I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14277

Effects of Measures of Teachers' Quality on Tertiary Education Attendance: Evaluation Tests versus Value Added

Juan Díaz
Rafael Sánchez
Gabriel Villarroel
Mauricio G. Villena

APRIL 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14277

Effects of Measures of Teachers' Quality on Tertiary Education Attendance: Evaluation Tests versus Value Added

Juan Díaz
Universidad de Chile

Rafael Sánchez
Universidad Diego Portales and IZA

Gabriel Villarroel
Ministry of Finance of Chile

Mauricio G. Villena
Universidad Diego Portales

APRIL 2021

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Effects of Measures of Teachers' Quality on Tertiary Education Attendance: Evaluation Tests versus Value Added*

Using Chilean administrative datasets for the period 2011-2017, we study which of the most used tools to evaluate teacher quality, namely teachers' evaluation tests (TET) and teacher's value added (TVA), predicts more accurately not only short run (as most of the literature focus on) but also middle run students' outcomes. For this evaluation we follow the same cohorts of students and teachers. Our results suggest that the correlation between (TET) and (TVA) appears to be null in school outcomes. However, our analysis also reveals that both measures, TET and TVA, positively affect the probability of tertiary education attendance, indicating that both measures are complementary in measuring teacher quality in the middle run. These results have relevance from the public policy point of view as unlike countries (e.g. USA) where TVA is used for teacher's promotions and personnel decisions, in countries where TVA is not used for teacher's personnel decisions (e.g. Chile), TVA seems to be useful to measure teacher quality. Furthermore, our findings are consistent with the argument of the multidimensionality of teaching quality, because even though in the short run TVA and TET seem to be orthogonal, in the medium run they seem to be complementary tools to measure teacher effectiveness.

JEL Classification: I20, I23, I26, I28

Keywords: teacher quality, value added, teacher evaluation test, Chile

Corresponding author:
Gabriel Villarroel
Budget Office
Ministry of Finance of Chile
Teatinos 120
Santiago
Chile

E-mail: gvillarroel@dipres.gob.cl

* We thank the Ministry of Education of Chile and the Budget Office of Chile for providing access to the data used in this work.

1 Introduction

The existing literature on teacher effectiveness suggests that, after controlling for school and class variables, teachers are more important to student learning than any other factor (Rivkin, et al. 2005). However, knowledge about what works in teacher effectiveness is less well understood, as teachers significantly vary in their ability to improve students' performance (Hanushek and Rivkin, 2010; Grossman et al., 2013). Consequently, teacher effectiveness measures have been introduced to differentiate teacher performance better. To date, the most commonly used instruments in the US and abroad have been teacher value-added (TVA) and teacher evaluation tests (TET), and in particular, a combination of both, mainly because it is not completely clear what dimension of teaching they are truly measuring (Chin and Goldhaber, 2015).

Having in mind that good teaching is multidimensional and that each instrument of evaluation may vary in what dimension of teacher performance they measure, a key policy question is how different teacher effectiveness measures help explain children's future performance and what they really measure. In order to do this in a right way it is important to correctly compare these measures, using data of the same cohorts and for the same period of time, evaluating the same performance variable. Unfortunately, to date this has not been possible due to data has limited research on this issue.

This paper addresses this gap. We examine two measures of teacher's quality and their impact on tertiary education attendance using a novel education data set constructed by merging several Chilean administrative data sets for the period 2011-2017. The two alternative instruments that measure teachers' effectiveness for the same sample of teachers and students are: the Chilean national teachers' evaluation test (TET) and the traditional teacher value added results approach (TVA) used in the literature (see: Chetty et al. 2014a, 2014b).[1]

We deviate from previous studies in four dimensions. First, to our knowledge this is the first attempt to evaluate and compare two different approaches to measure teacher's quality using the same sample of students and teachers to study each of them, including all students who graduated from public schools for the same period. Second, we also contribute to the literature by being the first to compare these two instruments by their medium-run effects, in this case by examining tertiary education attendance. This is important as the traditional test scores used in the literature for the short run analysis are imperfect measures of learning and may be weakly linked to future outcomes such as tertiary education attendance, adult salaries and success in life. Indeed, recent studies have documented that effects on long run outcomes may go undetected by test scores (e.g. Jackson et al. 2014, Chetty et al. 2014a, 2014b, Heckman et al. 2014). This is why we address the limitation of focusing not only in the short run school performance but also on the medium run tertiary education attendance. Third, we study the Chilean case in which the current policy, mandated by law, uses TET results to evaluate teachers, having a direct impact on teachers' salaries and careers. By contrast, TVA is not used at all in Chile, in spite of the fact that all necessary data is available in Chilean public databases to estimate teachers' value-added. Thus, it would be interesting to find out whether TVA positively affects the probability of tertiary education attendance, corroborating that this measure is a good indicator of teacher quality in this context, despite not being in the Chilean teacher's policy economic incentive scheme. Hence, Chile could provide an interesting case study to contrast with the United States one, in which several school districts across the country had already adopted the TVA system, including the Chicago Public Schools, New York City Department of Education and District of Columbia Public Schools. In some of these districts the TVA rankings have been used to decide on issues of teacher retention and the awarding of bonuses, as well as a tool for identifying those teachers who would benefit most from teacher training. Fourth, evaluating and comparing two different approaches to measure teacher's quality in the context of a developing country like Chile, could allow us to weigh the potential costs

[1] For previous studies evaluating school value added in less developed countries, see *inter alia*: Andrabi et al. (2011), Singh (2015), Singh (2019) and Bau and Das (2020)

of each, giving an account of certain instruments that could be applied to the extent that their economic resources allow. Indeed, this analysis could give us some insight into which measurement is more cost effective.

Teachers' Evaluation tests are a set of teacher evaluations based on protocols (such as CLASS, MQI, PLATO, and FFT) that have been developed around the world to improve teachers' effectiveness based on the idea that teacher evaluation can be a way to improve teachers' performance, either by making it possible to provide them with useful feedback or by creating incentives to implement better practices (Isoré, 2009; Taylor and Tyler, 2012). Other authors, such as Wyness et al. (2018), have questioned teachers' evaluations, arguing that teacher observation and feedback cannot solve the policy maker's problem of vast variations in teacher effectiveness in the absence of teacher incentives and non-peer feedback and thus cannot be used to reduce differences in teacher effectiveness significantly. Additionally, teachers' evaluations take many different forms across the world and vary greatly in terms of resources involved per teacher. For example, observation protocols used in research and practice, such as the FFT, the MQI, or CLASS, all broadly capture a teacher's effectiveness in delivering quality instruction. However, each evaluates teachers on different subjects and classroom practices. These characteristics have led to no consensus on what a good evaluation system should be and how intensive it should be (Isoré, 2009; OECD, 2013a and 2013b; Jackson et al., 2014).

By contrast, value-added measures circumvent the need to identify specific teacher characteristics related to quality and shift the focus to identifying overall teacher contributions to learning. However, it introduces additional complications and has sparked an active debate, especially about validity and reliability measures (see, for example, Chetty et al., 2014a and 2014b; Rothstein, 2010; Koedel et al., 2015; Bau and Das, 2020, among others). The traditional value-added approach rests on the assumption of selection in observables, which depends on the available information since the method consists of isolating the contribution that each teacher has in the test score of their students from the residualization of the test score once eliminating the effects of the rest of the observable variables that affect academic performance in addition to the teacher's quality. The challenges are data availability, measurement precision and bias, and dimensionality of what is understood by teaching quality. Issues that have generated debate because of the consequences of these results on teachers, such as dismissal from school.

Few studies have investigated the relationship between these two instruments, concluding that the relationship between the two is "modest" or "weak" (e.g., Kane et al., 2013; Lynch et al., 2017). In particular, the usual correlation between these two measures ranges between 0.1 and 0.3. These findings contradict what many scholars and practitioners might expect because theory and intuition suggest that teachers' strong instructional practices should improve student test performance. The determinants of the weak correlation between the instruments are not clear, mainly because true teacher quality is unobservable. However, there are some attempts to explain it. Chin and Goldhaber (2015) use simulations to study three possible scenarios that result in weak correlations between value-added and teacher evaluations. The first is that one or both measures could provide unreliable estimates of one or more dimensions of teacher quality due to sampling error. The second is that teacher quality may be multidimensional, and the measures provide reliable estimates of different dimensions of teacher quality. The third is that one or more of the measures may be invalid because they do not provide a reliable estimate of any dimension of teacher quality. They find that the simulations did not allow them to rule out any of the scenarios for the weak correlations seen in prior research. Chaplin et al. (2014) analyze the case of the Pittsburgh Public Schools teacher evaluation system, which includes teacher observation based on protocols, value-added measures, and student surveys. They find that i) all three measures can differentiate among teachers and ii) correlations among measures suggest they are valid and complementary. Other authors have pointed out that given the multidimensionality of teaching quality, since the multiple measures approach helps create a composite that is more representative of stakeholders' value, validity and reliability would improve (Darlin-Hammond et al., 2015; Harris, 2012). To

the best of our knowledge, no comparison of their medium-run effects, such as tertiary education attendance as we present here, has been made between these two instruments.

In particular, in this work we measure teachers' quality using TET and TVA for the same sample of teachers and estimating the effect of both measures on the probability of tertiary education attendance of their students in Chile. We do this by building a new and unique longitudinal data set based on an administrative database of educational records of 3.4 million Chilean students. For each student, we have nine years of data, from 4th year of primary school to 12th year (final) of secondary school, as well as detailed records of all the teachers. The resulting data set covers the period 2011-2018 and includes (i) all students who were enrolled in the Chilean educational system between 2011 and 2016 (4th primary school grade to the last grade of secondary school), for whom we add their tertiary education records for the period 2012-2018; (ii) all teachers who were instructing in the Chilean educational system between 2011 and 2016; and (iii) an identifier that allows us to link students with their teachers in each grade for the period 2011-2016. We also have school history information of students for the period 2011-2016. This data set has 6 years of data for 3,415,100 students enrolled in any of the 9 grades that we look at (4th primary year to the last year of secondary school). Importantly, this data set contains the census of students in the Chilean school system, which means an average of 234 thousand students by cohort.

These administrative records include reliable information on the grades that students obtained in each school year between 2011 and 2016. This information allows us to know students' final grades from a specific year-school-school grade-class-subject perspective. We also use two sources of administrative records about teachers for the period 2011-2016 . The first corresponds to the yearly teacher census, which includes a characterization of the whole teacher universe in the system in each period of study. In total, we follow 84,719 unique teachers. This database has information on the year-school-school grade-class-subject in which the instructor taught, allowing us to link it with each student's grades. Importantly, the teachers' administrative records can be linked with the teachers' evaluation tests. Since these evaluations are mandatory for the public system, we have a full record of public school teachers' results for 2011-2016.

Finally, we use an administrative database that provides information on the enrollment of students in tertiary education institutions. This data set allows us to identify each high school graduate's tertiary education enrollment status, i.e., whether they enrolled after graduation or not. Additionally, for the students who entered tertiary education, the database indicates which type of institution they enrolled in, differentiating between universities, technical formation centers, and professional institutions. Based on students who graduated from high school between 2012 and 2017, 1,163,343 students are represented in the database through 7,988,659 observations. We observed the entire academic school record for each student who graduated. Of these, 615,977 students were enrolled in tertiary education immediately after graduating, e.g., an attendance rate of 53

In terms of methodology, first, in order to estimate the value-added of a teacher, we follow Chetty et al. (2014a) and (2014b), as well as the contributions of 17. Kane and Staiger (2008), Rothstein (2010) and Kane et al. (2013). Essentially, the teacher's value-added is defined as each teacher's contribution to their students' academic performance. If teachers' and students' class allocation were random, estimating this contribution would not be challenging. However, it is well known that allocation is not random in Chile but rather highly selective. Parents choose their children's schools, schools select which students can enroll in them, and school directors choose which teachers the school hires and assigns them to a specific class. Therefore, if students with certain characteristics (higher skills, better grades in previous years, previous school, higher-income group, and religious beliefs) are systematically assigned to a specific type of teacher (longer experience, better performance, men instead of women, among others), not adjusting for students and teachers' characteristics will bias our value-added estimations.

To identify the teacher's contribution to their students' academic achievements, one could adjust the estimations for a substantial number of measured variables before the interaction between

a teacher and a student. In other words, one can assume that the teacher-student match is random conditional on a set of observables. More precisely, these variables will control students' characteristics, students' parents' characteristics, other teachers' and other schools' contribution to the students' academic performance, schools' student selection, directors' teachers' selection, and assignment to a class. The plausibility of this assumption depends on the quality of the variables that we control.

Second, the methodology used for estimating the impact of teacher evaluation also consider the fact that in order to identify causal effects, the assumption of no unmeasured confounders must hold. This means that unobservable variables should be uncorrelated with the treatment variable and the variable of interest (potential outcome) (Rubin, Stuart and Zanutto, 2004). In our case, the assumption would imply that there are no unmeasured variables that affect the treatment (i.e., teaching evaluation) and potential outcome (i.e., student performance). Unfortunately, we do not have a randomized experiment to ensure the assumption just mentioned. Thus, as observational data are used in our study, several potential sortings between and within schools might complicate our analysis as they challenge the no unmeasured confounders assumption, as has been explained during this investigation.

All potential selection sources (student-school selection, teacher-school selection, and student teacher selection) may generate non-random student-teacher sortings. In fact, there are several previous pieces of empirical evidence showing that student-teacher sorting (between and within schools) is not random (see Jackson, 2014 and Rothstein, 2010, among others). To address these concerns, we follow two approaches. The first of these, based on previous research (Kane et al., 2008, Kane et al., 2011, Briole and Maurin, 2019, among others), attempts to quantify teacher evaluation impacts on certain outcomes through a linear estimation controlling for the potential biases mentioned above. For the second approximation, we will use a strategy similar to that used to estimate the impact of value-added on educational outcomes.

Finally, in order to compare the two measures of teacher's quality and their impact on entry into tertiary education we use the same regression of the previous analyses, including both measurements simultaneously to corroborate whether the found results are maintained. Considering both instruments' application to the same sample, this time we only consider students who have graduated from public schools.

Our results suggest that teachers' quality based on TET is not related to those based on the TVA approach, with a correlation between these two measures of -0.02. Our results also suggest that both instruments are helpful predictors of tertiary education attendance: An standard deviation (SD) increase of 1 in a teacher's true TVA test score in a single grade increases the probability of tertiary education attendance by 0.6-0.7 percentage points, i.e., an increase of 1.3%-1.5% with respect to mean tertiary education attendance in public schools. An SD increase of 1 in a teacher's evaluation in a single grade increases the probability of tertiary education attendance by 1.7-1.9 percentage points, i.e., an increase of 3.5%-4.1% with respect to mean tertiary education attendance in public schools, where two (portfolio and external references) out of four parts of the TET are the best predictors for graduate students' tertiary education attendance.

The key finding of our study is that, unlike the USA where TVA is used for teacher's promotions and personnel decisions, in countries like Chile, where TVA is not used for teacher's personnel decisions, TVA also seems to be a useful tool to measure teacher quality in the medium run. Furthermore, our findings are consistent with the argument of the multidimensionality of teaching quality, because even though in the short run TVA and TET seem to be orthogonal, in the medium run they seem to be complementary tools to measure teacher effectiveness.

The structure of the article is as follows. Section 2 describes Chile's institutional background. Section 3 presents and describes the data sets used in our analyses as well as some summary statistics. Section 4 presents the empirical approach and results for the TVA methodology, while section 5 does the same for teacher evaluation methodology. Section 6 shows a comparison of both previous results. Finally, section 7 concludes.

2 Institutional Background

2.1 General Structure of the Chilean Educational System

The Chilean educational system consists of four educational levels: preschool (0 to 4 years old), primary school (kindergarten and 1st to 8th grade), secondary school (9th to 12th grade), and tertiary education. Primary school and secondary school (high school) are mandatory, while preschool and tertiary education are optional.

There are 11 subjects in each mandatory level, including mathematics, language, and science. The grading scale for each subject corresponds to a numeric scale that ranges from 1.0 to 7.0. A grade of 4.0 or higher and a minimum attendance of 85% are required to pass a subject. For being promoted to the next grade, students must pass all subjects; however, a student could still be promoted if she fails one subject but has an average grade across all subjects of 4.5 or higher or if she fails two subjects but has an average grade across all subjects of 5.0 or higher.

It is important to note that students' grades are not used to determine teacher quality or give the educational community any incentive but are only used for grade promotion or admissions to other schools or tertiary education. Nevertheless, standardized test results impact schools and teachers; first, in terms of their autonomy and support received from the Ministry of Education and, for the second, in terms of monetary incentives associated with the National Performance Evaluation System (NPES)(Sistema Nacional de Evaluación de Desempeño in Spanish[2].

Since the decentralization of the educational system in 1980, the Chilean educational experience has been one of the most extreme cases of the introduction of market-oriented reforms at a national level (e.g., universal parent school choice, voucher schools, copayment system, standardized measurements, and multiple for-profit and not-for-profit private schools). The Chilean educational system considers three kinds of administrative alternatives: public establishments under municipal administration (i.e., public schools); private subsidized establishments funded by a voucher system and administered by the private sector (i.e., voucher schools); and private fee-paying establishments funded and administered by the private sector. Regarding students' distribution, approximately 40% of students are in public schools, 8% are in private, non-voucher (i.e., fee-paying) schools, and 52% are in private voucher schools.

