147***	0.4780***	0.6693***
		(0.0041)	(0.0367)	(0.0401)		(0.0037)	(0.0351)	(0.0378)
2nd grade ECE score - Math^2			-0.0004***	-0.0004***			-0.0004***	-0.0004***
			(0.0000)	(0.0000)			(0.0000)	(0.0000)
2nd grade ECE score - Reading^2			-0.0005***	-0.0007***			-0.0005***	-0.0007***
			(0.0000)	(0.0000)			(0.0000)	(0.0000)
2nd grade ECE score - Math*Reading			0.0006***	0.0006***			0.0006***	0.0006***
			(0.0001)	(0.0001)			(0.0000)	(0.0001)
Classroom aggregates								
High-SES students (%)				44.5125***				39.6162***
				(1.3347)				(1.1297)
Mean 2nd grade ECE score - Math				0.0403***				0.0260***
				(0.0042)				(0.0037)
School's characteristics								
Rural school					-29.8386**	-23.3627*	-20.5935*	-7.3486***
					(12.2042)	(12.0609)	(12.0627)	(1.0172)
Total enrollment					-0.0096	-0.0124	-0.0131	0.0070***
					(0.0082)	(0.0081)	(0.0081)	(0.0006)
Constant	286.4287***	203.7029***	62.9515***	-52.5349***	290.1007***	208.4345***	76.4767***	-26.5384***
	(1.4600)	(2.1408)	(10.1859)	(10.3879)	(5.1962)	(5.2782)	(10.9303)	(9.7216)
Teacher FE	Yes	Yes	Yes	No	Yes	Yes	Yes	No
N.	103057	103055	103055	103055	161946	161940	161940	161940
N_g (number of teachers)	9101	9101	9101	9101	13049	13049	13049	13049
R2	0.6192	0.6299	0.6324	0.4266	0.6080	0.6191	0.6215	0.4237

***p<0.01; **p<0.05; *p<0.1. Standard errors in the full sample estimations are clustered at the student level.

Table 10. TVA estimation regression results – Reading.

	4th grade ECE score - Reading							
	Same teacher in 3rd and 4th grade				*Full sample*			
	One-step	One-step	One-step	Two-step	One-step	One-step	One-step	Two-step
	Reading	Reading	Reading	Reading	Reading	Reading	Reading	Reading
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student's characteristics								
High-SES (mother completed secondary school)	9.5471***	9.8334***	9.9321***	10.4317***	9.5032***	9.8022***	9.8749***	10.3200***
	(0.5028)	(0.4966)	(0.4957)	(0.5529)	(0.4564)	(0.4506)	(0.4494)	(0.4958)
2nd grade note in subject (Ref. Expected (A))								
Outstanding (AD)	50.2850***	44.9416***	45.0767***	40.3967***	50.5491***	45.2067***	45.3584***	40.6372***
	(0.5861)	(0.5891)	(0.5903)	(0.5869)	(0.5226)	(0.5242)	(0.5247)	(0.5156)
In progress (B)	-40.9282***	-37.3291***	-36.5000***	-27.1911***	-40.4575***	-37.0135***	-36.1605***	-27.2344***
	(0.9484)	(0.9397)	(0.9396)	(0.9700)	(0.7780)	(0.7792)	(0.7819)	(0.8082)
Initial (C)	-53.8660***	-49.9094***	-48.3524***	-44.3371***	-53.7870***	-49.8885***	-48.3018***	-43.9577***
	(1.7876)	(1.7678)	(1.7680)	(1.8703)	(1.5932)	(1.6064)	(1.6196)	(1.7883)
2nd grade ECE score - Reading	0.5675***	0.4618***	1.0887***	1.2891***	0.5726***	0.4650***	1.0956***	1.2540***
	(0.0036)	(0.0042)	(0.0378)	(0.0392)	(0.0036)	(0.0040)	(0.0379)	(0.0392)
2nd grade ECE score - Math		0.1296***	0.0040	-0.0265		0.1307***	0.0009	-0.0279
		(0.0027)	(0.0220)	(0.0225)		(0.0024)	(0.0209)	(0.0215)
2nd grade ECE score - Reading^2			-0.0002***	-0.0001***			-0.0002***	-0.0002***
			(0.0000)	(0.0000)			(0.0000)	(0.0000)
2nd grade ECE score - Math^2			-0.0008***	-0.0009***			-0.0008***	-0.0009***
			(0.0000)	(0.0000)			(0.0000)	(0.0000)
2nd grade ECE score - Reading*Math			0.0005***	0.0005***			0.0006***	0.0006***
			(0.0001)	(0.0001)			(0.0001)	(0.0001)
Classroom aggregates								
High-SES students (%)				43.5580***				39.1430***
				(1.3548)				(1.1582)
Mean 2nd grade ECE score - Reading				0.0201***				-0.0083
				(0.0068)				(0.0061)
School's characteristics								
Rural school					-33.3324***	-35.6391***	-33.3684***	-9.7914***
					(9.6475)	(9.9101)	(9.9209)	(0.9813)
Total enrollment					-0.0134*	-0.0164**	-0.0174**	0.0069***
					(0.0079)	(0.0079)	(0.0079)	(0.0006)
Constant	149.4102***	130.6856***	-21.6945**	-100.5712***	154.9094***	138.5574***	-12.8858	-76.6044***
	(2.1856)	(2.1926)	(10.4656)	(10.2521)	(5.2833)	(5.2476)	(11.6147)	(10.1944)
Teacher FE	Yes	Yes	Yes	No	Yes	Yes	Yes	No
N.	103103	103090	103090	103090	162018	161991	161991	161991
N_g (number of teachers)	9106	9106	9106	9106	13047	13047	13047	13047
R2	0.5909	0.6010	0.6025	0.4426	0.5814	0.5918	0.5934	0.4399

***p<0.01; **p<0.05; *p<0.1. Standard errors in the full sample estimations are clustered at the student level.

Table 11. Correlation between TVA specifications.

