POSTGRADO
ECONOMÍA Y NEGOCIOS
UNIVERSIDAD DE CHILE

Teacher value-added in Chile

Tesis para optar al grado de

Magíster en Economía

Alumno: Ignacio Urrea

Profesor Guía: Daniel Hojman

Santiago, May 2018

Abstract

This paper estimates teacher value-added measures of teacher quality in Chile. Using administrative data we link individual student test score results to teacher assignment for 6th and 8th grade students, and control for a rich vector of covariates for the value-added estimation, including previous score and tuition fees. We evaluate the degree of accuracy of our teacher value-added estimates for predicting teachers' impacts on student achievement, by means of a teacher switching quasi-experiment devised by Chetty et al. (2014a), and find no significant bias in these estimates when controlling for our full set of variables. We evaluate next which controls are most important for the unbiasedness of our estimates, and find that previous score, and as a novelty in the literature, tuition fees, are essential for these purposes. We study the sorting on teacher value-added and report positive sorting between socioeconomic measures and teacher value-added estimates, meaning that better off students and schools get the highest performing teachers.

JEL: I21, I24, J24

Keywords: Analysis of Education, Education and Inequality, Human Capital • Skills • Occupational Choice • Labor Productivity.

Acknowledgments

I would like to thank my thesis advisor Daniel Hojman for his immense and insightful support, as well as the members of my thesis committee Susana Claro and Alejandra Mizala for their helpful comments. Financial support from Centro COES CONICYT/FONDAP/15130009 is acknowledged.

Contents

Chapter 1

Introduction

The economic effect of cognitive skills has been recognized as a strong policy factor explaining differences in economic growth among countries. Hanushek & Woessmann (2011) find that neither institutional nor regulatory differences can account convincingly for differences in the long-run economic growth among rich countries, while cognitive skills present themselves as a strong candidate underlying these differences between OECD members. School is one of the main determinants of cognitive skills formation, and in turn teacher quality is one of the most important assets of a school, if not the most important. Determining how good a teacher is at improving his students' achievement, however, is not an easy task, yet one of the utmost importance. Higher teacher quality can have a tremendous impact on life earnings, of the order of $250,000 on lifetime income per classroom, only by replacing a teacher in the bottom 5% of value-added quality measure by an average teacher (Chetty et al. (2014b)).[1] Reliable teacher quality measures would let us be able to recognize the differences in effectiveness among teachers, giving us the possibility to ensure better student performance, and thus better lifetime income and higher long-run economic growth.

Value-added modelling is more and more widely used as a measure of teacher quality. Part of the interest stems from its intuitive and straightforward approach. A teacher's value-added corresponds to the unique contribution she makes to her students' achievement (Corcoran (2010)), and is defined as the impact of a teacher once we control for other explicative variables of a student's score such as the student/family background, classroom, school and community factors. Until recently, the focus was primarily on determining the teacher's observable characteristics that had an impact on student achievement, but

[1] Hanushek (2011) find that the economic impact of teachers is over $400,000 annually for a teacher with a value-added one standard deviation higher for a class of 20 students.

the research agenda has been lately exploring less parametric approaches to identify the teachers' unique contribution to the students' overall performance (Hanushek Rivkin, 2010, 2012; Chetty et al. (2017)).

Value-added measures rely on a selection on observables assumption: once we control for a sufficiently rich set of covariates that determine the students' achievement on a particular test, teachers must be as good as conditional randomly assigned on these characteristics (Angrist et al. (2016)). To what extent are controllable characteristics able to account for sorting such as student-teacher sorting, school-teacher sorting or school choice is the key factor to determine how biased our value-added estimators will be. Resolving this is fundamental for public policy, as teachers may be rewarded or penalized for considerations that are out of their control, such as the mix of students in their classrooms (Chetty et al. (2014a)). Another important property of value-added estimates is the intertemporal stability of the estimates, but its relevance will depend on the policy question. For example, year to year correlations may be less relevant than year to career correlations in value-added measures, if we are interested in teacher-career performance (see Staiger & Kane (2014)).

Bias in value-added estimates can be understood as two different, yet related, kind of biases. Teacher-level bias is the degree to which individual teachers' VA estimates differ in expectation from their true effects (Chetty et al. (2017)), while forecast bias determines how inaccurate predictions of teachers' causal effects on student achievement on average are. While both are related, teacher-level bias is considerably harder to estimate than forecast bias, yet both are policy pertinent, depending on the policy problem we are trying to solve. We will focus on the latter and refer to it simply as "bias".