The introduction of the voucher system gave parents complete freedom to choose schools for their children. Essential for this decision was introducing a standardized census-type test for all schools and students in the country. This test is known as the SIMCE and covers mathematics and language. The mere existence of this national test, along with the fact that school's results are made public, introduces an element of competitive pressure into the system, as parents have objective indicators of results to assess educational school outcomes.

Unlike voucher schemes implemented in other countries, private voucher and non-voucher schools in Chile can choose their students; however, public schools are prohibited from choosing, except in those cases where the demand for seats exceeds availability. This scheme, where private schools can select students, generates a positive sorting of students from high- and middle-income families into private schools and most vulnerable students into public schools (as found by Contreras et al., 2010). One of the reasons for this is that private voucher schools in Chile can operate for profit and may, therefore, select students who are less expensive to educate. This resulted in high segregation between private and public schools in Chile (Valenzuela et al., 2013).

Regarding job contracts, teachers in public schools are governed by the Teacher Statute, which is legislation that includes a centralized collective-bargaining process, wages based on uniform pay-

[2]The main objective of the NPES is to establish a mechanism that relates collective performance with salaries of teachers and the rest of the school staff. The establishments in the best 25 percent of their region receive a subsidy called Excellence Grant(Subvención de Excelencia in Spanish), those between the best 26 and 35 percent receive the 60 percent of this subsidy, and those located under the 35 percent do not receive anything. For 2020, the magnitude of the Excellence Grant is of CLP 5,482.84 per month per student in the case of teachers, and CLP 381.83 per month per student in the case of the rest of the school staff.

scales with special bonuses for training, experience, and working under difficult conditions, and strong restrictions on dismissals. On the other hand, private schools (in both voucher and non-voucher schools) operate as firms, and their teachers come under the labor code, such as all other private-sector workers.

In the context of a market-oriented educational system and parental school choice, school quality becomes crucial. The OECD's historical data suggest that Chile's student learning outcomes have been considerably below the OECD average. Furthermore, students' results differ considerably across the socioeconomic groups and type of school attended. In this context, the government accorded significant importance to teacher evaluation and generated conditions to establish a national evaluation test as described below.

2.2 Teachers Evaluation Test

Before the second round of the presidential campaign of 2000, one of the candidates (Ricardo Lagos, who won the election a few weeks later) signed an agreement with the National Teachers Union that included evaluating teacher performance. After a few years of work, the Chilean TET, *Evaluación Docente* in Spanish, was introduced in 2003 as a pilot program. Since 2004, it has been mandatory for public school teachers and optional for nonpublic school teachers. The Ministry of Education is responsible for the implementation of this test.

To evaluate teacher effectiveness, a technical committee composed of the Chilean Ministry of Education, the Association of Municipalities, and the National Teachers Union established the first set of effective teaching standards based on the Framework For Teaching (FFT) protocol presented by Charlotte Danielson. This set of standards is summarized in four domains, known as the Framework for Good Teaching (FGT), *Marco para la Buena Enseñanza* in Spanish (see Table 13). Each of these domains has several criteria (19 in total), and each criterion contains several descriptors (71 in total). Finally, each descriptor is disaggregated into observable elements of teaching practices that are measured by four different instruments on the teachers' evaluation test. These instruments are a) self-assessment, b) peer interviews, c) external references, and d) portfolio. The mapping between the instruments and the observable elements of the descriptors is established using a double-entry matrix. Some descriptors are evaluated with a few instruments, while others are evaluated with several instruments. As we mentioned above, this is helpful, as there is a considerable agreement regarding the use of several instruments to evaluate teacher effectiveness due to single instruments hardly capturing all aspects of teacher quality (Grossman et al., 2013; Manzi et al., 2011).

The average annual cost per teacher for the period of analysis (2012-2018) is US $400, covering approximately 20% of the total number of teachers in the public sector, which includes preparation of the instruments, application, review, and payment for peer interviewers, among others. The total cost for 2019 was US $10.2 million for the coverage of 21 thousand teachers, i.e., an average annual cost of US $461[3].

2.3 Teachers' Classification

Each descriptor has a language that describes performance at each level of the rubric: Distinguished, Proficient, Basic, and Unsatisfactory, with an evaluator assigning the respective scores of 4, 3, 2, and 1 to these rubric levels.

For each instrument (except the portfolio), the score is calculated by averaging all descriptors associated with each criterion and then averaging all the criteria to obtain each domain's score.

[3]The information corresponds to the execution of the Teacher Professional Development (*Desarrollo Profesional Docente* in Spanish), assignment that consolidates the resources for the teacher evaluation obtained from the Budget Office, http://www.dipres.gob.cl/598/w3-channel.html. The number of teachers evaluated corresponds to information obtained from the page of the Ministry of Education, http://datosabiertos.mineduc.cl/.

Finally, the score of the instrument is simply the average score of the domains. Therefore, all domains have the same weight in the final score of each of these instruments.

In the case of the portfolio, the FGT domains are operationalized in different dimensions. There are 7 dimensions in the portfolio (A, B, C, D, F, G, and H), as presented in Table 15. The score of the portfolio is obtained by averaging the scores of all dimensions. Each dimension has the same weight in the final score of the portfolio. Each teaching practice is evaluated with a score between 1-4 (from unsatisfactory to distinguished) to obtain the dimension's score.

Each of these four instruments is weighted (in different ways depending on how many times a teacher has taken the test) to obtain the final score. For first-time takers, Portfolio weights 60%, Peer Interview 20%, Self-Evaluation 10%, and External References 10%. With the weighted overall final score, teachers are evaluated in one of the following categories: i) unsatisfactory (1-1.75 points), ii) basic (1.75-2.5 points), iii) proficient (2.5-3.25 points), and iv) distinguished (3.25-4 points).

It is important to point out that the Chilean TET is taken every four years for teachers who were classified as distinguished or proficient. Those classified as basic must take training plans to overcome their weaknesses (funded by the Ministry of Education) and retake the test in two years. For those classified as unsatisfactory, the test must be taken the following year with a different weighting scheme. In this case, Portfolio weights 80%, Peer Interview 10%, Self-Evaluation 5%, and External References 5%.

If the teacher is again classified as unsatisfactory, the person is fired. If the teacher is instead classified as Basic in the second test, the teacher will have a new opportunity the following year. The only exemptions from the TET are teachers who are close to retirement (three years or less from retirement age).

The main benefits of distinguished and proficient are that these categories allow teachers to have preferential access to professional development opportunities, such as visiting overseas, promotions, academic seminars, and becoming workshop tutors. Furthermore, they have access to the possibility of increasing their wages, known as the Variable Individual Performance Allowance (Asignación *Variable por Desempeño Individual* in Spanish) and the Pedagogical Excellence Allowance (*Asignación de Excelencia Pedagógica* in Spanish), depending on their performance on a written test about discipline and pedagogical knowledge.

2.4 Classroom Observations: Instruments

Self-Evaluation

This instrument consists of a series of questions that have the goal of making teachers reflect on their pedagogical practices, evaluating the quality of their relationship with the students and their parents, recognizing the quality of their performance in the classroom, and identifying their strengths, weaknesses, and need for professional development. Teachers also receive an example of the very specific rubric used for this instrument. Each teacher must evaluate each of the questions with the rubric (distinguished, proficient, basic, or unsatisfactory) and submit them online after it is completed.

An example of the general concept used to construct the rubrics (common for all the indicators in all the instruments) is:

1. Unsatisfactory: means that the evaluated teacher presents clear weaknesses in their performance evaluated by the indicator.

2. Basic: means that the evaluated teacher fulfills the expected performance, although in an irregular way.

3. Proficient: means a correct performance, a level that fulfills what is requested by the indicator, although not exceptional. This is the expected performance.

4. Distinguished: means that the teacher has a clear and consistently better performance concerning what is expected by the indicator.

Peer Interview

This instrument consists of an interview by a peer, another teacher who works under similar conditions, known as Peer Evaluator(PE). The PE is specially trained for this task. Training includes the application of the interview and the rubric to ensure a standardized evaluation. Questions included in this instrument are based on the FGT mentioned above and can be classified into three parts: i) general information about the evaluated teacher and the interviewer, ii) questions regarding performance in the classroom, and iii) questions regarding the context under which the performance is done. The PE should register every answer, and immediately after the interview, the PE must evaluate each of them with the rubric (distinguished, proficient, basic, or unsatisfactory) and submit them online after its completion. This interview takes approximately 60 minutes.

External References

This instrument consists of an external evaluation from the teacher's two hierarchical superiors (who in general are the principal of the school and the chief of the pedagogical unit of the school (UTP)). This instrument is a very precise and structured guide based on the FGT that includes several questions about the evaluated teacher's performance. For each question, the principal and the chief of the UTP separately must assess the performance of the evaluated teacher using the same rubric of the two previous instruments (unsatisfactory, basic, proficient, and distinguished) using the values 1-4, respectively, and submit each score online after its completion. The results for this instrument are obtained by averaging both scores.

Portfolio

This instrument is an evaluation where teachers must present evidence of their pedagogical practice. The portfolio must be done in the teacher's subject (and the one registered in the TET). Teachers have 12 weeks for the elaboration of the following material separated into two modules:

Module 1: It includes two products. First, the design and implementation of an 8-hour pedagogical unit. In this product, teachers must develop clear and specific goals for the pedagogical unit and its classes. Additionally, they have to describe each of the classes implemented within the unit, indicating the date, duration, realized activities, and used resources, among other things. Additionally, teachers must answer some questions about a) their experience implementing the unit and b) their pedagogical performance in the classroom. Second, teachers must present an evaluation of the student's learning in that pedagogical unit. The goal is to gather information regarding what students have learned in that unit. If, for example, a written test was used to evaluate students, the teacher must send a copy of that test with the correct answers or the appropriate criteria used to evaluate each test's answer. Additionally, teachers must answer questions related to a) their experience in applying the evaluation and b) their analysis and use given to the results (such as students' feedback and the teacher's feedback for improving their practice, among others).

Module 2: Includes only one product: a complete 40-minute video recording (without cuts or interruptions) of one of the classes the evaluated teacher usually works with. The teacher then completes a form with information relative to the recorded class. Additionally, the evaluated teacher must attach a photocopy of the resources used in that class.

For the portfolio, each evaluated teacher receives a very detailed instruction manual to elaborate on each of the requested products so that every teacher delivers the requested outputs in a standardized format. The class's recording is the responsibility of a trained cameraman, and it is free of charge for the evaluated teacher. Additionally, there are several measures in place to

ensure sound and image quality. The cameraman has specific instructions so that the recording must clearly show the teacher and the students to reflect what happens in the classroom.

From these two modules, 7 dimensions are obtained (see Table 15). Each includes the indicators of teaching practices with rubrics based on the FGT. Later, the modules are evaluated by specially trained peers designated by the Ministry of Education. In their evaluations, these peers must follow very strict rules to guarantee an objective and blind process, including a blind double evaluation and score recalibration in rare cases, among several other processes (see Chapter 2 of Manzi et al., 2011 for full details).

3 Data

We build a new and unique longitudinal data set based on an administrative database of educational records of 3.4 million Chilean students. For each student, we have nine years of data, from 4th year of primary school to 12th year (final) of secondary school, as well as detailed records of all the teachers.

The data were collected from administrative records of the Education and Finance Ministries of Chile. The resulting data set covers the period 2011-2018 and includes (i) all students who were enrolled in the Chilean educational system between 2011 and 2016 (4th primary school grade to the last grade of secondary school), for whom we add their tertiary education records for the period 2012-2018; (ii) all teachers who were instructing in the Chilean educational system between 2011 and 2016; and (iii) an identifier that allows us to link students with their teachers in each grade for the period 2011–2016. The administrative records used to build the longitudinal data set that we use in this paper and the variables included in it are described below.

School history information of students for the period 2011-2016

This data set has 6 years of data for 3,415,100 students enrolled in any of the 9 grades that we look at (4th primary year to the last year of secondary school). Importantly, this data set contains the census of students in the Chilean school system, which means an average of 234 thousand students by cohort.

These administrative records were provided by the Chilean Ministry of Education and included reliable information on the grades that students obtained in each school year between 2011 and 2016. This information allows us to know students' final grades from a specific year-school-school grade-class-subject perspective. For our analysis, we focus on language (Spanish) and mathematics. Moreover, the administrative records contain information on the final grade for each school year, the final student's situation (promoted to the next grade, repetition of grade or school transfer), gender (1=woman), vulnerability condition (1=if the student belongs to the most vulnerable 40% of the population), geographic location of the school, administrative dependency (public, private voucher or private non-voucher), and rurality condition (1=urban).

Based on this information, we construct variables that indicate school and classroom sizes and their composition in terms of sex, vulnerability proportion, and a grade average of the previous and contemporary classes.

Information on teachers for the period 2011-2016

We use two sources of administrative records about teachers. The first corresponds to the yearly teacher census, which includes a characterization of the whole teacher universe in the system in each period of study. In total, we follow 84,719 unique teachers. This database has information on the year-school-school grade-class-subject in which the instructor taught, allowing us to link it with each student's grades. Furthermore, it includes data on each teacher's characteristics, such

as gender (1=woman), age (years), pedagogical hours, and their role in the school, which can be teaching or other (e.g., managerial duties).

Importantly, the teachers' administrative records can be linked with the teachers' evaluation tests described in section 3.2. Since these evaluations are mandatory for the public system, we have a full record of public school teachers' results for 2011-2016.

Information on tertiary education enrollment

We use an administrative database that provides information on the enrollment of students in tertiary education institutions. This data set allows us to identify each high school graduate's tertiary education enrollment status, i.e., whether they enrolled after graduation or not. Additionally, for the students who entered tertiary education, the database indicates which type of institution they enrolled in, differentiating between universities (that offer high-level professional and technical degrees), technical formation centers (that offer only high-level technical degrees), and professional institutions (that offer professional and technical studies that do not lead to academic degrees). Both modalities, technical formation centers and professional institutions will be called Vocational Education.

Based on students who graduated from high school between 2012 and 2017, 1,163,343 students are represented in the database through 7,988,659 observations. We observed the entire academic school record for each student who graduated. Of these, 615,977 students were enrolled in tertiary education immediately after graduating, e.g., an attendance rate of 53%.

Table 1 summarizes the final database that results after merging all the information described above. This panel data set has 25,286,565 observations for 6 years (3.4 million students with one observation per year for the 2012-2016 period). Half of the students have data for their grades in the language subject, and the other half have data for mathematics. Additionally, students are distributed across 9,590 schools, of which 84,719 teachers taught.

Variables of Dataset	**Observations**
Years	6
Grades (by year)	9
Subject (by year)	2
Students' Cohort (by year, grade and subject)	234,135
Total Observations	25,586,565
Unique Students	3,415,010
Schools	9,590
Public Schools	5,270
Subsidized private Schools	3,862
Private Schools	458
Teachers	84,719

Table 1: Summary Statistics for Dataset.

Table 2 presents the descriptive statistics of the variables studied. The mean classroom size is 32.9 students per class, while the mean number of students per school is 233.5, with the distribution between girls and boys being equitable. Moreover, 42% of students are vulnerable, and 41% of the schools are in rural areas. Regarding grades, the mean grade is 5.1 (standard deviation of 0.86), and 6% of students present grade repetition at some point in their schooling life.

As seen in Table 2, teachers present a larger share of women than men (71% vs. 29%), their mean age is 42.9 years old (12.1 years standard deviation), and teaching is the primary function of most of them.

Variable	Mean	SD	Observations
Student Data:			
Class Size	32.94	10.40	25,286,565
Test Score	5.13	0.86	25,284,795
Language	*5.20*	*0.78*	*12,631,865*
Math	*5.06*	*0.92*	*12,652,930*
Public Schools	*5.07*	*0.87*	*9,889,014*
Subsidized Private Schools	*5.11*	*0.84*	*13,426,498*
Private Schools	*5.57*	*0.83*	*1,969,283*
Test Score SIMCE	258.41	53.46	5,672,936
Repeating grade	0.06	0.23	25,286,565
Female	0.50	0.50	25,286,563
Vulnerability Condition	0.42	0.49	25,286,565
Schools Data:			
Grade Size	239,299	13,855	54,115
School Size	233.5	295.8	54,115.0
Urban	0.41	0.49	54,115
Teacher Data:			
Female	0.71	0.45	264,250
Primary Function	0.89	0.31	264,250
Age (years)	42.94	12.13	256,936
Outcome data:			
Students Graduated from High-School			7,988,659
Students Attending Tertiary Education	53.85	49.85	7,988,659
Students Attending Vocational Education	20.08	40.06	7,988,659
Students Attending University	33.77	47.29	7,988,659
Students Attending Public University	10.53	30.70	7,988,659
Students Attending Top-3 University	1.54	12.31	7,988,659

Table 2: Summary Statistics for Sample Used to Value-Added Model.