	Same teacher in 3rd and 4th grade			*Full sample*		
	(1)	(2)	(3)	(4)	(5)	(6)
Math	0.991	0.988	0.939	0.992	0.989	0.938
Reading	0.989	0.986	0.953	0.989	0.987	0.945

Note: The Table shows correlation coefficients between the basic VAM specification (column 1 of Tables 9 and 10 for the restricted sample, and column 5 of Tables 9 and 10 for the full sample) against each of the following specifications.

5.1.1. Teacher Characteristics and TVA

Table 12 shows the relationship between mean TVA and teacher characteristics.[28] Differently from columns 1 and 3, columns 2 and 4 include an indicator variable for whether the instructor teaching in 2018 had three or more years of experience. Including this variable reduces our sample by around 80 percent given that information on years of experience was only collected for those teachers who participated in the teacher hiring process in either 2015 or 2017.

We find that teacher gender, type of contract, and salary scale are significantly correlated with the estimated mean TVA. Specifically, the outcomes of students of female teachers are between 0.05-0.1 SD higher than those taught by male teachers. Meanwhile, TVA appears to be 0.04-0.06 SD lower for teachers with a temporary contract than those with a permanent position. Temporary teachers and those who begin their careers through the teacher hiring process receive a salary corresponding to the 1st (lowest) level. We observe a positive, although not robust, correlation between TVA and salary scale, with higher TVA for teachers at higher salary levels. Although information on years of experience in the public or private sector is available

[28] We report results from the two-step estimation of equation (4). One-step estimation results and results from specifications where school FE were not included are similar. Results are available upon request.

for only a restricted group of teachers, we find that having more than three years of experience (in the public or private sector) does not appear to be correlated with our measure of TVA.

Table 12. Relationship between mean TVA and teacher characteristics.

	Mean TVA			
	Same teacher in 3rd and 4th grade		*Full sample*	
	(1)	(2)	(3)	(4)
Female	0.0984***	0.1220	0.0538***	0.0355
	(0.0173)	(0.0961)	(0.0114)	(0.0390)
Age	0.1835	0.5060	0.1521	0.3585
	(0.1753)	(1.1300)	(0.1046)	(0.3891)
Age2	-0.3042*	-0.5280	-0.2195**	-0.3725
	(0.1694)	(1.3027)	(0.1029)	(0.4506)
Temporary contract	-0.0627**	-0.0747	-0.0444***	-0.0536
	(0.0246)	(0.0836)	(0.0141)	(0.0328)
Reference: 1st salary level (lowest)				
2nd salary level	0.0107	-0.2439*	0.0069	0.1358***
	(0.0185)	(0.1330)	(0.0129)	(0.0461)
3rd salary level	0.0412**	0.6797***	0.0191	0.4820***
	(0.0174)	(0.0930)	(0.0119)	(0.0350)
4th salary level	0.0401**	0.1240	0.0216**	-0.3528
	(0.0158)	(0.0792)	(0.0109)	(0.2185)
5th salary level	0.0388***		0.0292***	
	(0.0137)		(0.0101)	
6th salary level	0.0240*		0.0176*	
	(0.0136)		(0.0090)	
Had > 3 years		0.0585		-0.0062
of experience in 2018		(0.0993)		(0.0350)
Constant	0.9308***	0.3319	0.9135***	0.4736***
	(0.0201)	(0.3122)	(0.0134)	(0.1298)
School FE	Yes	Yes	Yes	Yes
N.	8654	1674	12132	2835
Clusters	3936	1281	4320	1762
R2	0.6678	0.8662	0.6584	0.8386

***p<0.01; **p<0.05; *p<0.1. Standard errors are clustered at the school level.

5.2. Robustness Checks

As explained in detail in Section 4, the quality of the student-teacher linkage data tends to be poorer for small and rural schools. To validate our VAM results, we consequently perform two robustness checks. First, we run equation (1) over the

sample of instructors who teach a class with at least five students, or the minimum number of students per classroom for the ECE evaluation to be implemented. This reduces our sample of teachers by about 4 percent (400 teachers) in the restricted sample and by about 8 percent (1,000 teachers) in the full sample. Second, we run equation (1) over the sample of urban schools. This reduces our sample of schools by about 15 percent (600 schools) in the restricted sample and by about 16 percent (700 schools) in the full sample. Tables A8 and A9 present the results of this analysis, showing that findings are robust to these sample restrictions.

5.3. Correlations Between TVA and Teacher Evaluation Instruments

Tables 13 and 14 show the correlation coefficients between estimated TVA and the different teacher evaluation instruments for math and reading, respectively. The VAM specifications over which we run these correlations are models (3) and (4) of Table 9 (math) and Table 10 (reading) for the restricted sample of students and models (7) and (8) of the same tables for the full sample of students. We present correlation results for non-adjusted and selection-adjusted coefficients, where the latter were estimated as explained in Section 3.2.1.

Overall, among the centralized stage instruments, the curriculum knowledge sub-test presents the highest correlation with our measure of TVA (0.09-0.28). The reading comprehension sub-test shows the lowest correlation with TVA (0.04-0.22) in all specifications and for both subjects—not surprising given that this instrument is simply testing whether the teacher can read and comprehend a text in her own language. Moreover, the weighted average of the three PUN sub-tests has a higher predictive power than the single subtests, suggesting that the weighted combination of

the different instruments possibly increases the ability to predict the added value of a teacher. For math specifically, the correlation between TVA and the total PUN score varies between 0.17 and 0.34, depending on the specification and the sample used. This magnitude is similar to that found in Chile by Taut et. al. (2014), where the authors find significant correlations between 0.18 and 0.20 in math for the written instrument of the teacher evaluation.

When run over the full sample of students, the correlation coefficients appear to be lower. This could be due to the inclusion of a higher proportion of temporary teachers or schools with lower teacher retention rates.

Once we correct for the bias due to non-random selection into teaching, correlation coefficients between TVA and the aggregate PUN score increases slightly. This may be due to inclusion of applicants who received lower PUN scores on the three instruments (Figure 2) and non-observable characteristics making the pool of candidates more heterogeneous. Among the three instruments, the reading comprehension coefficients drop significantly once we correct for selection into teaching, suggesting that the predictive power of this instrument is specific to a more homogeneous group of candidates. The logical reasoning coefficients also decrease when correcting for selection bias but by a much smaller amount. On the contrary, correlation coefficients of the curriculum knowledge instrument increase slightly when selection bias is taken into account, and is likely what drives the effect of the aggregate PUN score.