This paper makes three primary contributions. First, we provide teacher value-added estimation and characterization of their (un)biasedness in Chile, where this characterization, to our knowledge, has not yet been implemented. Even though recent efforts have estimated teacher value-added measures in Chile, none has evaluated whether they are unbiased estimates, an essential property that needs to be acknowledged shall teacher value-added measures be implemented with public policy purposes. In other parts of the world, their use as public policy input has become widespread due to its simplicity to understand and low implementation cost, what makes it particularly attractive to use in a developing country.[2]

Second, this paper shows the bias estimates in teacher value-added measures in a totally new context, and provides new empirical evidence about their reliability. Former literature, has consistently found no bias in value-added measures, or very little bias (Kane & Staiger (2008); Kane et al. (2013); Chetty et al.

[2] Hsieh & Urquiola (2006) implement a methodology to evaluate effects of unrestricted choice on educational outcomes in Chile, similar in spirit to the one used in this paper (and described in Section 6) to evaluate bias in teacher value-added estimates, though at the community level.

(2014a); Rothstein (2017); Bacher-Hicks et al. (2014), Bacher-Hicks et al. (2017)), with the use of experimental (Kane & Staiger (2008); Kane et al. (2013); Bacher-Hicks et al. (2017)) or quasi-experimental (Chetty et al. (2014a); Rothstein (2017); Bacher-Hicks et al. (2014)) methods. However, even though they are implemented in diverse settings (different districts and states), the previous studies all explore bias in value-added measures within the same country, the United States. Whether value-added measures maintain their unbiasedness in considerable different contexts, such as in a developing country with strong competitive forces and unrestricted choice in the educational system as Chile, is an issue that remains to be explored.

Third, we study the importance of tuition charged for value-added estimation. As a novelty in literature, we include the tuition charged by the schools to parents in the estimation of value-added measures and their bias. In Chile, this tuition (copayment, or *copago*) is charged by private-voucher schools following a financing reform to the chilean voucher system.[3] It is used as a supplement to the subsidy provided by the government, in a system known as *financiamiento compartido* (shared financing). The availability of data on tuition charged by schools from this important subpopulation of schools in Chile[4] allows us to evaluate the effect of add-ons charged to the families on bias in teacher value-added estimations, in a highly socioeconomically stratified system. The remainder of this paper is organized as follows. In section 2, we provide a brief review on the relevant literature. In section 3, we provide the conceptual framework and methodology, in section 4, we present the data. In section 5 we describe our value-added estimates. In section 6 we implement the quasi-experimental estimates of bias. In section 7 we conclude.

[3] Tuition is also charged in private unsubsidized schools (*particulares pagados*) but we do not include them in the analysis, as we don't have data availability on the tuition charged by these schools. These schools represent a small share of total schools in Chile and tend to serve the chilean elite.

[4] Mainly the subsidized private schools (*particulares subvencionados*).

Chapter 2

Background and literature

Up to this date, many studies have estimated the distribution of teacher effectiveness (see Hanushek & Rivkin (2012) or Koedel et al. (2015) for a review), but only a few have tried to determine the bias in these estimates. Kane & Staiger (2008) provide an experimental evaluation of the bias in non-experimental value-added measures, by evaluating whether non-experimental estimates could predict differences in achievement between classrooms that were randomly assigned. They do so by regressing the difference in average achievement among pairs of classrooms randomly assigned to treatment and control groups, on the within-pair difference in the non-experimental teacher effect, computed previous random assignment. They implement this random-assignment experiment in the Los Angeles Unified School District (LAUSD), where 78 pairs of elementary school classrooms were randomly assigned to different teachers. Their value-added estimation uses only within school randomization, but uses both within and between evaluation of the bias in value-added estimates. They find no bias for those non-experimental measures that conditioned on prior student achievement in some manner.

Another study that took advantage of random assignment is Kane et al. (2013): as part of the Measures of Effective Teaching (MET) project, the authors randomize on a much greater basis, with rosters of students assigned to 1591 teachers across six districts. They implement the same intuitive approach that Kane & Staiger (2008) to evaluate bias, but built upon a composite measure of teacher quality comprised of test scores, student surveys, and classroom observations in the prior school year. The authors find no bias in the composite measure of teacher effectiveness, and provide evidence to recommend the inclusion of prior achievement as a control in value-added models.

While the previous studies randomize within-schools, Glazerman & Protik (2015) use a between school randomization, to address also between-school sorting. They do so by using a subsample drawn from the transfer incentives study from the Glazerman et al. (2013) study. The authors' intervention consisted in a multisite randomized field trial, where they identified the highest performing teachers (as measured by their value-added score), and randomly assigned them to vacancies in schools with very low achievement, either to a treatment group that could hire from the high value candidates, or to a control group where the school filled vacancies as they usually did. Glazerman & Protik (2015) find no bias at the elementary level, but obtain evidence that their middle school value-added measures are biased.