4 Teachers' Value-Added Estimation

4.1 Conceptual Framework

To estimate the value-added of a teacher, we follow the methodology developed by Chetty et al. (2014a) and (2014b), as well as the contributions of Kane and Staiger (2008), Rothstein (2010) and Kane et al. (2013). Following the literature, we assume that the teacher-student match is random conditional on a set of observables. Our four stages approach ensures that the teacher-student allocations are random, and the value-added estimation is unbiased. We describe our four stages approach below:

First stage: teachers' contribution to students' grades

We observe student i, who during year t is assigned to classroom c from grade g of school sch. For simplicity, let us assume that $c = c(i,t) = c(i,t,sch,g)$. Additionally, in year t, professor j teaches in school sch, in grade g and class c, for the subject s, language and mathematics. To facilitate understanding of the model, we will assume that each professor j teaches in only one grade and only one subject per year, i.e., that each instructor is assigned only to one classroom $(j = j(c(i,t)) = j(c(i,t,sch,g)))$[4].

Consider that the final grade of a student is Q_{it}, which, as noted in the previous sections, ranges from 1.0 to 7.0 for the whole school system. We standardize according to year and grade, such that for each year degree, it has a mean of zero and variance of one, obtaining A^*_{it}.

Following the methodology developed in Chetty et al. (2014a), controlling for the previous grade score and adding a fixed effect per teacher, we estimate the following equation:

$$A^*_{it} = \alpha_j + \gamma A^*_{(it-1)} + \beta X_{it} + \mu_{jt} + \epsilon_{it} \tag{1}$$

Rearranging 1, we verify that:

$$A_{it} = A^*_{it} - \alpha_j + \gamma A^*_{(it-1)} + \beta X_{it} = \mu_{jt} + \epsilon_{it} \tag{2}$$

The above is estimated separately for each subject (language and mathematics) and according to level (primary and secondary), where A^*_{it} corresponds to a standardized test score, α_j the fixed effect per teacher,$A^*_{(it-1)}$ the standardized test score in the immediately preceding grade. These are introduced using a cubic polynomial for the previous language and mathematics grades and interacts with the student's grade level. The vector X_{it} contains information on the student-level characteristics, including gender and grade repetition in the contemporary and previous grades, average attendance in the contemporary grade, and a set of discrete variables by year and grade.

Namely, the residuals of the grades, A_{it}, eliminating the effect of observable characteristics X_{it}, including the previous grade, would be our best predictor of the value-added of teachers, as long as it is fulfilled that $Cov(e_{it}, \hat{\mu}_{jt}) = 0$, a subject that we will deal with later.

It is important to highlight that, in the recovery of the residuals of the previous regression, these incorporate the estimated coefficients of the fixed effect of the teacher, since otherwise, we will be underestimating the impact of the value-added by residualizing the grade, including the effect that the teacher has. Viewed another way, estimating the regression without the inclusion of a fixed teacher effect would overestimate the impact of X_{it} on A^*_{it}, since the correlation between X_{it} and μ_{jt} would be different from zero. To exemplify the above, let us imagine that a student changes schools that coincides with taking classes from a high value-added teacher. In this case, if in our set of observables, we include variables related to the school and do not include a fixed effect per teacher, part of the teacher's improvement could be attributed to the school's characteristics.

[4]The estimates incorporate the cases in which a teacher takes classes for more than one grade in a given school year.

Considering that we obtain unbiased teacher quality estimators, we move on to the methodology's second stage.

Second stage: Predictor of student outcomes.

From the set of values A_{it} associated with each student i in period t, we calculate the average associated with the respective teacher j in year t.

$$\overline{A}_{Jt} = \frac{1}{n}\sum_{i \in j} A_{it} \tag{3}$$

Performing the same procedure for other years, we obtain the vector $A_j^{-t} = (\overline{A}_{j1}, ..., \overline{A}_{jt-1})$, which corresponds to the average of the residues of professor j in different periods of t.

Chetty et al., (2014a) proposed the best linear predictor was $\overline{A}_{Jt}$, based on earlier-than-contemporary information, i.e., a prediction of $E(\overline{A}_{Jt}|A_j^{-t})$, which can be written as:

$$\hat{\mu}_{jt} = \Psi \overline{A}_{jt-1} \tag{4}$$

Where $\Psi = \frac{Cov(A_{jt}, \overline{A}_{jt-1})}{Var(\overline{A}_{jt-1})}$ corresponds to the coefficient that minimizes the sum of the squared errors of the prediction of academic results. The above is obtained from a linear regression between $\overline{A}_{jt}$ and $\overline{A}_{jt-1}$, as presented in equation 4.

The coefficient Ψ, also known as shrinkage, is aimed at correcting the temporal variation that the quality of the teacher may have. Intuitively, if a teacher's results do not change from one year to another, Ψ should have a value close to one. On the other hand, if the results change considerably, this factor will be close to zero; therefore, the extreme values will be taken to zero.

For the estimation of equation 4, we will consider a vector A_j^{-t}, with information for all years except the one for which we are making the prediction. Following Chetty et al., (2014a) we use information from all periods, both before and after, to predict the value-added of period t. The values of the respective covariances between A_{it} and A_j^{-t}, vary according to the subject and level at which we are making the estimate, but in general, these range from 0.69 for $Cov(A_{it}, A_{it-1})$ to 0.46 for $Cov(A_{it}, A_{it-4})$ in the case of language and from 0.31 to 0.18 for mathematics. The covariances in primary education are higher in both cases.

Figures 1a and 1b present the distribution of value-added for each of the teachers. For the case of primary levels, the standard deviation is 0.15 for language and 0.13 for mathematics, while for the case of secondary levels, the standard deviation is 0.24 for language and 0.16 for mathematics. Both results are similar to those obtained by Chetty et al. (2014a) for the case of primary levels but higher for the case of secondary levels.

Having obtained an expression for the value-added of teacher j in period t, $\hat{\mu}_{jt}$, the next step is to check if it is a good predictor of the students' grades.

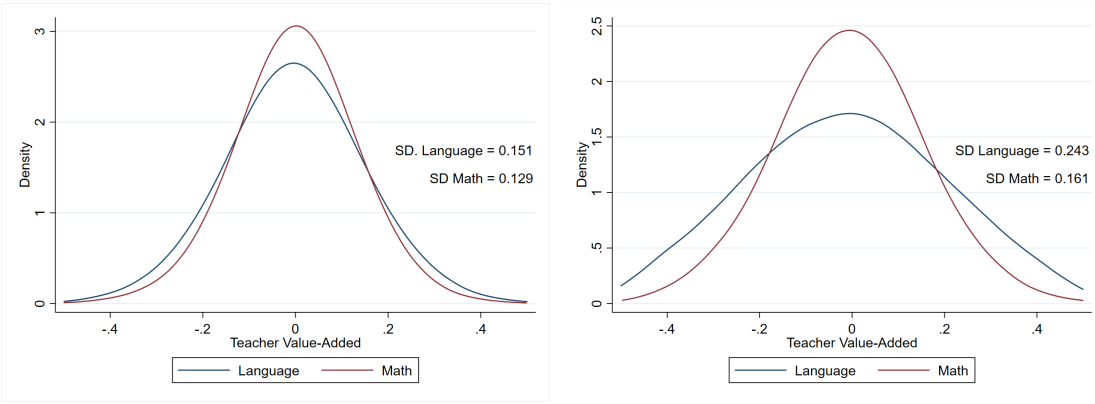
Third stage: Prediction Bias of the Value-Added Estimator.

We could estimate the following regression of the results corrected for observables for period t, A_{it}, with respect to $\hat{\mu}_{jt}$, which was constructed from different information in year t. Therefore, the following equation will correspond to the predictive potential in the student's results for period t but without considering information from this period.

$$A_{it} = \alpha_t + \lambda \hat{\mu}_{jt} + \xi_{it} \tag{5}$$

Where regression of A_{it} in $\hat{\mu}_{jt}$ includes controls by level, subject, and their interaction.

If the students were randomly assigned to period t, we have $E(e_{it}|\hat{\mu}_{jt}) = 0$ from equation 2; thus, the coefficient λ would measure the relationship between the true effect of teacher μ_{jt} and



(a) Primary School **(b)** Secondary School

Figure 1: Distribution of Teachers' Value-Added by Subject

For the case of primary levels, the standard deviation is 0.151 for language and 0.129 for mathematics, while for the case of middle levels, the standard deviation is 0.243 for language and 0.161 for mathematics.

the estimator $\hat{\mu}_{jt}$. Additionally, under random assignment, it is satisfied that $\lambda = \frac{Cov(A_{it},\hat{\mu}_{jt})}{Var(\hat{\mu}_{jt})} = \frac{Cov(\mu_{it},\hat{\mu}_{jt})}{Var(\hat{\mu}_{jt})}$.

Taking the above result, we define the degree of bias of $\hat{\mu}_{jt}$ as $B(\hat{\mu}_{jt}) = \frac{Cov(e_{it},\hat{\mu}_{jt})}{Var(\hat{\mu}_{jt})} = 1 - \lambda$

If $B(\hat{\mu}_{jt}) = 0$, $\hat{\mu}_{jt}$ gives us an unbiased estimator for the prediction of teacher quality, and therefore an improvement of the value-added $\hat{\mu}_{jt}$ has the same causal effect on grades as an improvement of the true value-added μ_{jt}, which is of the same magnitude.

Under the assumption of stationarity of μ_{jt}, from the regression 5, we obtain a coefficient of λ equal to one per construction. The above is corroborated in column 1 of Table 3, which presents a coefficient of 1.004, with a 95% confidence level where the standard errors are calculated considering the cluster at the cohort-school level, to adjust for the correlation that occurs for students in the same classroom and the one associated with multiple observations per student[5]. In the case of Chetty et al. (2014a), their results from their baseline are 0.998.

Figure 2 presents the conditional means of the residual of the grades of year t within quantiles constructed from the prediction of the value-added for period t with information from $t-1$. As expected, considering that our estimator $\hat{\mu}_{jt}$ corresponds to the best linear prediction of $\overline{A}_{Jt}$, we have a practically unitary slope.

Finally, returning to equation 5, we had that only in case the assignment was random between teachers and students would we have the certainty that $Cov(e_{it}, \hat{\mu}_{jt}) = 0$; however, in case there are nonobservable variables that are determining the assignment between teachers and students, we would have that our estimator $\hat{\lambda}$ would be different from one. An indirect way to check this assignment would be to add new variables in our estimation of the residual test score but not to consider them in the construction of our estimator, $\hat{\mu}_{jt}$.

Fourth stage: Selection on excludes observables and no observables.

Let us imagine that observable variables of the student are determining the assignment between students and teachers, i.e., the school follows the same assignment rule for teachers and students

[5]Chetty et al. (2014a) find that clustering according to these variables provides a more conservative and computationally manageable confidence interval than clustering according to classroom and student.

Variables	(1)	(2)	(3)
	Score in t	Score in t	Score in t
Teachers' Value-Added	1.004	0.999	1.019
	(0.000)	(0.000)	(0.000)
Baseline controls	x	x	x
Vulnerability conditions control		x	
Year t-2 test score control			x
Observations	16,758,760	16,758,760	11,750,202
R-squared	0.0687	0.0682	0.078

Table 3: Baseline Model Results and Selection on Exclude Observables.

Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort and p-value in parentheses. The regressions are run on the sample used to estimate the baseline VA model, restricted to observations with a non-missing leave-out teacher VA estimate. There is one observation for each student-subject-grade-school year in all regressions. Teacher VA is scaled in units of student test score standard deviations and is estimated using data from classes taught by the same teacher in other years. Teacher VA is estimated separately for each subject (language and mathematics) and according to level (primary and high school) and using the baseline control vector, which includes at student level: score in the immediately preceding grade, introduced using a cubic polynomial for the previous language and mathematics grades, interacted in turn with discrete variables of the grade in question; gender and grade repetition in the contemporary and previous grade, and average attendance in contemporary grade. Additionally, a set of discrete variables is added per year and per grade. In each columns, the dependent variable is the student's test score in a given year and subject. In Column 2, we add in the estimation of the Teachers' Value-Added the condition of vulnerability of the students, a discrete variable equal to 1 if the student belongs to the most vulnerable 40% of the population. In Column (3) we add in the estimation of the Teachers' Value-Added twice-lagged test scores.

in forecast period t as in previous periods, for example, according to the socioeconomic condition of the student's family or based on lagged test score gains. Adding any of these variables at the time of calculating the residual of the test score should explain an important part of the results and therefore lead to different results when regressing this new residual, $A^{'}_{it}$, in our value-added, $\hat{\mu}_{jt}$.

In this case, we estimate the following equation, a variant of equation 2, adding the observable variable Z_{it}:

$$A^{'}_{it} = A^{*}_{it} - \alpha_j + \gamma A^{*}_{(it-1)} + \beta X_{it} + \rho Z_{it} \tag{6}$$

Later, we regressed $A^{'}_{it}$ in $\hat{\mu}_{jt}$, as in equation 5, again including controls by level, subject, and interaction. The observable variables that we will include separately will be the condition of the vulnerability of the students (1=if the student belongs to the most vulnerable 40% of the population) and the grade after the previous one, A^{*}_{it-2}. This last one is if the ordering between students and teachers is made by the management teams of the establishments from the students' previous results.

The results of estimates of value-added considering equation 6 and then replicating equation 5 are presented in columns 2 and 3 of Table 3 and show that the estimation is practically unaltered. Specifically, for the students' vulnerability condition, the coefficient is 0.999, with a 95% confidence level. For the case of including the score after the previous one, it is 1.019 with a 95% confidence level. This is corroborated in Figures 3a and 3b, which repeats what was done in Figure 2 and

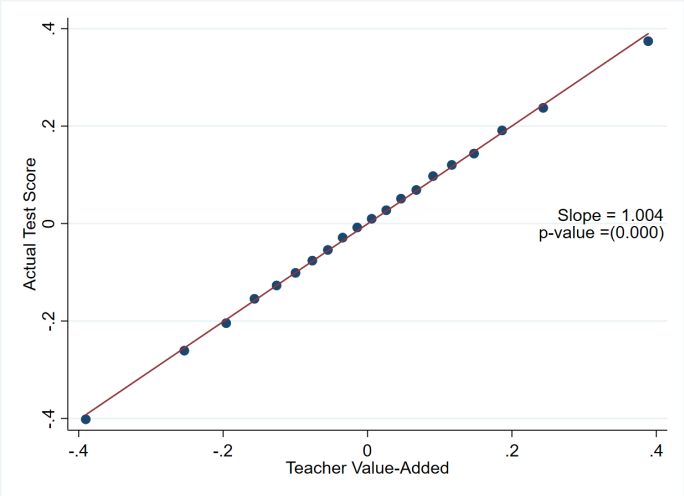


Figure 2: Effect of Teachers' Value-Added on Actual Scores.

This Figure are constructed using the sample used to estimate VA model, 16,758,760 observations. This plot correspond to the regression in column 1 of Table 3. To construct this binned scatter plot, we first residualisze the actual test score with respect to the baseline control vector (detailed in the note of Table 3) separately within each subject and using within-teacher variation to estimate the coefficients. Then divide the VA estimates $\hat{\mu}_{jt}$ into twenty equal-sized groups (vingtiles) and plot the means of the actual test score residuals within each bin against the mean value of $\hat{\mu}_{jt}$ within each bin. The line shows the best linear fit estimated on the underlying micro data using OLS

shows an adjustment practically equal to our base model.

We can conclude that selection due to the students' socioeconomic status is quite marginal, mainly because the student comes from a nonvulnerable family. Let us assume that this means the student regularly obtains good results and will present good results in the contemporary grade and the previous ones, which is largely captured by including the grade of the immediately previous one for control.

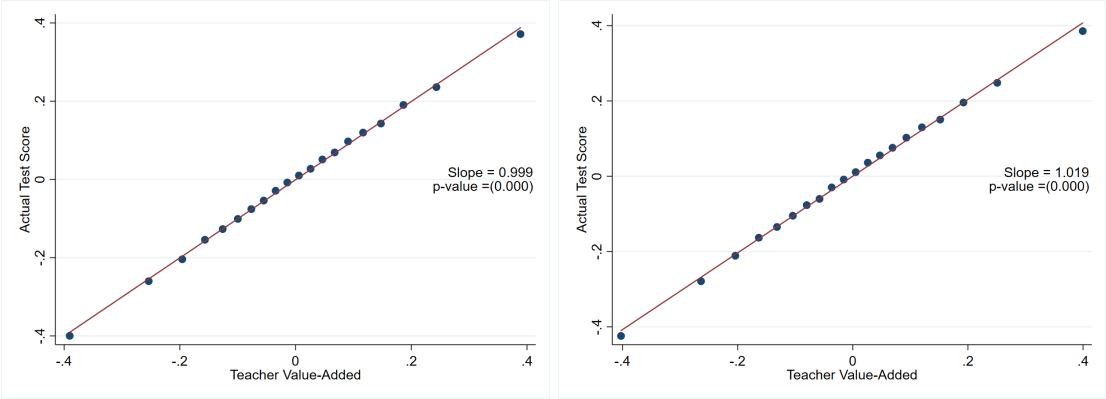
In the case of the impact that selection from previous results would have, our explanation again points in the same direction. Much of the variation observed in the grades after previous ones are captured by the set of controls that we include in our regression when estimating the grade residue.

The results presented are consistent with those found by Chetty et al. (2014a), in the sense that, in the case of parental characteristics, in cases measured mainly through household income, it finds values for $\hat{\lambda}$ of 0.996 (and for the case of including the score after previous, 0.976).