Moreover, we find no significant correlation between the decentralize stage instruments and our measures of TVA for math, as well as non-robust correlations for

the professional experience and classroom observation instruments. We do find a positive and significant correlation between the classroom observation component and TVA for reading (around 0.08 across the different specifications). For the US, Kane and Staiger (2012) report that the correlation between math teacher value-added and score on the classroom observation (measured across different classrooms) ranges from 0.16 to 0.26, depending on the observation rubric. However, they find higher correlations with other instruments, such as the teacher's portfolio (0.24-0.31) and video-taped lessons (0.20-0.24). Although not directly comparable to the decentralized stage in Peru, other studies for the US analyze a decentralized evaluation instrument in the form of school principal assessments. For example, Jacob and Lefgren (2008) report a correlation of 0.32 between estimates of teacher value-added in math and ratings based on principals' beliefs about the ability of teachers to raise math achievement. The analogous correlation for reading is 0.29 (correlations are reported after adjusting for estimation error in the value-added measures). Meanwhile, Harris and Sass (2014) report slightly larger correlations for the same principal assessment—0.41 for math and 0.44 for reading—as well as correlate value-added with "overall" principal ratings, documenting correlations in math and reading of 0.34 and 0.38, respectively.

It is important to stress that in the case of Peru the decentralized stage evaluations are only taken up by the teachers who passed the PUN.[29] This means that the sample of teachers at this stage is much smaller and more homogeneous, possibly affecting

[29] The passing rate of the PUN has been consistently low in all of the teacher selection years: 13% in 2015, 11% in 2017, 12% in 2018, and 7% in 2019.

predictive power and comparability with other studies based on hiring systems where all the instruments are applied to the full set of applicants (e.g. Bertoni et al., 2020). Indeed, we observe that for the group of teachers that pass to the second stage, the PUN is no longer predictive. This suggests that the homogeneity of the pool of candidates can affect the predictive power of the evaluation instruments at both the centralized and decentralized stages. Additionally, though its implementation should follow ministerial guidelines, the decentralized stage of the teacher recruitment in Peru is a more arbitrary process than the centralized stage. The dynamics of the classroom observation leave, for example, room for subjective assessment. Moreover, schools have the freedom to tailor the content of the interview to their specific needs, making it more difficult to preserve an entirely objective outcome of the evaluation.

This result justifies the fact that, by construction, the PUN weighs 67 percent of the total score. It also points to a need to reflect on the scope of the decentralized stage and, in the short run, to strengthen the monitoring of its implementation.

Table 13. Correlation coefficients of TVA measures and teacher evaluation instruments – Math

		Teacher Value Added (TVA)							
		Same teacher in 3rd and 4th grade				*Full sample*			
	Covariate Mean	One-step VAM		Two-step VAM		One-step VAM		Two-step VAM	
		Non-adjusted	Selection-adjusted	Non-adjusted	Selection-adjusted	Non-adjusted	Selection-adjusted	Non-adjusted	Selection-adjusted
		Math	Math	Math	Math	Math	Math	Math	Math
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teachers in Centralized stage (PUN)									
Total PUN score	125.0	0.3014***	0.3406***	0.2703***	0.2812***	0.2067***	0.2249***	0.1754***	0.1688***
Reading comprehension	36.3	0.2236***	0.1029***	0.1984***	0.0820***	0.1491***	0.0583***	0.1229***	0.0372*
Logical reasoning	29.0	0.2489***	0.1545***	0.2290***	0.1423***	0.1851***	0.1247***	0.1650***	0.1132***
Curriculum knowledge	59.6	0.2785***	0.2421***	0.2469***	0.1827***	0.1890***	0.1445***	0.1572***	0.0897***
N.	2031	2031	2020	2031	2020	3550	3526	3550	3526
Teachers in Decentralized stage									
Total PUN score	149.2	0.1029	0.1406	0.0436	0.0641	0.0229	0.0825	-0.0332	0.0068
Reading comprehension	41.4	0.0509	0.0661	0.0246	0.0413	0.0146	0.0334	-0.0137	0.0041
Logical reasoning	37.0	0.0379	0.0250	0.0185	0.0033	0.0284	0.0434	0.0000	0.0132
Curriculum knowledge	70.8	0.0831	0.1265	0.0317	0.0570	0.0034	0.0502	-0.0378	-0.0055
N.	750	750	750	750	750	1136	1136	1136	1136
Total decentralized stage	71.1	0.0372	0.0358	0.0549	0.0536	0.0493*	0.0525*	0.0542*	0.0576**
Classroom observation	42.8	0.0269	0.0246	0.0322	0.0298	0.0521*	0.0533*	0.0450	0.0463
Interview	21.1	0.0310	0.0269	0.0330	0.0285	0.0247	0.0241	0.0217	0.0210
Professional experience	7.2	0.0257	0.0292	0.0596*	0.0651*	0.0240	0.0301	0.0514*	0.0583**
N.	750	750	750	750	750	1136	1136	1136	1136

***p<0.01; **p<0.05; *p<0.1.

Table 14. Correlation coefficients of TVA measures and teacher evaluation instruments – Reading