Implementing experimental evaluations is not always feasible. Chetty et al. (2014a) (henceforth, CFR) presented a quasi-experiment analog to the Kane & Staiger (2008) experiment. The quasi-experiment exploits teacher turnover at the school-grade level, where teachers from a school-grade level are replaced with a teacher of varying quality (as measured by his value-added). Whenever a school-grade teacher is replaced by a higher (lower) value-added teacher, the average value-added increases (decreases) in that school-grade. As long as a one-unit change in average teacher value-added is able to predict a one unit change in mean scores at the school-grade level, value-added estimates are unbiased. This plausibly exogenous quasi-experiment relies on the assumption that teacher staff changes from year to year are uncorrelated with school and teachers characteristics.

While in an experimental setting we can ensure by design that teacher assignment is uncorrelated with students' or schools' characteristics, it is not clear that this is the case for the naturally occurring changes in staff within a school-grade cell. Though intuitively plausible, as the authors argue parents are not likely to switch their children to a new school just because some year a teacher arrives or leaves, it is still an empirical issue whether the design of the quasi-experiment is valid. CFR present some diagnostic tests to evaluate the exogeneity of the quasi-experiment. They found that changes in VA estimates are uncorrelated to changes in cross-cohort parent characteristics. The authors also present the relationship between the changes in the average of the other subject score in a school-grade cell and the change in average VA in the cell. If their key assumption holds, there should be no or little relationship for middle school (considering different teachers teach different subjects to a same school grade), unless the VA of a teacher in a subject also influences the outcome in the other subject: in that case the assumption may or may not hold. The evidence the authors collect points toward a valid quasi-experiment design, and thus presents itself as an attractive alternative to randomized trials when these are not feasible, and also to take advantage of the considerably larger and readily available data from administrative sources and periodic standardized tests.

CFR find no bias in value-added measures in their calculations, and are able to obtain considerably more precise estimates of the bias than the ones implemented up to that date, by taking advantage of a 7.6 million observations sample for computing the value-added measures, and roughly 60 thousand observations for the evaluation of bias in their preferred specification (Chetty et al. (2014a), Table 4, Column 1). The authors also test whether there are differences in average value-added across different socioeconomic measures, and fail to find evidence of sorting of teacher effectiveness across these various measures.

Following CFR, authors have replicated their analysis in other settings. Rothstein (2017), while questioning the validity of Chetty et al. (2014a) methodology, replicates the analysis in North Carolina, focusing only in a sample of elementary school-grades. When computing Chetty et al. (2014a) preferred specification, they find no evidence of bias. The author argues nonetheless that the quasi-experiment design is invalid, because teacher switching is correlated with changes in students prior score, what Rothstein refers to as "student preparedness". Rothstein points out as an important factor to this placebo effect the sorting that introduces the sample selection from dropping teachers observed in only one period, which also implies dropping his students altogether. He argues that this pushes $(1-B)$ where B is the bias, upward. By adjusting for this, Rothstein finds moderate bias.

Bacher-Hicks et al. (2014) replicate CFR analysis in the LAUSD. As CFR, they similarly find unbiasedness in VA estimates, and acknowledge the correlation between changes in mean prior score and changes in average VA reported by Rothstein (2014, 2017). They also test whether the predictive validity of value-added estimates changes depending on the school-teacher matches. They do so by decomposing the value-added measures into a component reflecting the information obtained from the same school, a second one from another schools, and a third one from another but considerably different (in terms of mean test scores) school. Even though the predictive validity of the portion of the value-added estimate from the same school is higher than the one from other schools (whether the whole pool of other schools or just the considerably different), they find no evidence that these different components were not equally predictive of their students' achievement. The authors also test for differences in mean VA across different socioeconomic measures, and unlike CFR, find that both within and across schools, there is student-teacher sorting on the students' socioeconomic characteristics.

The analysis the authors did is complemented in Bacher-Hicks et al. (2017) with an experimental evaluation of value-added measures, classroom observations and student surveys. Based on a prediction of teacher effects combination of different achievement measures based on test scores, classroom observations, and student surveys, they find that both score-based measures of effectiveness are unbiased, and as a novelty in the literature, the unbiasedness of value-added measures based on classroom observations ratings are unbiased when predicting a teacher's performance. The authors explore the unbiasedness of student surveys, but due to a lack of statistical power, are unable to argue unbiasedness in these estimates. In particular, the authors cannot reject that teachers' predicted effectiveness using pre-random assignment data, has a one on one effect on students' score, or teachers' classroom observation rates.