The above does not rule out the possibility that students are sorted to teachers based on unobservable characteristics orthogonal to the Z_{it} variables. For this purpose, we replicate the quasi-experiment realized by Chetty et al. (2014a), after considering the impossibility of conducting a random experiment as Kane and Staiger (2008) and Kane et al. (2013).

This quasi-experiment exploits teacher turnover between schools and classes from one year to the next. A good way to understand this method's idea is to exemplify it, such as Chetty et al. (2014a), considering a school with three 8th grade classrooms (the last grade in the primary level). Suppose one of the teachers leaves the school in 2012 and is replaced by a teacher whose VA estimate in mathematics is 0.3 higher. Assume that the distribution of unobserved determinants of scores e_{it} does not change between 2011 and 2012. If forecast bias $B = 0$, this change in teaching staff should raise average 8th-grade math scores in the school by $0.3/3 = 0.1$. More generally, we can estimate B by comparing the change in mean scores across cohorts to the change in mean VA driven by teacher turnover, provided that student quality is stable over time.

We estimate the degree of forecast bias B by regressing changes in mean test scores across



(a) Added includes Two Lags of Test Score **(b)** Added includes Vulnerability Conditions

Figure 3: Effect of Teachers' VA on Actual Scores

These Figures are constructed using the same procedure explained in note of Figure 2. The two panels are binned scatter plots of actual test scores vs teacher VA including condition of the vulnerability of the students for his estimation (panel a) and including the twice-lagged test scores (panel b).

cohorts on changes in mean teacher VA:

$$\Delta A_{sch,gt} = a + b\Delta Q_{sch,gt} + \Delta\chi_{sch,gt} \tag{7}$$

Where A_{sgt} denote the mean value of A_{it} for students in school sch in grade g in year t and define the change in mean residual scores as $\Delta A_{sch,gt} = A_{sch,gt} - A_{sch,gt-1}Q_{sch,gt}$ denote the (student-weighted) mean of $\hat{\mu}_{jt}^{-(t,t-1)}$ across teachers in school sch in grade g. We define the change in mean teacher value-added from year $t-1$ to year t in grade g in school sch as $\Delta Q_{sch,gt} = Q_{sch,gt} - Q_{sch,gt-1}$.

The coefficient b in equation 7 identifies the degree of forecast bias as defined in equation 5 under the following identification assumption that changes in teacher VA across cohorts within a school grade are orthogonal to changes in other determinants of student scores, $Cov(\Delta Q_{sch,gt}, \Delta\chi_{sch,gt}) = 0$.

In general, student sorting at an annual frequency is minimal because of the costs of changing schools. Considering that families would be unlikely to change their school simply because a single teacher leaves or enters a given grade, we believe in this assumption's plausibility.

Table 4 presents the results of the quasi-experiment detailed previously. We estimate six alternative value-added models, reporting correlations with baseline value-added estimates in column 1, and forecast bias for each model, defined as $B = 1 - \lambda$, in column 2. In this case, the first row presents the baseline model results, which estimates a forecast bias of 5.4%. These results are similar to those obtained by Chetty et al. (2014a) for the specification that includes only a lagged test score of 4.8%. In the second row, following the same three stages detailed above, we use all the baseline controls but omit teacher fixed effects and obtain a forecast bias of 0.3%. The next row adds information on lagged cross-subjects through a cubic polynomial interacting with the grade, marginally increasing the forecast bias by 6.8%. In row 4, we replicate our baseline specification by adding information on the same variables as our baseline but considering the averages of the classes and the school; in this case, the bias remains in ranges similar to our baseline model, 6.6%. Row 5 removes all controls related to the test score from the baseline specification, leaving only non-score controls at the student level; in this case, we confirm how relevant it is to include these controls in the estimate to obtain unbiased forecast estimates since the bias increases to 40.1%. The last row drops all controls except grades and year fixed effects, showing a forecast bias of 46.5%.

Specification	Correlation with Value-Added Baseline Estimate	Quasi-Experimental Estimate Bias %
(1) Value-Added Baseline	1.000	5.4%
		(0.000)
(2) VA Baseline without Teacher Fixed Effect	0,997	0.3%
		(0.000)
(3) VA Baseline and student's lagged score in other subject	0,992	6.8%
		(0.000)
(4) VA Baseline and Class and School scores	0,958	6.6%
		(0.000)
(5) Non-sore Controls	0,868	40.1%
		(0.000)
(6) No controls	0,788	46.5%
		(0.000)

Table 4: Comparison of Forecast Bias Across Value-Added Models.

In this table we estimate six alternative Value-Added models reporting correlations with the baseline Value-Added estimates in column 1. In column 2, we report quasi-experimental estimates of forecast bias for each model, defined as 1 minus the coefficient in a regression of the cross-cohort change in scores on the cross-cohort change in mean teacher VA. The regressions are run on the sample used to estimate the baseline VA model, restricted to observations with a non-missing leave-out teacher VA estimate. All models are estimated separately by school level and subject; the correlations and estimates of forecast bias pool VA estimates across all groups. Each model only varies the control vector used to estimate student test score residuals in equation 2; the remaining steps of the procedure used to construct VA estimates are the same for all the models. Model 1, baseline model, includes at student level: score in the immediately preceding grade, introduced using a cubic polynomial for the previous language and mathematics grades, interacted in turn with discrete variables of the grade in question; gender and grade repetition in the contemporary and previous grade, and average attendance in contemporary grade. Additionally, a set of discrete variables is added per year and per grade. Model 2 uses all of the baseline controls but omits teacher fixed effects. Model 3 uses all of the baseline controls adding a cubic in lagged cross-subject scores (for the case of language we adding mathematics and vice versa), interacted with the student's grade level. Model 4 uses all of the baseline controls and includes the same variable at student level at class and school level (mean of the student variables). Model 5 removes all controls related to test scores from the baseline specification, leaving only non-score controls at the student level. Finally, Model 6 drops all controls except grade and year fixed effects.

Two conclusions are obtained from this last exercise. First, although our second specification exhibits a lesser bias than our baseline model, it is because this method exploits variation both within and across teachers to identify the coefficients on the control vector and thus can understate teacher effects by overattributing test score growth to covariates if there is sorting, which is why we consider the specification including teacher fixed effects as our baseline model. Second, we corroborated that controlling for prior student-level test scores is fundamental to obtaining unbiased value-added estimates. As explained by Chetty et al. (2014a), one potential explanation for this result is that classroom assignment in large schools is made primarily on the basis of prior-year test performance and its correlates.

Reviewing the robustness of our results, conditional on the available observable variables and considering the quasi-experiment's results, we have that even when there are no random assignments to teachers in forecast year t, by including prior student-level test scores, the estimate exhibits minimal predictive error.

4.2 Impact of Value-added in Educational Outcomes

This section estimates the impact of value-added on student outcome variables once they leave school. Following the previous notation, let us consider that Y_{it}^* corresponds to the tertiary education attendance (or university attendance or vocational education) of student i during their first year after graduation. We are interested in measuring the impact of value-added on this outcome variable. The following linear specification is proposed.

$$Y_{it}^* = \alpha + \tau m_{jt} + \varepsilon_{it} \tag{8}$$

Where the variable m_{jt} corresponds to $m_{jt} = \mu_{jt}/\sigma_\mu$, normalized teacher value-added j, such that the τ coefficient of equation 8 represents the reduced form of the impact of an increase of one standard deviation of teacher value-added for a given year, or grade, on tertiary education attendance. For more detail on the formalization of this reduced parameter's interpretation, see Appendix A of Chetty et al. (2014b).

It should be noted that τ will correspond to the value-added impact, measured through the students' grades, on their future university attendance. In other words, a teacher may affect students' university attendance in ways other than those associated with their score, such as their confidence or aptitude when applying for university.

Assuming that the value-added is unbiased, as detailed in section 3.5.1, we corroborate $\frac{Cov(\mu_{jt},\hat{\mu}_{jt})}{Var(\hat{\mu}_{jt})} = \frac{Cov(m_{jt},\hat{m}_{jt})}{Var(\hat{m}_{jt})}$, and we can verify from equation 8 that:

$$\frac{Cov(Y_{it},\hat{m}_{jt})}{Var(\hat{m}_{jt})} = \tau\frac{Cov(m_{it},\hat{m}_{jt})}{Var(\hat{m}_{jt})} + \frac{Cov(\varepsilon_{it},\hat{m}_{jt})}{Var(\hat{m}_{jt})} = \tau + \frac{Cov(\varepsilon_{it},\hat{m}_{jt})}{Var(\hat{m}_{jt})} \tag{9}$$

provided that unobserved determinants of attendance, ε_{it}, are orthogonal to teacher VA estimates $\hat{m}_{jt}$.

For estimation, and similar to the calculation of the value-added, we residualize the dependent variable, Y_{it}^*, including the baseline covariates and teacher fixed effect:

$$Y_{it} = Y_{it}^* - \hat{\beta}^Y X_{it} \tag{10}$$

Note that again, in recovering the residual from the previous equation, these must incorporate the teacher fixed effect. Otherwise, we will underestimate the value-added's impact by not incorporating the teacher's effect on the respective dependent variable.

Finally, we estimate the following regression:

$$Y_{it} = \alpha + \tau\hat{m}_{jt} + \varepsilon_{it} \tag{11}$$

The interpretation of $\hat{\tau}$ will be the effect that an increase of one standard deviation in teacher quality, as measured by grades, has on their graduating students' university attendance. This can be considered a direct impact of the variation of the value-added on the probability of university attendance since our results of $\hat{\tau}$ are close to the ($\hat{\tau} = 1.004$).

Table 5 presents the results for the estimates of value-added through equation 11 for the entire educational system in the rates of tertiary, vocational, university, and top-3 university attendance and their respective robustness as detailed below.

The first column of each educational outcome specification is the result of equation 11 using our baseline controls. For the case of tertiary education attendance, the effect is that a 1 SD increase in a teacher's true VA test score in a single grade increases the probability of tertiary education attendance by 1.9 percentage points relative to a mean tertiary education attendance rate of 54.4%, which means an increase of 3.5% in mean tertiary education attendance in the regression sample. The null hypothesis that teacher VA does not affect tertiary education attendance is rejected with a $p - value < 0.001$.

For the case of vocational education attendance, the effect is that a 1 SD increase in a teacher's true VA test score in a single grade diminishes the probability of vocational education attendance by 1.5 percentage points relative to a mean vocational education attendance rate of 20.2%. For university attendance, the effect in a single grade is 3.4 percentage points, relative to a mean university attendance rate of 34.2%, i.e., an increase of 9.9% for mean university attendance in the regression sample almost triples if we consider any tertiary education institution type in column 1. Finally, column 10 presents the result for top-3 universities, where a 1 SD increase in a teacher's true VA test score in a single grade increases the probability of attendance at these universities by 0.5 percentage points relative to a mean top-3 university attendance rate of 1.9%, implying a 25.7% mean for top-3 university attendance in the regression sample. In all the above cases, the null hypothesis that teacher VA does not affect tertiary education attendance is rejected with a $p - value < 0.001$.

Taking the previous results, we can infer that the effect is a 1 SD increase in a teacher's true VA test score in a single grade has a general positive effect. However, the composition of this is not, affecting almost 3 times more than the effect on the average of the dependent variable in the case of universities. This effect is 2.5 times higher within this group if we consider top-3 universities, leading us to infer that the higher the institution's quality, the more important the professor's quality is in determining a student's enrollment.

We evaluate this estimate's robustness to alternative control vectors in the second and third columns of each educational outcome specification, replicate the specification with the baseline control vector, and add the student's vulnerability condition and twice-lagged test scores, respectively (equation 6). Again, the coefficient does not change appreciably. Both variables are strong predictors of tertiary education attendance rates even conditioned on the baseline controls, with a $p - value < 0.001$. Hence, despite controlling for these variables, they do not significantly affect the estimates of τ, supporting the identification assumption of selection on observables.

Figures 5 to 4d plot the residual of each educational attendance rate for students in school year t against $\hat{m}_{jt}$. To construct this binned scatter plot, a nonparametric representation of the conditional expectation function is used. We divide the VA estimates $\hat{m}_{jt}$ into twenty equal-sized groups (vingtiles) and plot the mean of the attendance residuals in each bin against the mean of $\hat{m}_{jt}$ in each bin. Finally, we add back the mean attendance rate in the estimation sample to facilitate the scale's interpretation. The regression coefficient and standard error are reported in this, and all subsequent figures are estimated on the class-level, with standard errors clustered by school cohort.

As the objective of this research is to study the effects of two measures on educational variables, and one of them, teacher evaluation, is applied only to public schools, Table 6 replicates the previous results but only to this subsample of teachers.

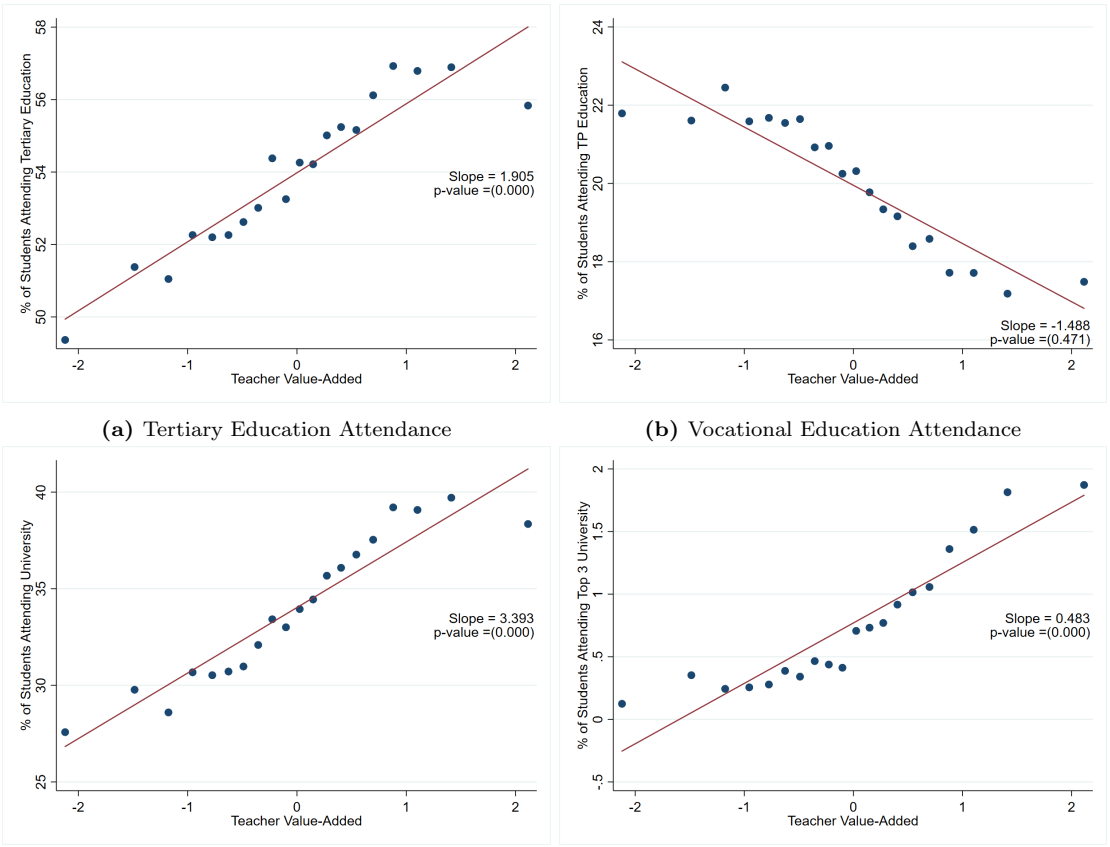
The results are in the same direction, although more moderate, than those obtained for the entire educational system. Specifically, the effect is that a 1 SD increase in a teacher's true VA

Variable	(1)	(2)	(3)	(4)	(5)	(6)
	Tertiary Education Attendance			Vocational Education Attendance		
Teachers' Value-Added	1.905	1.829	1.761	-1.488	-1.471	-1.604
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mean Dep. Var.	54.35	54.35	54.58	20.16	20.16	20.14
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions		x			x	
Year t-2 Test Score			x			x
Observations	5,331,323	5,331,323	3,583,633	5,331,323	5,331,323	3,583,633
VA Over Mean Dep. Var	3.5%	3.4%	3.2%	-7.4%	-7.3%	-8.0%

Variable	(7)	(8)	(9)	(10)	(11)	(12)
	University Attendance			Top-3 University Attendance		
Teachers' Value-Added	3.393	3.300	3.365	0.483	0.480	0.540
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mean Dep. Var.	34.19	34.19	34.44	1.88	1.88	2.269
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions		x			x	
Year t-2 Test Score			x			x
Observations	5,331,323	5,331,323	3,583,633	5,331,323	5,331,323	3,583,633
VA Over Mean Dep. Var	9.9%	9.7%	9.8%	25.7%	25.6%	23.8%

Table 5: Teachers' Value-Added Outcomes.