		Teacher Value Added (TVA)							
		Same teacher in 3rd and 4th grade				*Full sample*			
	Covariate Mean	One-step		Two-step		One-step		Two-step	
		Non-adjusted	Selection-adjusted	Non-adjusted	Selection-adjusted	Non-adjusted	Selection-adjusted	Non-adjusted	Selection-adjusted
		Reading (1)	Reading (2)	Reading (3)	Reading (4)	Reading (5)	Reading (6)	Reading (7)	Reading (8)
Teachers in Centralized stage (PUN)									
Total PUN score	124.9	0.2298***	0.2878***	0.2074***	0.2431***	0.1374***	0.1905***	0.1168***	0.1529***
Reading comprehension	36.3	0.1601***	0.0725**	0.1404***	0.0549*	0.0934***	0.0421*	0.0746***	0.0257
Logical reasoning	29.0	0.1848***	0.1130***	0.1716***	0.1054***	0.1252***	0.0973***	0.1128***	0.0906***
Curriculum knowledge	59.6	0.2212***	0.2469***	0.1986***	0.2023***	0.1274***	0.1435***	0.1070***	0.1086***
N.	2032	2032	2021	2032	2021	3551	3527	3551	3527
Teachers in Decentralized stage									
Total PUN score	149.2	0.1671**	0.1868*	0.1161	0.1148	0.0235	0.0663	-0.0203	0.0039
Reading comprehension	41.4	0.0280	0.0281	-0.0113	-0.0116	-0.0252	-0.0151	-0.0489	-0.0404
Logical reasoning	37.0	0.0717	0.0556	0.0596	0.0397	0.0438	0.0562	0.0174	0.0265
Curriculum knowledge	70.7	0.1566**	0.1905**	0.1183*	0.1334	0.0149	0.0509	-0.0135	0.0101
N.	750	750	750	750	750	1135	1135	1135	1135
Total decentralized stage	71.1	0.0776**	0.0749**	0.0934**	0.0905**	0.0658**	0.0686**	0.0711**	0.0737**
Classroom observation	42.8	0.0735*	0.0720*	0.0813**	0.0796**	0.0810***	0.0821***	0.0769***	0.0778***
Interview	21.1	0.0701*	0.0670*	0.0692*	0.0655*	0.0514*	0.0515*	0.0474	0.0471
Professional experience	7.2	0.0249	0.0222	0.0534	0.0517	-0.0043	-0.0008	0.0202	0.0242
N.	750	750	750	750	750	1135	1135	1135	1135

***p<0.01; **p<0.05; *p<0.1.

6. DISCUSSION

In this article, we estimate value-added measures for public primary school teachers in Peru and test for their correlation with the results of two national teacher selections (2015 and 2017). In doing so, we assess whether the instruments used to assign vacancies in the teacher hiring process are effective in identifying the most competent teachers.

Our findings indicate that among the three sub-tests of the first, centralized stage of the teacher recruitment, the curricular and pedagogical knowledge component has the highest (and significant) correlation with the TVA measure, while the lowest is found for the reading comprehension component. This first result suggests that assigning a higher weight to the curricular and pedagogical knowledge component is a good strategy to identify the most effective teachers. Second, we find that the aggregate PUN score has a higher correlation with our TVA measure than do the individual sub-tests, implying that the current design (a weighted combination of the three different instruments where the weights are assigned by the MINEDU) possibly increases the ability to predict the added value of a teacher. Additionally, when we correct for the bias due to non-random selection into teaching, the correlation coefficients of the aggregated centralized stage and the curriculum knowledge instrument increase, while those of the reading comprehension and the logical reasoning instruments decrease.

Among the decentralized, second stage instruments, we find no significant correlation with our measures of TVA for math, as well as non-robust correlations for the professional experience and classroom observation instruments. The lack of

correlation between the decentralized stage and TVA may be driven by several factors. First, the homogeneity of the group of applicants who pass on to the decentralized stage can affect the predictive power of these instruments. Second, though regulated by ministerial guidelines, the implementation of the decentralized stage is also shaped by local aspects, such as the freedom schools have to tailor the content of the interview to the specificities of their institution, or the availability of personnel in forming the Evaluation Committee on the day of the test, among others. Interestingly, we do find a low positive and significant correlation between the classroom observation component and TVA for reading. Further analysis of the implementation of the classroom observation instrument at the school level would help to disentangle the variation in the correlation results by subject.

We also find that our measure of TVA is higher for female teachers and for instructors at higher salary levels. At the same time, TVA measures are lower for teachers with temporary contracts compared to those with permanent positions. We find no correlation between our measure of TVA and teacher experience (when greater than 3 years).

Future research could test the stability of these TVA estimates by, for example, estimating TVA models for an additional cohort of students. This analysis could also be further developed by studying the impact of TVA on student achievement and its effectiveness in bridging learning gaps. These extensions would require more systematic student and teacher classroom data. Additionally, more information on the quality of the implementation of the decentralized stage of the teacher hiring process

would improve the set of inputs necessary to evaluate its correlation with teacher effectiveness.

Certainly, defining the optimal weights of the different evaluation components ultimately depends on the MINEDU's policy objectives. Over the last decade, Peru has made significant efforts to reform its public-school teacher selection, promoting a more meritocratic and effective process, which has helped to recover the prestige of the profession and improve the quality of teaching (Elacqua et al., 2018). Our results suggest that adjusting the scope and monitoring the implementation of the evaluation instruments employed in the teacher selection process would not only help to successfully identify the most effective teachers, but would represent a significant step in increasing equitable educational opportunities in the country.

REFERENCES

Aaronson, D., Barrow, L. and Sander, W. (2007), Teachers and student achievement in the Chicago public high schools. Journal of Labor Economics, 25, 95-135.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., and Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. The Quarterly Journal of Economics, 131, 1415-1453.

Bertoni, E., Elacqua, G., Méndez, C., Montalva, V., Munevar, I., Olsen, S. and Román, A. (2020). Seleccionar y asignar docentes en America Latina y el Caribe. Un camino para la calidad y equidad en Educación. Inter-American Development Bank Technical Note N.01900.

Bertoni, E., Elacqua, G., Hincapié, D., Méndez, C., and Paredes, D. (2019). Teachers' preferences for proximity and the implications for staffing schools: Evidence from Peru. IDB Working Paper Series N.01073.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., and Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. Washington, DC: Economic Policy Institute.

Bau, N. and Das, J. (2020). Teacher value added in a low-income country. American Economic Journal: Economic Policy, 12, 62-96.

Betts, J. R., Zau, A., and Rice, L. (2003). Determinants of student achievement: New evidence from San Diego, San Francisco: Public Policy Institute of California, 1-5821.

Bietenbeck, J., Piopiunik, M. and Wiederhold, S. (2018). Africa's skill tragedy: Does teachers' lack of knowledge lead to low student performance? Journal of Human Resources, University of Wisconsin Press, 53, 553-578.