Each column reports coefficients from an OLS regression between the residual of dependent variable and Teacher VA using the baseline control vector, with standard errors clustered by school-cohort and p-value in parentheses. Columns 1-3 use an indicator for tertiary education attendance; columns 4-6 use an indicator for vocational education attendance; columns 7-9 use an indicator for university attendance; columns 10-12 use an indicator for Top-3 university attendance. In the first column of each outcome-specification, we residualize each dependent variable using the baseline control vector detailed in note of Table 3. In the second column of each outcome-specification, we use the baseline control vector adding the condition of vulnerability of the students, a discrete variable equal to 1 if the student belongs to the most vulnerable 40% of the population. In the third column of each outcome-specification, we use the baseline control vector adding the twice-lagged test scores. The last row of each table corresponds to the ratio between the impact of the Teacher's Value-Added on the average of the dependent variable.



(a) Tertiary Education Attendance

(b) Vocational Education Attendance

(c) University Attendance

(d) Top-3 Universities Attendance

Figure 4: Conditional mean of Teachers' VA on Educational Outcomes

Panels (a) to (d) are binned scatter plots of tertiary education attendance rates, vocational education attendance rates, university attendance rates and top-3 university attendance rates vs. normalized teacher VA $\hat{m}_{jt}$. These plots correspond to the regressions in the first column of each outcome-specification of Table 5 and use the same sample restrictions and variable definitions, considering 15,568,432 observations. To construct these binned scatter plots, we first residualize the dependent variable with respect to the baseline control vector separately within each subject, using within-teacher variation to estimate the coefficients. We then divide the VA estimates $\hat{m}_{jt}$ into twenty equal-sized groups (vingtiles) and plot the means of the dependent variable residuals within each bin against the mean value of $\hat{m}_{jt}$ within each bin. Finally, we add back the unconditional mean of the dependent variable in the estimation sample to facilitate interpretation of the scale. The solid line shows the best linear fit estimated on the underlying micro data using OLS.

Variable	(1)	(2)	(3)	(4)	(5)	(6)
	Tertiary Education Attendance			Vocational Education Attendance		
Teachers' Value-Added	0.626	0.603	0.587	-0.222	-0.217	-0.419
	(0.008)	(0.010)	(0.021)	(0.181)	(0.189)	(0.028)
Mean Dep. Var.	47.12	47.12	47.79	22.98	22.98	22.99
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions		x			x	
Year t-2 Test Score			x			x
Observations	1,815,678	1,815,678	1,216,653	1,815,678	1,815,678	1,216,653
VA Over Mean Dep. Var	1.3%	1.3%	1.2%	-1.0%	-0.9%	-1.8%

Variable	(7)	(8)	(9)	(10)	(11)	(12)
	University Attendance			Top-3 University Attendance		
Teachers' Value-Added	0.847	0.819	1.006	0.116	0.116	0.144
	(0.019)	(0.021)	(0.010)	(0.134)	(0.136)	(0.142)
Mean Dep. Var.	24.15	24.15	24.80	1.24	1.24	1.55
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions		x			x	
Year t-2 Test Score			x			x
Observations	1,815,678	1,815,678	1,216,653	1,815,678	1,815,678	1,216,653
VA Over Mean Dep. Var	3.5%	3.4%	4.1%	9.4%	9.4%	9.3%

Table 6: Teachers' Value-Added Outcomes in Public Schools.

Each column reports coefficients from an OLS regression between the residual of dependent variable and Teacher VA using the baseline control vector for public schools, with standard errors clustered by school-cohort and p-value in parentheses. Columns 1-3 use an indicator for tertiary education attendance of students graduated from public schools; columns 4-6 use an indicator for vocational education attendance; columns 7-9 use an indicator for university attendance; columns 10-12 use an indicator for Top-3 university attendance. In the first column of each outcome-specification, we residualize each dependent variable using the baseline control vector detailed in note of Table 3. In the second column of each outcome-specification, we use the baseline control vector adding the condition of vulnerability of the students, a discrete variable equal to 1 if the student belongs to the most vulnerable 40% of the population. In the third column of each outcome-specification, we use the baseline control vector adding the twice-lagged test scores. The last row of each table corresponds to the ratio between the impact of the Teacher's Value-Added on the average of the dependent variable for students graduated from public schools.

test score in a single grade increases the probability of tertiary education attendance of graduate students who attended public schools by 0.6 percentage points relative to a mean tertiary education attendance rate of 47.1%, which means an increase of 1.3% in mean tertiary education attendance in the regression sample. In the case of vocational education attendance, the effect is not nonzero ($p-value < 0.181$), as is the case for top-3 universities ($p-value < 0.134$). University attendance increases by 0.85 percentage points if teachers' true VA test score in a single grade increases by 1 SD, that is, an increase of 3.5% in mean university attendance.

Figures 9a to 9d in Appendix plot the residual of each educational attendance rate for graduate students who attended public schools in year t against $\hat{m}_{jt}$, as explained above.

To contextualize our results, moving a student from a teacher in the fifth to the ninety-fifth percentile of the true VA test score distribution would lead to increases the probability of tertiary education attendance of graduate students who attended public schools by 2.21 percentage points in a single grade[6]. Figure 5 illustrates the estimated effects of moving a student from a teacher in the fifth to the twenty-fourth (+1 SD), sixtieth (+2 SD) and ninety-fifth (+3.5 SD) percentile of the true VA test score distribution, respectively, on the probability of tertiary education attendance.

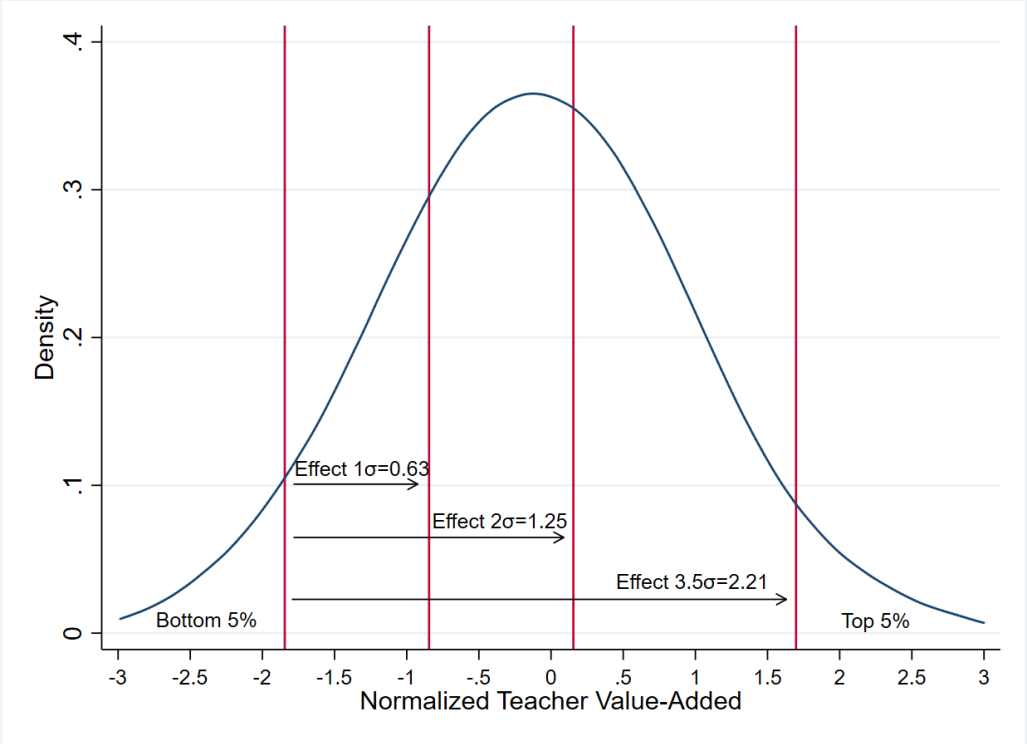


Figure 5: Effects of moving a student from Teachers' Value-Added in the fifth percentile of the true VA test score distribution in public schools.

Distribution of normalized Teachers' Value-Added for language in public schools . The arrows in the figure correspond to the effect of moving to a student from a teacher in the fifth to the twenty-fourth, sixtieth and ninety-fifth percentile of the true VA test score distribution. Its calculation corresponds to $1 \times \tau = 0.626$, $2 \times \tau = 1.25$, and $3.54 \times \tau = 2.21$, respectively.

[6] A teacher who is at the ninety-fifth percentile is 3.5 standard deviation better than one at the fifth percentile. Therefore, the effect is 3.54 x τ, i.e., 3.54 x 0.626 = 2.21

5 Methodology for Estimating the Impact of Teacher Evaluation

To identify causal effects, the assumption of no unmeasured confounders must hold. This means that unobservable variables should be uncorrelated with the treatment variable and the variable of interest (potential outcome) (Rubin, Stuart and Zanutto, 2004). In our case, the assumption would imply that there are no unmeasured variables that affect the treatment (i.e., teaching evaluation) and potential outcome (i.e., student performance). Unfortunately, we do not have a randomized experiment to ensure the assumption just mentioned. Thus, as observational data are used in our study, several potential sortings between and within schools might complicate our analysis as they challenge the no unmeasured confounders assumption, as has been explained during this investigation.

All potential selection sources (student-school selection, teacher-school selection, and student-teacher selection) may generate non-random student-teacher sortings. In fact, there are several previous pieces of empirical evidence showing that student-teacher sorting (between and within schools) is not random (see Jackson, 2014 and Rothstein, 2010, among others).

To address these concerns, we follow two approaches. The first of these, based on previous research (Kane et al., 2008, Kane et al., 2011, Briole and Maurin, 2019, among others), attempts to quantify teacher evaluation impacts on certain outcomes through a linear estimation controlling for the potential biases mentioned above. For the second approximation, we will use a strategy similar to that used to estimate the impact of value-added on educational outcomes developed in section 3.5.1.

5.1 First Approach for Teacher Evaluation

For the first approach, we estimate the following equation:

$$Y_{it}^{*} = \alpha + \gamma A_{it-1}^{*} + \pi X_{it} + \theta ED_{jt} + \epsilon_{ijt} \tag{12}$$

When Y_{it}^{*} represents the tertiary educational attendance of graduating student i at time t. Similar to value-added estimates on educational outcomes, we include A_{it-1}^{*}, which corresponds to standardized test scores in the immediately preceding grade, and is introduced using a cubic polynomial for the previous language and mathematics grades, and interacts with the student's grade level. The vector X_{it} contains the information for student-level characteristics, including gender and grade repetition in the contemporary and previous grades, average attendance in the contemporary grade, and a set of discrete variables by year and grade. ED_{jt} is the normalized teacher evaluation, wich is a continuous variable that takes values between 1 and 4 depending on the teacher evaluation result. The error term is ϵ_{ijt}. Given this model, we are interested in the value of the coefficient θ, which, in the case of correct identification, should capture the impact of an improvement in teacher overall test score on student achievement.

In our estimate, we do not add a teacher fixed effect since this effect is expressed directly from the coefficient θ, which accompanies the variable ED. Incorporating a teacher fixed effect would underestimate the ED's effect since part of the teacher's effectiveness, not captured through the TET, would be absorbed through this effect.

Similarly, we do not include school or student fixed effects. Although the inclusion of school fixed effects is appealing because of its potential to reduce bias in teacher value-added estimates, the best practices to date suggest that their practical importance is limited (Chetty et al., 2014a and Koedel et al., 2015). In particular, Koedel et al. (2015) point out that once students' lagged performance has been included, adding these layers of fixed effects narrows the identifying variation used to estimate teacher value-added, which can increase the imprecision in estimation. Further evidence supports this view; Chetty et al. (2014a) show that value-added models with students‘ lagged test

scores but without school and student fixed effects produce teacher value-added estimates with no significant bias. Similarly, Kane et al. (2008) and Kane et al. (2013) show that teacher value-added models without school and student fixed effects perform well versus experimental studies[7]. To address within-school sorting, we control the classroom averages of student characteristics (as suggested by Altonji and Mansfield, 2014).

Table 7 presents the results of our first methodology to evaluate the impact of teacher evaluation on the rates of tertiary, vocational, university, and top-3 university attendance estimates through equation 12.

The first column of each educational outcome specification uses baseline controls used in value-added estimates. The second column adds controls associated with the vulnerability condition of students and twice-lagged test scores. The third column adds class and school controls through cubic in-class means of prior-year test scores in each subject (language and mathematics). Each interacted with the grade and class means of all the other individual covariates.

For tertiary education attendance, a 1 SD increase in a teacher's evaluation in a single grade increases the probability of attendance of graduate students who attended public schools by 1.25-1.67 percentage points, depending on the specification, relative to a mean tertiary education attendance rate, which means an increase between 2.7% and 3.5% of mean tertiary education attendance in the regression sample.

In the case of vocational education, a 1 SD increase in a teacher's evaluation in a single grade decreases the probability of vocational education attendance of graduate students who attended public schools by 0.21-0.42 percentage points, depending on the specification, relative to a mean vocational education attendance rate, which means a decrease between 1.8% and 0.9% of mean attendance in the regression sample.

For the university analysis, we found that a 1 SD increase in a teacher's evaluation in a single grade increased the probability of university education attendance of graduate students who attended public schools by 1.46-2.09 percentage points, depending on the specification, which means an increase between 6.0% and 8.6% in mean university attendance in the regression sample. If we focus on top-3 universities, the effect is between 0.22 and 0.26 percentage points, depending on the specification, i.e., an increase between 17.8% and 18.5% of mean attendance in the regression sample.

In all the above cases, the null hypothesis that teacher VA does not affect tertiary education attendance is rejected with a $p - value < 0.001$, except for the specification in column 6, which presents a $p - value = 0.071$.

Similar to the results for the case of value-added, we observed a positive impact, at a general level, of a 1 SD increase in a teacher's evaluation in a single grade of 3.5% of mean attendance in the regression sample; however, this effect is heterogeneous depending on the type of tertiary education institution. From the above, we can infer that the better the quality of the institutions in which graduate students enroll, the greater the influence that a good teacher, measured through the TET, can have on those results.

As detailed in sections 3.3.3 and 3.3.4, the variable used in equation 12, ED_{jt}, corresponds to a weighted total score based on 4 instruments. We are interested in verifying the impact of each instrument separately on attendance to tertiary educational attendance. To do this, we estimate an alternative presentation of equation 12:

$$Y_{it}^{*} = \alpha + \gamma A_{it-1}^{*} + \pi X_{it} + \sum_{v=1}^{4} \rho_v Instrument_{vjt} + \varepsilon_{ijt} \tag{13}$$

where we include the four instruments of teacher evaluation in the same equation: self-evaluation, peer interview, external references, and portfolio.

[7] The intuition of this is that past test scores act as proxies for the unobserved heterogeneity; hence, the inclusion of student fixed effects is not particularly useful.

Variable	(1)	(2)	(3)	(4)	(5)	(6)
	Tertiary Education Attendance			Vocational Education Attendance		
Weighted Total Score	1.670	1.367	1.253	-0.418	-0.457	-0.212
	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.071)
Mean Dep. Var.	47.27	47.92	47.27	22.97	22.97	22.97
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions and Year t-2 Test Score		x			x	
Class and School Controls			x			x
Observations	1,805,012	1,165,653	1,805,012	1,805,012	1,165,653	1,805,012
Over Mean Dep. Var	3.5%	2.9%	2.7%	-1.8%	-2.0%	-0.9%

Variable	(7)	(8)	(9)	(19)	(20)	(21)
	University Attendance			Top-3 University Attendance		
Weighted Total Score	2.088	1.824	1.464	0.225	0.235	0.217
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mean Dep. Var.	24.30	24.95	24.30	1.22	1.57	1.22
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions and Year t-2 Test Score		x			x	
Class and School Controls			x			x
Observations	1,805,012	1,165,653	1,805,012	1,805,012	1,165,653	1,805,012
Over Mean Dep. Var	8.6%	7.3%	6.0%	18.5%	15.0%	17.8%

Table 7: Teachers' Evaluation Test Outcomes - Lineal Regression.

Each column reports coefficients from an OLS regression between the dependent variable and TET, with standard errors clustered by school-cohort and p-value in parentheses. Columns 1-3 use an indicator for tertiary education attendance of students graduated from public schools; columns 4-6 use an indicator for vocational education attendance; columns 7-9 use an indicator for university attendance; columns 10-12 use an indicator for Top-3 university attendance. In the first column of each outcome-specification, we use the baseline control vector detailed in note of Table 3. In the second column of each outcome-specification, we use the baseline control vector adding the condition of vulnerability of the students, a discrete variable equal to 1 if the student belongs to the most vulnerable 40% of the population. In the third column of each outcome-specification, we use the baseline control vector adding the twice-lagged test scores. The last row of each table corresponds to the ratio between the impact of the Teacher Evaluation on the average of the dependent variable for students graduated from public schools.

From regression, we want to know which instrument explains in a better way the probability of tertiary educational attendance. It should be noted that the weighted total score corresponds to a weighted average in a different way for each instrument; however, in these regressions, we include the normalized score by year and grade of each instrument, not considering the respective weights used in the weighted total score of equation 12.

Again, the identification assumptions are the same as previously considered, the covariance between each instruments and ε_{ijt} is zero, i.e., $Cov(\varepsilon_{ijt}, Instrument_{vjt})$.