Boyd, D., Goldhaber, D., Lankford, H., and Wyckoff, J. (2007). The effect of certification and preparation on teacher quality. Future Child, 17, 45-68.

Bruno, P., and Strunk, K. O. (2019). Making the cut: The effectiveness of teacher screening and hiring in the Los Angeles Unified School District. Educational Evaluation and Policy Analysis, 41, 426-460.

Brutti, Z. and Sánchez Torres, F. (2017). Does better teacher selection lead to better students? evidence from a large-scale reform in Colombia. Documento CEDE N. 2017-11.

Chetty, R., Friedman, J. N., and Rockoff., J. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. American Economic Review, 104, 2593-2632.

Clotfelter, C., Ladd, H. and Vigdor, J. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. The Journal of Human Resources, 45, 655–81.

Clotfelter, C., Ladd, H. and Vigdor, J. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. Economics of Education Review, 26, 673–82.

Clotfelter, C., Ladd, H. and Vigdor, J. (2006). Teacher-student matching and the assessment of teacher effectiveness. Journal of Human Resources, 41, 778–820.

Cruz-Aguayo, Y., Hincapié, D., and Rodríguez., C. (2020). Profesores a Prueba: Claves para una evaluación docente exitosa. Washington DC: Inter-American Development Bank.

Cruz-Aguayo, Y., Ibarrarán, P., and Schady, N. (2017). Do tests applied to teachers predict their effectiveness? Economics Letters, 159, 108-111.

Ehlert, M., Koedel, C., Parsons, E., and Podgursky, M. (2016). Selecting growth measures for use in school evaluation systems: should proportionality matter? Educational Policy, 30, 465–500.

Ehlert, M.; Koedel, C., Parsons, E. and Podgursky, M. (2014). The sensitivity of value-added estimates to specification adjustments: Evidence from school- and teacher-level models in Missouri. Statistics and Public Policy, 1, 19-27.

Goldhaber, D., Grout, C., Huntington-Klein, N. (2017). Screen twice, cut once: Assessing the predictive validity of applicant selection tools. Education Finance and Policy, 12, 197–223.

Goldhaber, D. D., and Brewer, D. J. (1999). Teacher licensing and student achievement. In M. Kanstoroom and C. E. Finn, Jr. (Eds.), Better teachers, better schools (pp. 83–102). Washington, DC: The Thomas B. Fordham Foundation.

Guarino, C., Santibañez, L., Daley, G., and Brewer, D. (2004). A review of research literature on teacher recruitment and retention. TR-164-EDU. Santa Monica, CA: RAND.

Hanushek, E. A., Piopiunik, M. and Wiederhold, S. (2017). The Value of Smarter Teachers: International Evidence on Teacher Cognitive Skills and Student Performance. NBER Working Paper 20727. Cambridge, MA: National Bureau of Economic Research.

Hanushek, E. A., and Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. Annual Review of Economics, 4, 131-157.

Harris, D. N., and Sass, T. R. (2014). Skills, productivity, and the evaluation of teacher performance. Economics of Education Review, 40, 183-204.

Harris, D.N. and Sass, T.R. (2011). Teacher training, teacher quality and student achievement. Journal of Public Economics, 95, 798-812.

Hock, H., and Isenberg, E. (2017) Methods for accounting for co-teaching in value-added models. Statistics and Public Policy, 4, 1-11.

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., and Rosen, R. (2018). Teacher applicant hiring and teacher performance: Evidence from DC public schools. Journal of Public Economics, 166, 81–97.

Jacob, B. A. and Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluations in education. Journal of Labor Economics 26, 101-136.

Kalogrides, D., and Loeb, S. (2013). Different teachers, different peers: The magnitude of student sorting within schools. Educational Researcher, 42, 304–316.

Kane, T. J., and Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper. MET Project. Bill and Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., and Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. Journal of human Resources, 46, 587-613.

Kane, T., and Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Working paper N.14607. Washington, DC: National Bureau of Economic Research.

Koedel, C., Mihaly, K., and Rockoff, J. E. (2015). Value-added modeling: A review. Economics of Education Review, 47, 180–195.

Leigh, A. (2010). Estimating teacher effectiveness from two-year changes in students' test scores. Economics of Education Review, 29, 480-488.

Lockwood, J. R., and McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA Models for estimating treatment effects. Journal of Educational and Behavioral Statistics, 39, 22–52.

Herrmann, M., Walsh, E., and Isenberg, E. (2016). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. Statistics and Public Policy, 3, 1-10.

McCaffrey, D. F., Lockwood, J. R., Mihaly, K., and Sass, T. R. (2012). A review of Stata routines for fixed effects estimation in normal linear models. The Stata Journal, 12, 1–27.

Metzler, J. and Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. Journal of Development Economics, 99, 486–496.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. Peabody Journal of Education, 79, 33-53.

Morris, C. N. (1983). Parametric empirical bayes inference: Theory and applications. Journal of the American Statistical Association, 78, 47–55.

OECD (2014). Indicator D6: What does it take to become a teacher? in Education at a Glance 2014: OECD Indicators, OECD Publishing.

Rivkin, S. G., Hanushek, E. A. and Kain, J. (2005), teachers, schools, and academic achievement. Econometrica, 73, 417-458.

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

Slater, H., Davies, N.M. and Burgess, S. (2012). Do teachers matter? Measuring the variation in teacher effectiveness in England. Oxford Bulletin of Economics and Statistics, 74, 629-645.

Taut, S., Valencia, E., Palacios, D., Santelices, M. V., Jiménez, D., and Manzi, J. (2016). Teacher performance and student learning: linking evidence from two national assessment programmes. Assessment in Education: Principles, Policy and Practice, 23, 53–74.

Treviño, E., Valenzuela, J. P., Villalobos, C., Béjares, C., Wyman, I., and Allende, C. (2018). Agrupamiento por habilidad académica en el sistema escolar. Nueva evidencia para comprender las desigualdades del sistema educativo chileno. Revista mexicana de investigación educativa, 23, 45-71.

Tyler, J., Taylor, E., Kane, T., and Wooten, A. (2010). Using student performance data to identify effective classroom practices. The American Economic Review, 100, 256-260.

APPENDIX TABLES

Table A1. Logistical model results

	P(*employed*)	
	2016	2018
Teacher characteristics		
Female	-0.0799***	-0.1301***
	(0.0038)	(0.0036)
Age	-0.0009***	0.0006**
	(0.0003)	(0.0002)
Has some experience (public/private)	0.2053***	0.2060***
	(0.0058)	(0.0070)
Reference: Studied only in a University		
Studied at an institute and at a university	-0.0367***	-0.0209**
	(0.0133)	(0.0098)
Studied only in an institute	0.0009	0.0011
	(0.0038)	(0.0036)
Centralized stage results		
Reading comprehension score	-0.0014***	0.0038***
	(0.0004)	(0.0003)
Logical reasoning score	-0.0016***	-0.0015***
	(0.0004)	(0.0003)
Curriculum knowledge score	0.0039***	0.0044***
	(0.0002)	(0.0002)
Passed the threshold (0/1)		
Reading comprehension	0.0025	-0.0230***
	(0.0062)	(0.0058)
Logical reasoning	0.0306***	0.1188***
	(0.0062)	(0.0060)
Curriculum knowledge	0.2381***	0.1946***
	(0.0073)	(0.0059)
Constant	0.1841***	-0.0699***
	(0.0154)	(0.0153)
N.	77594	78758
R2	0.0801	0.1605

***p<0.01; **p<0.05; *p<0.1.

Table A2. Students by teacher dosage

	All students	Dosage=1 (A)	Dosage=.5 (B)	p-value A=B	N.
Student characteristics					
Low-SES (student's mother did not complete secondary school)	0.76	0.76	0.76	0.953	145105
Main language (Castilian Spanish)	0.82	0.81	0.83	0.000	146000
Household size	5.29	5.29	5.31	0.401	64189
School characteristics					
Most rural (Rural 1)	0.02	0.01	0.01	0.00	176997
Moderate rural (Rural 2)	0.05	0.04	0.05	0.00	176997
Least rural (Rural 3)	0.08	0.08	0.08	0.29	176997
Urban	0.86	0.87	0.86	0.00	176997
Single-teacher	0.00	0.00	0.00	0.01	177221
Multigrade	0.02	0.02	0.02	0.97	177221
Multi-teacher	0.98	0.98	0.98	0.74	177221
Bilingual	0.09	0.10	0.09	0.00	32984
Vraem	0.07	0.07	0.09	0.00	32984
Frontier	0.26	0.28	0.24	0.00	32984
Total enrollment	559.35	576.98	566.38	0.00	177221
Basic services	0.94	0.95	0.94	0.00	177228
N.	192,226	136,340	40,888		

Note: The universe of this analysis comprises students in the SIAGIE student panel database.

Table A3. Teachers in SIAGIE vs Nexus database.

	All teachers	Nexus only (A)	Nexus and SIAGIE (B)	p-value A=B	N.
Teacher characteristics					
Female	0.62	0.46	0.72	0.000	130,815
Age	48.37	46.92	49.25	0.000	130,815
Full-time contract	0.79	0.70	0.84	0.000	130,815
Temporary contract	0.21	0.30	0.16	0.000	130,815
Teacher salary scale					
1st (lowest)	0.49	0.52	0.47	0.000	130,815
2nd	0.20	0.20	0.20	0.462	130,815
3rd	0.19	0.16	0.21	0.000	130,815
4th	0.07	0.06	0.08	0.000	130,815
5th	0.02	0.01	0.03	0.000	130,815
6th	0.01	0.00	0.01	0.000	130,815
7th	0.00	0.00	0.00	0.191	130,815
8th (highest)	0.00	0.00	0.00	0.317	130,815
School characteristics					
Most rural (Rural 1)	0.15	0.31	0.06	0.000	130,103
Moderate rural (Rural 2)	0.17	0.28	0.11	0.000	130,103
Least rural (Rural 3)	0.13	0.13	0.13	0.004	130,103
Urban	0.55	0.27	0.71	0.000	130,103
Single-teacher	0.07	0.17	0.01	0.000	130,115
Multigrade	0.20	0.40	0.08	0.000	130,115
Multi-teacher	0.73	0.43	0.91	0.000	130,115
Bilingual	0.29	0.37	0.20	0.000	63,409
Vraem	0.06	0.07	0.06	0.025	63,409
Frontier	0.14	0.13	0.17	0.000	63,409
Total enrollment	336.57	177.83	431.18	0.000	130,115
Basic services	0.81	0.65	0.90	0.000	130,815
Main language (Castilian Spanish)	0.89	0.80	0.95	0.000	130,815
N.	130,928	49,314	81,614		

Note: The universe of this analysis comprises teachers in the 2018 Nexus database for whom a match in the 2018 SIAGIE teacher database is available.

Table A4. Student test score data (ECE).

	All students	2nd grade only (A)	2nd and 4th grade (B)	p-value A=B	N.
Student characteristics					
Low-SES (student's mother did not complete secondary school)	0.74	0.80	0.73	0.000	97,950
Main language (Castilian Spanish)	0.81	0.80	0.81	0.000	97,507
Household size	5.35	5.58	5.31	0.000	95,430
School characteristics					
Most rural (Rural 1)	0.03	0.08	0.02	0.000	267,482
Moderate rural (Rural 2)	0.07	0.13	0.05	0.000	267,482
Least rural (Rural 3)	0.09	0.11	0.08	0.000	267,482
Urban	0.81	0.68	0.84	0.000	267,482
Single-teacher	0.00	0.00	0.00	0.000	267,751
Multigrade	0.06	0.15	0.03	0.000	267,751
Multi-teacher	0.94	0.84	0.97	0.000	267,751
Bilingual	0.19	0.28	0.15	0.000	64,061
Vraem	0.07	0.07	0.07	0.748	64,061
Frontier	0.20	0.12	0.24	0.000	64,061
Total enrollment	528.03	434.26	552.39	0.000	267,751
Basic services	0.93	0.87	0.94	0.000	267,751
N.	267,751	55,223	212,528		

Note: The universe of this analysis comprises students in the SIAGIE database for whom at least the 2016 ECE 2nd grade score is available.