Tables 8 and 9 present these results. As in the previous cases, for each educational outcome specification, two columns are added with the robustness associated with the inclusion of the observable variables: second column add vulnerability condition of students and twice-lagged test scores and third column add class and school controls. In all cases, we highlight that the instruments that explain the probability of tertiary educational attendances are the portfolio and external references. In the case of Portfolio, a 1 SD increase in a teacher's instrument Portfolio in a single grade increases the probability of attendance of graduate students who attended public schools by 1.25-1.52 percentage points depending on the specification relative to a mean tertiary educational attendance rate, which means an increase between 3.2% and 2.6% of mean attendance in the regression sample. For External References, a 1 SD increase in a teacher's instrument in a single grade increases the probability of attendance of graduate students who attended public schools by 1.06-1.54 percentage points depending on the specification relative to a mean attendance rate, which means an increase between 3.2% and 2.2% of mean attendance in the regression sample. Finally, for the case of the Self-evaluation score, the results, although positive, are marginal, while for the Peer Interview score in all specifications, the results are negative.

5.2 Second Approach for Teacher Evaluation

For the second approach, we residualize the dependent variable, Y_{it}^{*}, in this case, the tertiary educational attendance, including baseline controls and teacher fixed effect, obtained Y_{it}, where Y_{it} isolates the impact that a certain teacher has on tertiary educational attendance, as explained in the case of value-added estimates.

Then, we estimate the linear specification:

$$Y_{it} = \alpha + \rho ED_{jt} + \omega ijt \tag{14}$$

Where the variable ED_{jt} corresponds to normalized teacher evaluation j and represents the reduced form of the impact of an increase of one standard deviation of teacher evaluation for a given year, or grade, on tertiary educational attendance.

It should be noted that ρ will correspond to the teacher evaluation impact, measured through the students' grades, on the students' future tertiary educational attendance. In other words, a teacher may affect students' attendance in ways other than those associated with their grades, such as their confidence or aptitude when applying for university.

Our second methodology for evaluating the impact of teacher evaluation on the rates of the tertiary, vocational, university, and top-3 university attendance estimates is presented in Table 10.

We observe stability in the results, which is slightly lower than those observed in our first methodology used in the case of teacher evaluation (Table 7). Specifically, we see a 1 SD increase in a teacher's evaluation in a single grade increases the probability of tertiary education attendance of graduate students who attended public schools by 1.63-1.91 percentage points (an increase between 4.0% and 3.4% of mean attendance in the regression sample), decreases the probability of vocational education attendance by 0.33-0.41 percentage points (a decrease between 1.4% and 1.8% of mean attendance in the regression sample), increases the probability of university education by 2.24-2.04 percentage points (an increase between 9.2% and 8.2% of mean attendance in the regression

Variable	(1)	(2)	(3)	(4)	(5)	(6)
	Tertiary Education Attendance			Vocational Education Attendance		
Portfolio Score	1.517	1.048	1.248	-0.387	-0.397	-0.269
	(0.000)	(0.000)	(0.000)	(0.001)	(0.006)	(0.021)
Self Evaluation Score	0.410	0.236	0.256	-0.019	-0.037	0.047
	(0.007)	(0.176)	(0.075)	(0.855)	(0.776)	(0.643)
Peer Interview Score	-0.441	-0.237	-0.375	0.472	0.402	0.450
	(0.013)	(0.204)	(0.020)	(0.000)	(0.007)	(0.000)
External References Score	1.536	1.498	1.061	-1.028	-0.959	-0.782
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions and Year t-2 Test Score		x			x	
Class and School Controls			x			x
Mean Dep. Var. (Public School)	47.27	47.92	47.27	22.97	22.97	22.97
Observations	1,805,012	1,165,653	1,805,012	1,805,012	1,165,653	1,805,012
Portfolio Score	3.2%	2.2%	2.6%	17.0%	14.7%	15.3%
Self Evaluation Score	0.9%	0.5%	0.5%	-0.7%	-1.1%	-0.9%
Peer Interview Score	-0.9%	-0.5%	-0.8%	-5.9%	-5.2%	-2.9%
External References Score	3.2%	3.1%	2.2%	21.8%	17.2%	19.3%

Table 8: Instruments of Teacher Evaluation on Outcomes - Linear Regression.

Each column reports coefficients from an OLS regression between the dependent variable and TET instruments (portfolio, self evaluation, peer interview and external references), with standard errors clustered by school-cohort and p-value in parentheses. Columns 1-3 use an indicator for tertiary education attendance of students graduated from public schools; columns 4-6 use an indicator for vocational education attendance. In the first column of each outcome-specification, we use the baseline control vector detailed in note of Table 3. In the second column of each outcome-specification, we use the baseline control vector adding the condition of vulnerability of the students, a discrete variable equal to 1 if the student belongs to the most vulnerable 40% of the population. In the third column of each outcome-specification, we use the baseline control vector adding the twice-lagged test scores. The last row of each table corresponds to the ratio between the impact of the Teacher Evaluation on the average of the dependent variable for students graduated from public schools.

Variable	(7)	(8)	(9)	(10)	(11)	(12)
	University Attendance			Top-3 University Attendance		
Portfolio Score	1.904	1.444	1.518	0.207	0.230	0.186
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Self Evaluation Score	0.429	0.273	0.209	-0.008	-0.017	-0.011
	(0.009)	(0.143)	(0.167)	(0.794)	(0.640)	(0.685)
Peer Interview Score	-0.913	-0.639	-0.825	-0.072	-0.081	-0.035
	(0.000)	(0.008)	(0.000)	(0.046)	(0.106)	(0.282)
External References Score	2.564	2.457	1.843	0.266	0.270	0.235
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions and Year t-2 Test Score		x			x	
Class and School Controls			x			x
Mean Dep. Var. (Public School)	24.30	24.95	24.30	1.219	1.567	1.219
Observations	1,805,012	1,165,653	1,805,012	1,805,012	1,165,653	1,805,012
Portfolio Score	7.8%	5.8%	6.2%	13.0%	12.9%	12.4%
Self Evaluation Score	1.8%	1.1%	0.9%	0.3%	0.3%	-0.2%
Peer Interview Score	-3.8%	-2.6%	-3.4%	-5.6%	-5.4%	-5.7%
External References Score	10.6%	9.8%	7.6%	19.8%	19.8%	15.4%

Table 9: Instruments of Teacher Evaluation on Outcomes - Residual Regression.

Each column reports coefficients from an OLS regression between the dependent variable and TET instruments (portfolio, self evaluation, peer interview and external references), with standard errors clustered by school-cohort and p-value in parentheses. Columns 7-9 use an indicator for university attendance; columns 10-12 use an indicator for Top-3 university attendance. In the first column of each outcome-specification, we use the baseline control vector detailed in note of Table 3. In the second column of each outcome-specification, we use the baseline control vector adding the condition of vulnerability of the students, a discrete variable equal to 1 if the student belongs to the most vulnerable 40% of the population. In the third column of each outcome-specification, we use the baseline control vector adding the twice-lagged test scores. The last row of each table corresponds to the ratio between the impact of the Teacher Evaluation on the average of the dependent variable for students graduated from public schools.

Variable	(1)	(2)	(3)	(4)	(5)	(6)
	Tertiary Education Attendance			Vocational Education Attendance		
Weighted Total Score	1.910	1.910	1.633	-0.327	-0.327	-0.407
	(0.000)	(0.000)	(0.000)	(0.007)	(0.007)	(0.007)
Mean Dep. Var.	47.27	47.27	47.92	22.97	22.97	22.97
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions		x			x	
Year t-2 Test Score			x			x
Observations	1,805,012	1,805,012	1,165,653	1,805,012	1,805,012	1,165,653
VA Over Mean Dep. Var	4,0%	4,0%	3,4%	-1,4%	-1,4%	-1,8%

Variable	(7)	(8)	(9)	(19)	(20)	(21)
	University Attendance			Top-3 University Attendance		
Weighted Total Score	2.237	2.237	2.040	0.183	0.183	0.194
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mean Dep. Var.	24.30	24.30	24.95	1.22	1.22	1.57
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions		x			x	
Year t-2 Test Score			x			x
Observations	1,805,012	1,805,012	1,165,653	1,805,012	1,805,012	1,165,653
VA Over Mean Dep. Var	9,2%	9,2%	8,2%	15,0%	15,0%	12,4%

Table 10: Teachers' Evaluation Test Outcomes - Residual Regression.

Each column reports coefficients from an OLS regression between the residual of dependent variable and Teacher Evaluation for public schools, with standard errors clustered by school-cohort and p-value in parentheses. Columns 1-3 use an indicator for tertiary education attendance of students graduated from public schools; columns 4-6 use an indicator for vocational education attendance; columns 7-9 use an indicator for university attendance; columns 10-12 use an indicator for Top-3 university attendance. In the first column of each outcome-specification, we residualize each dependent variable using the baseline control vector detailed in note of Table 3. In the second column of each outcome-specification, we use the baseline control vector adding the condition of vulnerability of the students, a discrete variable equal to 1 if the student belongs to the most vulnerable 40% of the population. In the third column of each outcome-specification, we use the baseline control vector adding the twice-lagged test scores. The regressions are run on the sample restricted to observations with a non-missing Teacher Evaluation and estimated for each subject and according to level. The Weighted Total Score of Teacher Evaluation is scaled in units of student test score standard deviations and is estimated using data from classes taught by the same teacher in other years. The last row of each table corresponds to the ratio between the impact of the Weighted Total Score on the average of the dependent variable for students graduated from public schools.

sample), and increases the probability of a top-3 university education by 0.18-0.19 percentage points (an increase between 12.4% and 15.0% of mean attendance in the regression sample).

Figures 7a to 7d plot the residual of each educational attendance rate for students in school year t against ED_{jt}. To construct this binned scatter plot, we follow the same procedure detailed in section 3.5.2.

In the case of Teacher's Evaluation, our results suggest that, moving a student from a teacher in the fifth to the ninety-fifth percentile of TET score distribution would lead to increases the probability of tertiary education attendance of graduate students who attended public schools by 6.23 percentage points in a single grade[8]. Figure 6 illustrates the estimated effects of moving a student from a teacher in the fifth to the twenty-second (+1 SD or +0.25 points of TET score), fifty-eighth (+2 SD or +0.50 points of TET score) and ninety-fifth (+3.26 SD or +0.81 points of TET score) percentile of TET score distribution, respectively, on the probability of tertiary education attendance.

Finally, we replicate the study associated with each of the TET instruments. The results of Tables 16 and 17 in Appendix, remain the same as those obtained in section 3.6.1, being the most relevant instruments in explaining the probability of tertiary educational attendance, portfolio (1.30-1.71 percentage points), and external references (1.74-1.86 percentage points).

6 Comparison of two mesuares of teacher's quality

In this section, we compare both measures of teacher quality and their impact on entry into tertiary education. In the same regression of the previous sections, we include both measurements simultaneously to corroborate whether the found results are maintained.

Considering both instruments' application to the same sample, we use the methodology used in sections 3.5.2 and 3.6.2 only considering students who have graduated from public schools. For the above, we estimate the following equation, including controls by level, subject, and their interaction, as shown in equations 11 and 14:

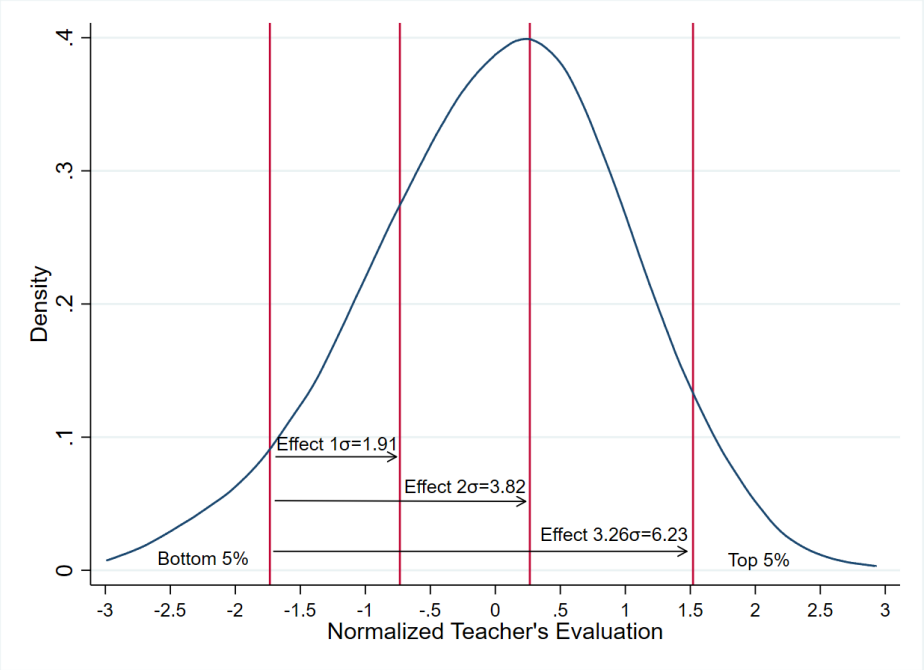
$$Y_{it} = \alpha + \tau \hat{m}_{jt} + \rho ED_{jt} + \omega ijt \tag{15}$$

Where Y_{it} is the residual of attendance, eliminating the effect of observable characteristics X_{it} from the regression (including the previous grade and teacher fixed effect), the same as we do for estimating equation 11 (Value-Added) and equation 14 (Teacher Evaluation). $\hat{m}_{jt} = \hat{\mu}_{jt}/\sigma_{\mu}$ is the normalized teacher value-added j, such that τ represents the reduced form of the impact of an increase of 1 SD on teacher value-added for a given year, or grade, on any variable of attendance. ED_{jt} is the normalized teacher evaluation j, such that ρ represents the reduced form of the impact of an increase of 1 SD on teacher evaluation for a given year, or grade, on any variable of attendance. Last, ωijt are the unobserved determinants of attendance, which are assumed to be orthogonal to teacher value-added estimates and teacher evaluation.

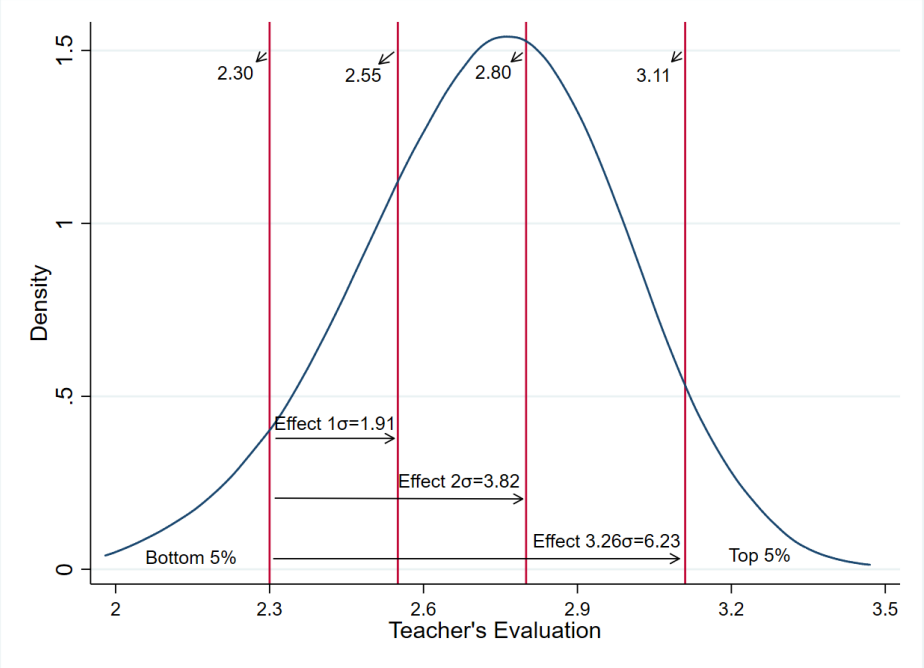
The Tables 11 and 12 present the results of the estimates for graduated students from public schools in the rates of tertiary, vocational, university, and top-3 university attendance and their respective robustness. The first column of each educational outcome specification is the result of equation 15 using our baseline controls. The second and third columns replicate the specification with the baseline control vector and add the vulnerability condition of student and twice-lagged test scores, respectively.

For the case of tertiary education attendance, a 1 SD increase in a teacher's evaluation in a single grade increases the probability of tertiary education attendance by 1.67-1.93 percentage points, which means an increase of 3.5%-4.1% of mean tertiary education attendance in the regression

[8] A teacher who is at the ninety-fifth percentile is 3.26 standard deviation better than one at the fifth percentile. Therefore, the effect is 3.26 x ρ, i.e., 3.26 x 1.91 = 6.23



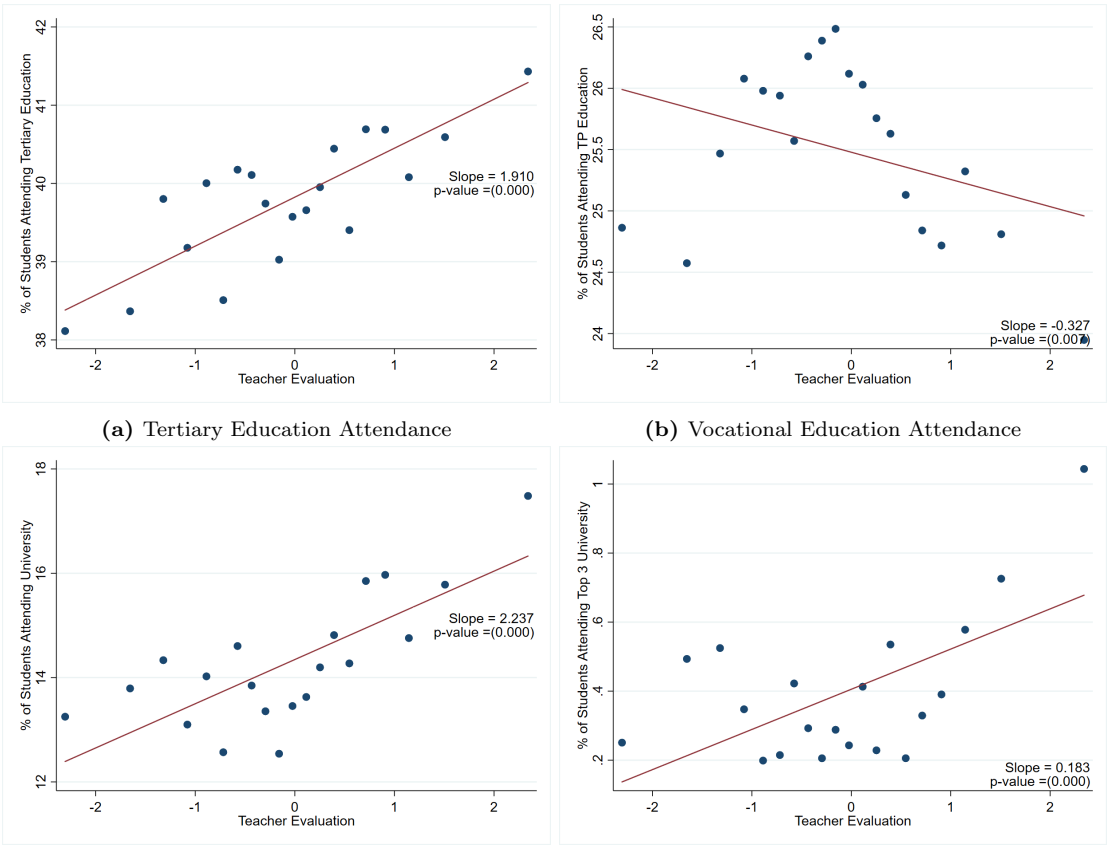
(a) Normalized TET score distribution.



(b) TET score distribution.

Figure 6: Effects of moving a student from Teacher's Evaluation in the fifth percentile of the TET score distribution in public schools.

Distribution of TET score and normalized TET score for language in public schools. The arrows in the figure correspond to the effect of moving to a student from a teacher in the 5th to the 22th, 58th and 95th percentile of the TET score distribution. Its calculation corresponds to $1 \times \rho = 1.91$, $2 \times \rho = 3.82$, and $3.26 \times \rho = 6.23$, respectively.



(a) Tertiary Education Attendance

(b) Vocational Education Attendance

(c) University Attendance

(d) Top-3 Universities Attendance

Figure 7: Conditional mean of Teachers' Evaluation on Educational Outcomes in Public Schools

Panels (a) to (d) are binned scatter plots of tertiary education attendance rates, vocational education attendance rates, university attendance rates and top-3 university attendance rates vs. Teacher's Evaluation ED_{jt}. These plots correspond to the regressions in the first column of each outcome-specification of Table 10 and use the same sample restrictions and variable definitions. To construct these binned scatter plots, we first residualize the dependent variable with respect to the baseline control vector separately within each subject, using within-teacher variation to estimate the coefficients. We then divide the VA estimates ED_{jt} into twenty equal-sized groups (vingtiles) and plot the means of the dependent variable residuals within each bin against the mean value of ED_{jt} within each bin. Finally, we add back the unconditional mean of the dependent variable in the estimation sample to facilitate interpretation of the scale. The solid line shows the best linear fit estimated on the underlying micro data using OLS.

Variable	(1)	(2)	(3)	(4)	(5)	(6)
	Tertiary Education Attendance			Vocational Education Attendance		
Weighted Total Score	1.936	1.936	1.669	-0.324	-0.324	-0.435
	(0.000)	(0.000)	(0.000)	(0.010)	(0.010)	(0.004)
Teachers' Value-Added	0.689	0.666	0.639	-0.230	-0.225	-0.435
	(0.004)	(0.005)	(0.014)	(0.173)	(0.180)	(0.024)
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions		x			x	
Year t-2 Test Score			x			x
Mean Dep. Var. (Public School)	47.27	47.27	47.95	22.96	22.96	22.95
Observations	1,726,710	1,726,710	1,152,901	1,726,710	1,726,710	1,152,901
Over Mean Dep. Ver.						
Weighted Total Score	4.1%	4.1%	3.5%	-1.4%	-1.4%	-1.9%
Teachers' Value-Added	1.5%	1.4%	1.3%	-1.0%	-1.0%	-1.9%

Table 11: Teachers' Value-Added and Teacher Evaluation on Outcomes.

Each column reports coefficients from an OLS regression between the residual of dependent variable and Teacher VA and Teacher Evaluation for public schools, with standard errors clustered by school-cohort and p-value in parentheses. Columns 1-3 use an indicator for tertiary education attendance of students graduated from public schools; columns 4-6 use an indicator for vocational education attendance. In the first column of each outcome-specification, we residualize each dependent variable using the baseline control vector. In the second column of each outcome-specification, we use the baseline control vector adding the condition of vulnerability of the students, a discrete variable equal to 1 if the student belongs to the most vulnerable 40% of the population. In the third column of each outcome-specification, we use the baseline control vector adding the twice-lagged test scores. The regressions are run on the sample restricted to observations with a non-missing in VA and Teacher's Evaluation model, and estimated for each subject and according to level. The score for VA and Weighted Total Score of Teacher Evaluation is scaled in units of student test score standard deviations and is estimated using data from classes taught by the same teacher in other years. The last row of each table corresponds to the ratio between the impact of each measure on the average of the dependent variable for students graduated from public schools.

Variable	(7)	(8)	(9)	(10)	(11)	(12)
	University Attendance			Top-3 University Attendance		
Weighted Total Score	2.260	2.260	2.104	0.196	0.196	0.202
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Teachers' Value-Added	0.919	0.891	1.074	0.123	0.122	0.158
	(0.011)	(0.013)	(0.007)	(0.121)	(0.123)	(0.121)
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions		x			x	
Year t-2 Test Score			x			x
Mean Dep. Var. (Public School)	24.31	24.31	25.00	1.25	1.25	1.57
Observations	1,726,710	1,726,710	1,152,901	1,726,710	1,726,710	1,152,901
Over Mean Dep. Ver.						
Weighted Total Score	4.8%	4.8%	4.4%	0.9%	0.9%	0.9%
Teachers' Value-Added	1.9%	1.9%	2.2%	0.5%	0.5%	0.7%

Table 12: Teachers' Value-Added and Teacher Evaluation on Outcomes.

Each column reports coefficients from an OLS regression between the residual of dependent variable and Teacher VA and Teacher Evaluation for public schools, with standard errors clustered by school-cohort and p-value in parentheses. Columns 7-9 use an indicator for university attendance; columns 10-12 use an indicator for Top-3 university attendance. In the first column of each outcome-specification, we residualize each dependent variable using the baseline control vector. In the second column of each outcome-specification, we use the baseline control vector adding the condition of vulnerability of the students, a discrete variable equal to 1 if the student belongs to the most vulnerable 40% of the population. In the third column of each outcome-specification, we use the baseline control vector adding the twice-lagged test scores. The regressions are run on the sample restricted to observations with a non-missing in VA and Teacher's Evaluation model, and estimated for each subject and according to level. The score for VA and Weighted Total Score of Teacher Evaluation is scaled in units of student test score standard deviations and is estimated using data from classes taught by the same teacher in other years. The last row of each table corresponds to the ratio between the impact of each measure on the average of the dependent variable for students graduated from public schools.

sample. For teachers' true VA test scores in a single grade, the effect is an increase of 0.64-0.69 percentage points, which means a 1.3%-1.5% increase in the mean tertiary education attendance. In both cases, the null hypothesis that teachers' evaluation and value-added have no effect on tertiary education attendance is rejected with a $p - value < 0.001$.

For the case of vocational attendance, the results for VA are not significant. At the same time, for teacher evaluation, a 1 SD increase in TET decreases the probability between 0.32 and 0.44 percentage points (1.4%-1.9% decrease in the mean vocational education attendance in the regression sample).

For universities, the result is that a 1 SD increase in a teacher's evaluation in a single grade increases the probability of university attendance by 2.10-2.26 percentage points (an increase of 4.8%-4.4% of mean university education attendance in the regression sample). For the case of teachers' true VA test scores in a single grade, the effect is an increase of 0.92-1.07 percentage points (an increase of 1.9%-2.2% of mean university attendance). In both cases, the null hypothesis that teachers' evaluation and value-added have no effect on tertiary education attendance is rejected with a $p - value < 0.001$.

Finally, the VA results are not significant for top-3 university attendance. At the same time, for teacher evaluation, a 1 SD increase in TET increases the probability by 0.19-0.20 percentage points (0.9% increase in the mean university attendance in the regression sample).

As shown by these results, we verify that i) both impacts remain in the same ranges of equations 11 and 14 for each of the variables of attendance; ii) a 1 SD increase in a teacher's evaluation affects between two or three times the effect of a 1 SD increase in a teacher's value-added; and iii) both measurements turn out to be orthogonal, from which we can infer that both would capture different dimensions or abilities of teachers when studying their impact on tertiary education attendance.

7 Concluding Remarks

This essay contributes to the discussion of the impact teachers have on their students' academic results once they graduate from secondary education. To this end, we analyze two measurements used by most countries: the value-added methodology and the teacher's evaluation. We use the same sample of students and teachers to study each of them, including all students who graduated from public schools for the same period.

In the case of the value-added methodology, all evidence available in the literature is verified to ensure that the results present minimal bias that allows us to conclude and infer from them. In this sense, the results of both observables and quasi-experiment bias are in ranges similar to those found by Chetty et al. (2014a) for the US.

Similarly, in the case of teacher's evaluation, its effects are estimated from two methodologies, finding practically the same results in both cases. The first is commonly used by the literature (Kane et al., 2008, Kane et al., 2011, Briole and Maurin, 2019, among others), and the second allows us to estimate both measures simultaneously. Additionally, this paper contributes to quantifying which of a series of instruments, used to a greater or lesser extent by all countries that apply this type of evaluation, has a greater relationship with tertiary, vocational and university attendance.

Our results suggest that the correlation between (TET) and (TVA) appears to be null in school outcomes. However, our analysis also reveals that both measures, TET and TVA, positively affect the probability of tertiary education attendance, indicating that both measures are complementary in measuring teacher quality in the middle run. These results have relevance from the public policy point of view as unlike countries (e.g. USA) where TVA is used for teacher's promotions and personnel decisions, in countries where TVA is not used for teacher's personnel decisions (e.g. Chile), TVA seems to be useful to measure teacher quality. Furthermore, our findings are consistent with the argument of the multidimensionality of teaching quality, because even though in the short run TVA and TET seem to be orthogonal, in the medium run they seem to be complementary

tools to measure teacher effectiveness.

The main conclusion of our study is that, unlike the USA where TVA is used for teacher's promotions and personnel decisions, in countries like Chile, where TVA is not used for teacher's personnel decisions, TVA also seems to be a useful tool to measure teacher quality in the medium run. Furthermore, our findings are consistent with the argument of the multidimensionality of teaching quality, because even though in the short run TVA and TET seem to be orthogonal, in the medium run they seem to be complementary tools to measure teacher effectiveness.

Finally, suppose we weigh the potential costs of each, especially for a developing economy. In that case, these results give an account of certain instruments that could be applied to the extent that their economic resources allow. In our evaluation and the availability to access certain students' socioeconomic variables and their grades, the value-added measurement is highly cost-effective. Likewise, an external evaluation from the hierarchical superiors, based on a precise and structured guide based on some good teaching framework, which includes several questions about the evaluated teacher's performance, turns out to be a cost-effective tool. Additionally, an instrument that presents evidence of the pedagogical practice of the teacher in an objective way designed and adapted for each context by a centralized institution, in our case the Portfolio, is a third tool that would help to identify the teachers who could have a better impact on their students' academic results. However, as reviewed during this paper, this tool may be time-consuming for teachers and costly from a fiscal perspective.

Appendix

Domain A: Teaching, Planning, and Preparation	**Domain B: Development of an Adequate Environment for Learning**	**Domain C: Instruction**	**Domain D: Professional Responsibilities**
A1. Masters the content of the discipline s/he teaches and the national curriculum framework	B1. Establishes an environment of acceptance, equity, trust, solidarity and respect	C1. Communicates learning goals in a precise and clear way	D1. Systematically reflects on his/her performance
A2. Consciousness of students' characteristics, knowledge and experiences	B2. Expresses high expectations about the possibilities of learning and development of all students	C2. The teaching strategies are challenging and significant for the students	D2. Forms teams and professional relationships with his/her colleagues
A3. Masters the didactics of the disciplines s/he teaches	B3. Establishes and keeps consistent rules of behavior in the classroom	C3. The content of the class is rigorously covered and it is understandable for the students	D3. Takes responsibility for his/her students' guidance
A4. Organizes the goals and content in a way that it is coherent with the national curriculum framework and the distinctive feature of his/her students	B4. Organizes a structured environment and uses the available resources	C4. Optimizes the available teaching time	D4. Fosters collaborative and respectful relationships with students' parents or representatives
A5. The evaluation strategies are coherent with the learning goals, the discipline taught and the national curriculum framework and allow all students to show what they have learned		C5. Promotes the development of thinking	
		C6. Evaluates and monitors the process of understanding and appropriation of the contents by the students	

Table 13: Domains and Criteria of the Framework for Good Teaching (FGT).

The FGT has 4 domains, 20 criteria, and 71 descriptors. To save space, the descriptors are not included in the table.

Domain A: Teaching Planning, and Preparation					
Criteria A2: Consciousness of students' characteristics, knowledge, and experiences					
Descriptor	Portfolio (Module 1)	Portfolio (Module 2)	External References	Peer Interview	Self-Evaluation
A.2.1.Consciousness of the developmental characteristics of his/her students according to their age	*			*	*
A.2.2. Consciousness of the family and cultural heterogeneity of the students	*	*	*	*	*
A.2.3. Consciousness of the strengths and weaknesses of the students relative to the content	*	*		*	*
A.2.4. Consciousness of the different learning methods of the students	*	*		*	*

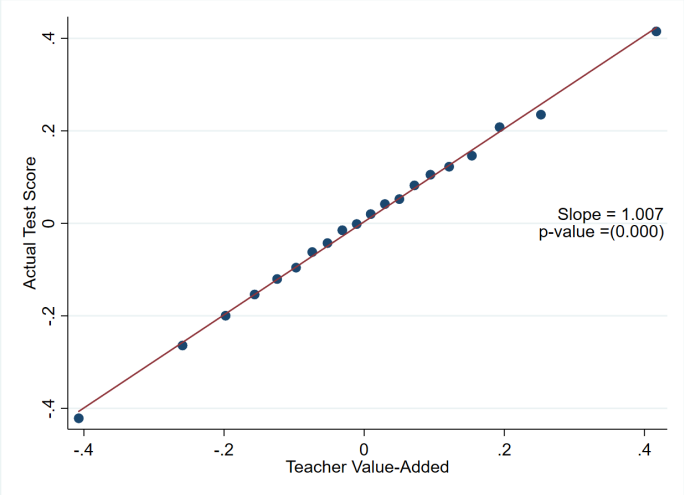
Table 14: Mapping between the Instruments and the FGT.

Taken from Manzi et al. (2011). In this example we see that descriptor A.2.2 is measured with all the instruments, while the other descriptors are measured with fewer instruments. ∗=measured by the instrument.

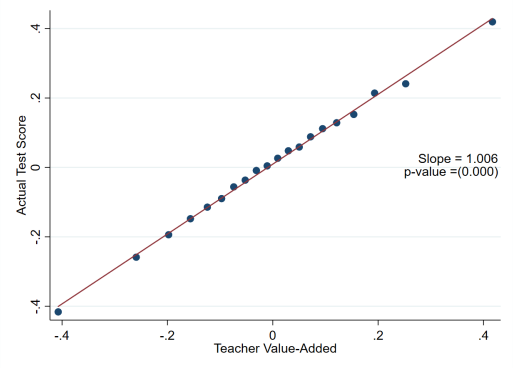
Module	Dimension	Teaching Practice
Module 1	A. Organization of the unit	a.a. Formulation of goals a.b. Relationship between activities and goals a.c. Sequence of the unit
	B. Analysis of the class	b.a. Analysis based on students' characteristics b.b. Analysis of the carried-out unit b.c. Analysis of the class
	C. Quality of the evaluation	c.a. Evaluations and rubrics used c.b. Relationship between evaluations and goals
	D. Reflection on students' results	d.a. Responsibility for students' results d.b. Students' feedback
Module 2 (video)	F. Environment of the class	f.a. Work environment f.b. Promotion of students' participation f.c. Activity's support and guidance
	G. Structure of the class	g.a. Quality at the beginning of the class g.b. Quality at the end of the class g.c. Activity's contribution to the fulfillment of goals
	H. Pedagogical interaction	h.a. Developed explanations h.b. Quality of the questions asked of the students h.c. Feedback quality h.d. Curricular emphasis

Table 15: Dimensions and Teaching Practices Evaluated by the Portfolio.