Table A5. Teacher hiring process data (centralized stage - PUN)

	All teachers	No PUN (A)	PUN (B)	p-value A=B	N.
Teacher characteristics					
Female	0.74	0.74	0.71	0.006	9,576
Age	50.06	52.70	38.08	0.000	9,576
Full-time contract	0.88	0.99	0.34	0.000	9,576
Temporary contract	0.12	0.00	0.66	0.000	9,576
School's characteristics					
Most rural (Rural 1)	0.03	0.02	0.07	0.000	10,407
Moderate rural (Rural 2)	0.08	0.06	0.12	0.000	10,407
Least rural (Rural 3)	0.10	0.10	0.11	0.061	10,407
Urban	0.80	0.82	0.70	0.000	10,407
Single-teacher	0.00	0.00	0.00	0.054	10,414
Multigrade	0.06	0.04	0.10	0.000	10,414
Multi-teacher	0.94	0.96	0.90	0.000	10,414
Bilingual	0.13	0.13	0.13	0.960	2,645
Vraem	0.07	0.06	0.08	0.157	2,645
Frontier	0.20	0.23	0.14	0.000	2,645
Total enrollment	493.92	507.43	449.60	0.000	10,414
Basic services	0.93	0.94	0.88	0.000	10,415
Main language (Castilian Spanish)	0.97	0.98	0.97	0.016	10,415
N.	10,415	7,981	2,434		

Note: The universe of this analysis comprises instructors in the SIAGIE student panel database teaching in 2018.

Table A6. Teachers in centralized vs decentralized stage of teacher hiring process

	All teachers	No decentralized stage (A)	Passed to decentralized stage (B)	p-value A=B	N.
Teacher characteristics					
Female	0.73	0.74	0.70	0.069	2,434
Age	36.65	37.15	35.06	0.000	2,434
Teaching experience in public schools					
No experience	0.13	0.13	0.13	0.963	2,434
< 2 years	0.25	0.26	0.24	0.436	2,434
3-5 years	0.32	0.30	0.36	0.007	2,434
6-10 years	0.23	0.22	0.23	0.561	2,434
>10 years	0.07	0.09	0.04	0.000	2,434
Teaching experience in private schools					
No experience	0.32	0.36	0.26	0.000	2,434
< 2 years	0.24	0.24	0.25	0.442	2,434
3-5 years	0.22	0.21	0.24	0.069	2,434
6-10 years	0.15	0.14	0.18	0.003	2,434
>10 years	0.06	0.05	0.06	0.532	2,434
Education					
Studied at an institute	0.03	0.03	0.03	0.969	2,434
Studied at a university	0.60	0.64	0.52	0.000	2,434
Studied at an institute and in a university	0.37	0.33	0.45	0.000	2,434
PUN score					
Total	124.07	110.01	148.90	0.000	2,434
Reading comprehension	36.14	33.20	41.33	0.000	2,434
Logical reasoning	28.75	24.12	36.92	0.000	2,434
Curriculum knowledge	59.19	52.70	70.65	0.000	2,434
N.	2,434	1,554	880		

Note: The universe of this analysis comprises instructors in the SIAGIE student panel database teaching in 2018 for whom centralized teacher evaluation scores are available.

Table A7. Descriptive statistics of TVA and teacher characteristics – restricted sample

	Same teacher in 3rd and 4th grade			Full sample		
	N.	Mean	Std. Dev.	N.	Mean	Std. Dev.
Female	1,674	0.74	0.44	2,835	0.75	0.43
Age	1,674	38.17	6.31	2,835	38.44	6.51
Temporary contract	1,674	0.71	0.46	2,835	0.74	0.44
1st salary level (lowest)	1,674	1.00	0.05	2,835	1.00	0.05
2nd salary level	1,674	0.00	0.02	2,835	0.00	0.02
3rd salary level	1,674	0.00	0.02	2,835	0.00	0.02
4th salary level	1,674	0.00	0.03	2,835	0.00	0.04
Had > 3 years of experience in 2018	1,674	0.89	0.31	2,835	0.89	0.32

Table A8. TVA estimation regression results – Teachers with more than 5 students

	4th grade ECE score							
	Same teacher in 3rd and 4th grade		Full sample		Same teacher in 3rd and 4th grade		Full sample	
	One-step	Two-step	One-step	Two-step	One-step	Two-step	One-step	Two-step
	Math	Math	Math	Math	Reading	Reading	Reading	Reading
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student characteristics								
High-SES (mother completed secondary school)	6.0424***	6.2269***	5.9049***	6.1271***	9.9352***	10.4524***	9.8677***	10.3238***
	(0.4827)	(0.5676)	(0.4347)	(0.5041)	(0.4961)	(0.5533)	(0.4489)	(0.4966)
2nd grade note in subject (Ref. Expected (A))								
Outstanding (AD)	48.1519***	44.0170***	48.0268***	43.7582***	45.0808***	40.4163***	45.3414***	40.6439***
	(0.5723)	(0.6043)	(0.5123)	(0.5329)	(0.5907)	(0.5882)	(0.5245)	(0.5181)
In progress (B)	-39.5066***	-28.5406***	-38.8745***	-28.2300***	-36.4926***	-27.0985***	-36.1771***	-27.2183***
	(0.8349)	(0.9028)	(0.6856)	(0.7375)	(0.9398)	(0.9715)	(0.7808)	(0.8100)
Initial (C)	-52.9222***	-46.4824***	-53.4592***	-46.5824***	-48.2871***	-44.1665***	-48.2693***	-43.8274***
	(1.6718)	(1.8629)	(1.5639)	(1.8156)	(1.7695)	(1.8738)	(1.6179)	(1.7951)
2nd grade ECE score in subject	0.4353***	0.4090***	0.4225***	0.4117***	1.0893***	1.2958***	1.0971***	1.2602***
	(0.0214)	(0.0232)	(0.0201)	(0.0217)	(0.0378)	(0.0393)	(0.0379)	(0.0395)
2nd grade ECE score in other subject	0.4864***	0.7441***	0.4791***	0.6761***	0.0046	-0.0258	0.0015	-0.0263
	(0.0368)	(0.0402)	(0.0351)	(0.0380)	(0.0220)	(0.0225)	(0.0209)	(0.0215)
2nd grade ECE score - Math^2	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0002***	-0.0001***	(0.0209)	-0.0002***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0209)	(0.0000)
2nd grade ECE score - Reading^2	-0.0005***	-0.0007***	-0.0005***	-0.0007***	-0.0008***	-0.0009***	(0.0209)	-0.0009***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0209)	(0.0000)
2nd grade ECE score - Math*Reading	0.0006***	0.0006***	0.0006***	0.0006***	0.0005***	0.0005***	(0.0209)	0.0006***
	(0.0001)	(0.0001)	(0.0000)	(0.0001)	(0.0001)	(0.0001)	(0.0209)	(0.0001)
Classroom aggregates								
High-SES students (%)		45.1461***		40.5415***		44.1489***		39.9657***
		(1.3454)		(1.1403)		(1.3671)		(1.1692)
Mean 2nd grade ECE score in subject		0.0399***		0.0261***		0.0198***		-0.0086
		(0.0042)		(0.0038)		(0.0068)		(0.0062)
School characteristics								
Rural school			-20.5810*	-7.1212***			-33.3556***	-10.0111***
			(12.0241)	(1.0367)			(9.8904)	(0.9974)
Total enrollment			-0.0131	0.0067***			-0.0175**	0.0067***
			(0.0081)	(0.0006)			(0.0079)	(0.0006)
Constant	63.1912***	-53.7221***	76.0878***	-28.8921***	-22.1245**	-102.9383***	-13.5707	-79.0539***
	(10.1974)	(10.4147)	(10.9342)	(9.7815)	(10.4771)	(10.2842)	(11.6220)	(10.2592)
Teacher FE	Yes	No	Yes	No	Yes	No	Yes	No
N.	102488	102488	160031	160031	102523	102523	160079	160079
N_g (number of teachers)	8687	8687	11945	11945	8691	8691	11942	11942
R2	0.6308	0.4281	0.6191	0.4247	0.6011	0.4440	0.5910	0.4410

***p<0.01; **p<0.05; *p<0.1. Standard errors in the full sample estimations are clustered at the student level.

Table A9. TVA estimation regression results – Urban schools

	4th grade ECE score							
	Same teacher in 3rd and 4th grade		*Full sample*		*Same teacher in 3rd and 4th grade*		*Full sample*	
	One-step	Two-step	One-step	Two-step	One-step	Two-step	One-step	Two-step
	Math	Math	Math	Math	Reading	Reading	Reading	Reading
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student's characteristics								
High-SES (mother completed secondary school)	6.1402***	6.2903***	5.9988***	6.1927***	10.0320***	10.5348***	10.0095***	10.4479***
	(0.4873)	(0.5716)	(0.4383)	(0.5070)	(0.5012)	(0.5576)	(0.4525)	(0.4993)
2nd grade note in subject (Ref. Expected (A))								
Outstanding (AD)	48.2158***	44.4097***	48.0999***	44.0714***	44.9737***	40.3525***	45.2904***	40.6608***
	(0.5833)	(0.6149)	(0.5232)	(0.5418)	(0.6024)	(0.5987)	(0.5350)	(0.5261)
In progress (B)	-39.6077***	-28.8951***	-38.8737***	-28.2988***	-36.6784***	-27.7549***	-36.3184***	-27.5268***
	(0.8434)	(0.9094)	(0.6913)	(0.7426)	(0.9507)	(0.9799)	(0.7888)	(0.8164)
Initial (C)	-53.4564***	-46.4174***	-53.8134***	-46.2409***	-49.0002***	-44.4311***	-48.9380***	-44.0184***
	(1.7139)	(1.9069)	(1.6183)	(1.8685)	(1.8201)	(1.9259)	(1.6753)	(1.8521)
2nd grade ECE score in subject	0.4362***	0.4076***	0.4229***	0.4106***	1.1015***	1.2605***	1.1088***	1.2509***
	(0.0220)	(0.0239)	(0.0208)	(0.0224)	(0.0389)	(0.0405)	(0.0391)	(0.0406)
2nd grade ECE score in other subject	0.4794***	0.7015***	0.4713***	0.6548***	-0.0149	-0.0445*	-0.0166	-0.0423*
	(0.0378)	(0.0414)	(0.0362)	(0.0392)	(0.0226)	(0.0232)	(0.0215)	(0.0222)
2nd grade ECE score - Math^2	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0002***	-0.0001***	-0.0002***	-0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
2nd grade ECE score - Reading^2	-0.0005***	-0.0007***	-0.0005***	-0.0007***	-0.0008***	-0.0009***	-0.0008***	-0.0009***
	(0.0000)	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
2nd grade ECE score - Math*Reading	0.0006***	0.0006***	0.0006***	0.0006***	0.0006***	0.0005***	0.0006***	0.0006***
	(0.0001)	(0.0001)	(0.0000)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Classroom aggregates								
High-SES students (%)		43.0933***		39.4933***		42.6633***		39.1757***
		(1.3554)		(1.1391)		(1.3737)		(1.1690)
Mean 2nd grade ECE score in subject		0.0414***		0.0301***		0.0092		-0.0076
		(0.0043)		(0.0038)		(0.0070)		(0.0063)
School's characteristics								
Total enrollment			-0.0124	0.0069***			-0.0190**	0.0068***
			(0.0082)	(0.0006)			(0.0081)	(0.0006)
Constant	66.2536***	-39.3714***	78.4394***	-24.0199**	-18.6311*	-78.6483***	-10.4399	-71.3323***
	(10.5453)	(10.9356)	(11.3264)	(10.2075)	(10.8415)	(10.8448)	(12.0432)	(10.7113)
Teacher FE	Yes	No	Yes	No	Yes	No	Yes	No
N.	98929	98929	154913	154913	98965	98965	154964	154964
N_g (number of teachers)	8448	8448	12026	12026	8453	8453	12025	12025
R2	0.6238	0.4195	0.6123	0.4163	0.5918	0.4330	0.5819	0.4299

***p<0.01; **p<0.05; *p<0.1. Standard errors in the full sample estimations are clustered at the student level.