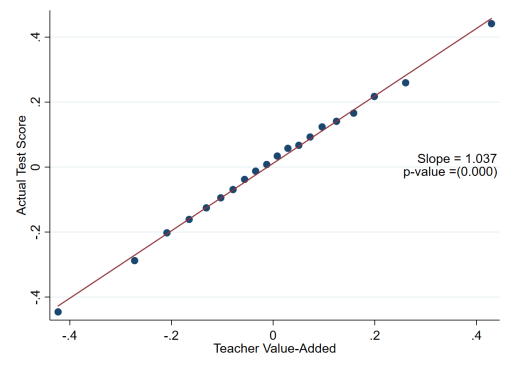
Own elaboration based on the document "Evaluación Docente como herramienta de gestión escolar." CPEIP, Mineduc. Dimension E was not considered in the estimation as it was discontinued.



(a) Baseline



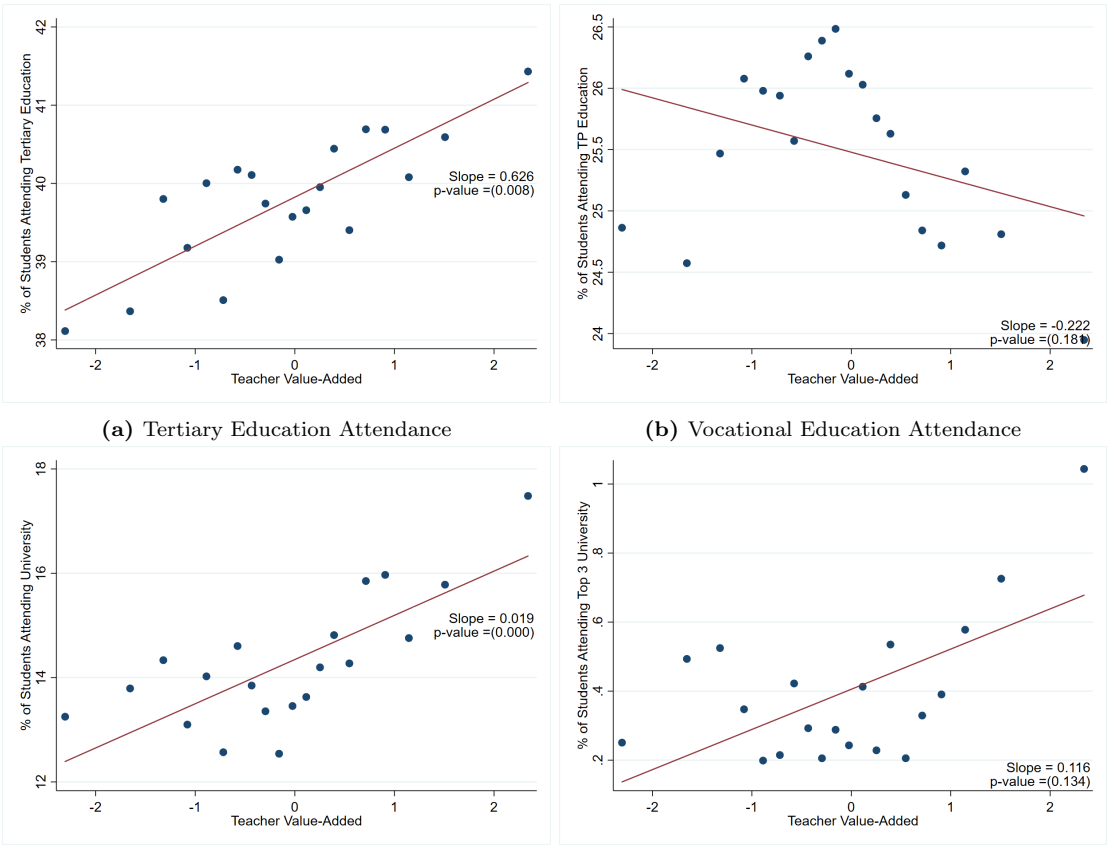
(b) Added includes Two Lags of Test Score



(c) Added includes Vulnerability Conditions

Figure 8: Effect of Teachers' Value-Added on Actual Scores in Public Schools

These Figures are constructed using the sample used to estimate VA model for public schools. To construct this binned scatter plot, we first residualisze the actual test score with respect to the baseline control vector (detailed in the note of Table 3) separately within each subject and using within-teacher variation to estimate the coefficients for public schools. Then divide the VA estimates $\hat{\mu}_{jt}$ into twenty equal-sized groups (vingtiles) and plot the means of the actual test score residuals within each bin against the mean value of $\hat{\mu}_{jt}$ within each bin. The line shows the best linear fit estimated on the underlying micro data using OLS



(a) Tertiary Education Attendance

(b) Vocational Education Attendance

(c) University Attendance

(d) Top-3 Universities Attendance

Figure 9: Conditional mean of Teachers' VA on Educational Outcomes in Public Schools

Panels (a) to (d) are binned scatter plots of tertiary education attendance rates, vocational education attendance rates, university attendance rates and top-3 university attendance rates vs. normalized teacher VA $\hat{m}_{jt}$ for public schools. These plots correspond to the regressions in the first column of each outcome-specification of Table 6 and use the same sample restrictions and variable definitions. To construct these binned scatter plots, we first residualize the dependent variable with respect to the baseline control vector separately within each subject, using within-teacher variation to estimate the coefficients. We then divide the VA estimates $\hat{m}_{jt}$ into twenty equal-sized groups (vingtiles) and plot the means of the dependent variable residuals within each bin against the mean value of $\hat{m}_{jt}$ within each bin. Finally, we add back the unconditional mean of the dependent variable in the estimation sample to facilitate interpretation of the scale. The solid line shows the best linear fit estimated on the underlying micro data using OLS.

Variable	(1)	(2)	(3)	(4)	(5)	(6)
	Tertiary Education Attendance			Vocational Education Attendance		
Portfolio Score	1.708	1.678	1.297	-0.305	-0.299	-0.346
	(0.000)	(0.000)	(0.000)	(0.014)	(0.015)	(0.018)
Self Evaluation Score	0.461	0.454	0.332	-0.002	-0.001	-0.024
	(0.003)	(0.003)	(0.071)	(0.981)	(0.994)	(0.856)
Peer Interview Score	-0.470	-0.426	-0.429	0.475	0.466	0.410
	(0.010)	(0.018)	(0.033)	(0.000)	(0.000)	(0.008)
External References Score	1.741	1.730	1.863	-0.977	-0.974	-0.938
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions		x			x	
Year t-2 Test Score			x			x
Mean Dep. Var. (Public School)	47.27	47.27	47.92	22.97	22.97	22.97
Observations	1,805,012	1,805,012	1,165,653	1,805,012	1,805,012	1,165,653
Over Mean Dep. Ver.						
Portfolio Score	3.6%	3.5%	2.7%	-1.3%	-1.3%	-1.5%
Self Evaluation Score	1.0%	1.0%	0.7%	0.0%	0.0%	-0.1%
Peer Interview Score	-1.0%	-0.9%	-0.9%	2.1%	2.0%	1.8%
External References Score	3.7%	3.7%	3.9%	-4.3%	-4.2%	-4.1%

Table 16: Instruments of Teacher Evaluation on Outcomes - Residual Regression.

Each column reports coefficients from an OLS regression only for public schools, with standard errors clustered by school-cohort and p-value in parentheses. Dependent variable of first column of each outcome-specification correspond to residuals of educational outcome using the same control vector used to estimate baseline VA model detailed in Table 3. Dependent variable of second column of each outcome-specification, add in the estimate of residuals of educational outcome the condition of vulnerability of the students, a discrete variable equal to 1 if the student belongs to the most vulnerable 40% of the population. Dependent variable of third column of each outcome-specification, add in the estimate of residuals of educational outcome the twice-lagged test scores.The regressions are run on the sample used to estimate the baseline Teacher Evaluation model, restricted to observations with a non-missing Teacher Evaluation. There is one observation for each student-subject-grade-school year in all regressions.The score for each instrument of Teacher Evaluation is scaled in units of student test score standard deviations and is estimated using data from classes taught by the same teacher in other years. Impact of each instrument of Teacher Evaluation is estimated for each subject (language and mathematics) and according to level (primary and high school). The last row of each table corresponds to the ratio between the impact of each instruments of Teacher's Value-Added on the average of the dependent variable.

Variable	(7)	(8)	(9)	(10)	(11)	(12)
	University Attendance			Top-3 University Attendance		
Portfolio Score	2.013	1.977	1.643	0.158	0.157	0.194
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Self Evaluation Score	0.464	0.454	0.356	0.004	0.004	-0.003
	(0.006)	(0.006)	(0.073)	(0.896)	(0.903)	(0.936)
Peer Interview Score	-0.945	-0.892	-0.839	-0.068	-0.066	-0.090
	(0.000)	(0.000)	(0.002)	(0.063)	(0.069)	(0.085)
External References Score	2.718	2.705	2.801	0.241	0.241	0.241
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Baseline Controls	x	x	x	x	x	x
Vulnerability Conditions		x			x	
Year t-2 Test Score			x			x
Mean Dep. Var. (Public School)	24.30	24.30	24.95	1.219	1.219	1.567
Observations	1,805,012	1,805,012	1,165,653	1,805,012	1,805,012	1,165,653
Over Mean Dep. Ver.						
Portfolio Score	8.3%	8.1%	6.6%	13.0%	12.9%	12.4%
Self Evaluation Score	1.9%	1.9%	1.4%	0.3%	0.3%	-0.2%
Peer Interview Score	-3.9%	-3.7%	-3.4%	-5.6%	-5.4%	-5.7%
External References Score	11.2%	11.1%	11.2%	19.8%	19.8%	15.4%

Table 17: Instruments of Teacher Evaluation on Outcomes - Residual Regression.

Each column reports coefficients from an OLS regression only for public schools, with standard errors clustered by school-cohort and p-value in parentheses. Dependent variable of first column of each outcome-specification correspond to residuals of educational outcome using the same control vector used to estimate baseline VA model detailed in Table 3. Dependent variable of second column of each outcome-specification, add in the estimate of residuals of educational outcome the condition of vulnerability of the students, a discrete variable equal to 1 if the student belongs to the most vulnerable 40% of the population. Dependent variable of third column of each outcome-specification, add in the estimate of residuals of educational outcome the twice-lagged test scores.The regressions are run on the sample used to estimate the baseline Teacher Evaluation model, restricted to observations with a non-missing Teacher Evaluation. There is one observation for each student-subject-grade-school year in all regressions.The score for each instrument of Teacher Evaluation is scaled in units of student test score standard deviations and is estimated using data from classes taught by the same teacher in other years. Impact of each instrument of Teacher Evaluation is estimated for each subject (language and mathematics) and according to level (primary and high school). The last row of each table corresponds to the ratio between the impact of each instruments of Teacher's Value-Added on the average of the dependent variable.

References

[1] Altonji, J. G., & Mansfield, R. K. (2014). "Group-average observables as controls for sorting on unobservables when estimating group treatment effects: The case of school and neighborhood effects" (No. w20781). National Bureau of Economic Research.

[2] Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc (2011). "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." American Economic Journal: Applied Economics 3 (3): 29–54.

[3] Bau, N. and Das, J. (2020) "Teacher Value Added in a Low-Income Country". American Economic Journal: Economic Policy, 12(1): 62–96.

[4] Briole, S. and E. Maurin (2019) "Does Evaluating Teachers Make a Difference?". IZA Discussion Paper Series No. 12307.

[5] Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). "Professional Practice, Student Surveys, and Value-Added: Multiple Measures of Teacher Effectiveness in the Pittsburgh Public Schools". REL 2014-024. Regional Educational Laboratory Mid-Atlantic.

[6] Chetty, R., Friedman, J. and Rockoff, J. (2014a). "Measuring the impacts of teacher I: Evaluating the bias in teacher value-added estimates". American Economic Review, 104(9): 2593–2632.

[7] Chetty, R., Friedman, J. and Rockoff, J. (2014b). "Measuring the impacts of teacher II: Teacher Value-Added and Student Outcomes in Adulthood". American Economic Review, 104(9): 2633–2679.

[8] Chin, M. and Goldhaber, D. (2015) "Exploring Explanations for the 'Weak' Relationship Between Value Added and Observation-Based Measures of Teacher Performance". Mimeo, Center for Education Policy Research at Harvard University.

[9] Contreras, D., Bustos, S. and Sepúlveda, P. (2010). "When schools are the ones that choose: The effect of screening in Chile". Social Science Quarterly, 91: 1349–1368.

[10] Darling-Hammond, L. (2015) "Can Value Added Add Value to Teacher Evaluation?" Educational Researcher Volume: 44 issue: 2, page(s): 132-137.

[11] Grossman, P., Loeb, S., Cohen, J. and Wyckoff, J. (2013). "Measure for measure: The relationship between measures of instructional practice in middle school English arts and teachers' value-added scores". American Journal of Education, 119: 445–470.

[12] Hanushek, E. and Rivkin, S. (2010). "Generalizations about using value added measures of teacher quality". American Economic Review, 100(2): 267–271.

[13] Harris, D. (2012). "How do Value-Added indicators compare to other ,easures of teacher effectiveness?". Draft. Carnegie Knowledge Network What we know series: Value-added methods and application. Stanford, CA. Retrieved from http://www.carnegieknowledgenetwork.org/wp-content/uploads/2012/10/CKN_201210_Harris.pdf

[14] Heckman, J. J., Humphries, J. E., Veramendi, G., & Urzua, S. S. (2014). "Education, health and wages" (No. w19971). National Bureau of Economic Research.

[15] Isoré, M. (2009). "Teacher evaluation: Current practices in OECD countries and a literature review". OECD Education Working Papers, No. 23, OECD Publishing, Paris.

[16] Jackson, K. (2014). "Teacher quality at the high-school level: The importance of accounting for tracks". Journal of Labor Economics 23(4): 645–684.

[17] Jackson, C. K., Rockoff, J. E., and Staiger, D. O. (2014). Teacher effects and teacher-related policies. Annu. Rev. Econ., 6(1), 801-825.

[18] Kane, T. J., and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation (No. w14607). National Bureau of Economic Research.

[19] Kane, T., Rockoff, J. and Staiger, D. (2008). "Estimating teacher impacts on student achievement: An experimental evaluation". NBER Working Paper 14607.

[20] Kane, T., Taylor, E., Tyler, J. and Wooten, A. (2011). "Identifying effective classroom practices using student achievement data". Journal of Human Resources, 46(3): 587–613.

[21] Kane, T., McCaffrey, D., Miller, T. and Staiger, D. (2013). "Have we identified effective teachers? Validating measures of effective teaching using random assignment". Seattle, WA: Bill and Melinda Gates Foundation.

[22] Koedel, C., Mihaly, K. and Rockoff, J. (2015). "Value added modeling: A review". Working Paper 1501, University of Missouri.

[23] Lynch, K., Chin, M. and Blazar, D. (2017). "Relationships between observations of elementary teacher mathematics instruction and student achievement: Exploring variability across districts". American Journal of Education, Volume 123, Number 4

[24] Manzi, J., González, R. and Sun, Y. (2011). "La Evaluación Docente en Chile". Mide UC, Facultad de Ciencias Sociales, Escuela de Psicología, Universidad Católica de Chile.

[25] OECD. (2013a). "Synergies for Better Learning: An International Perspective on Evaluation and Assessment". OECD Reviews of Evaluation and Assessment in Education, Editions OCDE, Paris.

[26] OECD. (2013b). "Teachers for the 21st Century: Using Evaluation to Improve Teaching". OECD Publishing.

[27] Rivkin, S., Hanushek, E. and Kain, J. (2005). "Teachers, schools, and academic achievement". Econometrica, 73(2): 417–458.

[28] Rothstein, J. (2010). "Teacher quality in educational production: Tracking, decay, and student achievement". Quarterly Journal of Economics, 125(1): 175–214.

[29] Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). "A potential outcomes view of value-added assessment in education". Journal of educational and behavioral statistics, 29(1), 103-116.

[30] Singh, Abhijeet. 2015. "Private School Effects in Urban and Rural India: Panel Estimates at Primary and Secondary School Ages." Journal of Development Economics 113: 16–32.

[31] Singh, Abhijeet. 2019. "Learning More with Every Year: School Year Productivity and International Learning Divergence." Journal of the European Economic Association, JVZ033.

[32] Taylor, E. and Tyler, J. (2012). "The effect of evaluation on teacher performance". American Economic Review, 102(7):3628–51.

[33] Valenzuela, J., Bellei, C. and De los Ríos D. (2013). "Socioeconomic Segregation in a market-oriented Educational System. The Case of Chile". Journal of Educational Policy, 29 (2):217-241

[34] Wyness, G., Murphy, R. and Weinhardt, F. (2018). "Who Teaches the Teachers? a RCT of Peer-to-Peer Observation and Feedback in 181 Schools". Discussion Paper No.116, Rationality & Competition.