As mentionned previously, Chetty et al. (2014a) quasi-experimental method has nonetheless been questioned. Rothstein (2014, 2017) argues that this method violates a placebo test they present, where changes in mean prior scores are regressed on current changes in teachers' value-added (see also Rothstein (2009), Rothstein (2010), where the author finds evidence of fifth teacher effects influencing fourth grade scores gains). If the value-added estimates capture causal impacts of teachers, they could not possibly influence students' scores in an earlier grade. Rothstein finds that there is a significant effect of changes in average value-added estimates on mean prior scores, and argue that this results in a non-valid quasi-experiment design. Chetty et al. (2017) respond by arguing that this is caused by a mechanical effect, rather than a valid placebo test. As Chetty et al. (2017) say: "the treatment effect in this setting (VA) is endogenously estimated from data on test scores", meaning that value-added measures are computed using prior scores, which makes it an invalid test to check on pre-treatment balance. The authors present Monte Carlo simulations to show that the placebo test detects this correlation even when the research design is valid. Chetty et al. (2016) provide several theoretical arguments and simulations to argue that prior outcomes balance tests do not provide robust information on bias in value-added models, rejecting balance in lagged gains across teachers' effects, in unbiased estimates. Koedel & Betts (2011) posit that future teacher effects are smaller when they focus on teachers with multiple cohorts of students' observations. Rothstein (2017) replies with an alternative placebo test, which uses only demographic characteristics – presumably unaffected by prior teachers' effectiveness or by school-level shocks - to predict a score. He then uses this predicted score to compute mean predicted scores, which he regresses on the change in mean predicted VA, and finds an association significantly different from zero. Rothstein finally argues that this is caused by the sample selection introduced by the dropping of teachers whose VA measure is seen only one year, which causes that entire classrooms disappear from the analysis, in a non-random way. This pushes to underestimate the degree of forecast bias in the value-added estimates. Rothstein (2017) introduces that a way to correct this non-random sample selection is to impute the grand mean to teachers VA (which corresponds to the Empirical Bayes estimator imputation when there is no signal at all), and to further obtain more precise estimates, to control for the changes in mean prior scores in the quasi-experiment regressions. Chetty et al. (2017) argue that both implementations are incorrect, as they introduce bias in the estimations. The first method, because the crucial assumption that VA is independent among teachers within a school does not hold empirically. If a high (low) performing teacher leaves a school-grade cell, the replacing teacher will tend to be also a high (low) performing teacher (the authors find a correlation of 0.2 in New York data). Contrary to Rothstein, CFR argue that with a correlation of this magnitude, the bias is considerable. They also implement a Monte Carlo simulation to contrast the effect of dropping missing data, and imputing the grand mean. They find that dropping the observations leads to no bias, while the imputation procedure yields a bias of 7%. As what respects the second method, CFR argue that the changes in mean prior scores constitute a bad control: an endogenous control, because the change in mean prior scores and the change in average VA correlate as some teachers follow students across grades.

Up to this date, only three out of the six studies on bias in value-added estimates mentioned above obtained unbiased value-added measures at the middle school level. Two studies, Kane & Staiger (2008) and Chetty et al. (2014a), did not disaggregate by school level, and one (Rothstein (2017)) did not use middle school data. The studies from Kane et al. (2013) and Bacher-Hicks et al. (2014) found no bias in middle school estimates, although with larger confidence intervals than for the elementary level. The study from Glazerman & Protik (2015), present results that support biased value-added measures at the middle school level, although with less statistical precision than Kane et al. (2013).

Until very recently there were also no VA estimates of teacher effects in Chile. Recent efforts have been implemented, notably in conjunction with evaluating the effect of teacher characteristics on teachers' contribution to student achievement (Santelices et al. (2015), Taut et al. (2016), Canales & Maldonado (2018)). Canales & Maldonado (2018) implement three teachers' contribution to student performance measures: intraclass correlations, proportion of students' test score variance explained by teachers, and VA estimates. They do so by using student-level scores in 8th grade for 2011, estimating with a set of covariates that comprises student characteristics, previous student's score from four years ago, among other controls. They then use these estimates to evaluate the role of specific teacher characteristics in explaining student achievement. They find important teachers' effects, as well as a relevant role for teacher experience in teacher quality, but acknowledge that there is substantial variation not explained by observable teacher characteristics.

An important limitation of the previous studies implemented in Chile, even though they use different methods to correct for sorting, is that they do not assess if there is bias in their estimates.

We take charge of this limitation. In the next sections we estimate VA measures of teacher quality, and assess whether VA estimates in middle school in Chile are unbiased. We also explore what factors are the most relevant to ensure their unbiasedness.

Chapter 3

Empirical Implementation

We will follow the methodology implemented by Chetty et al. (2014a)[1] and their notation closely. First, we will estimate a model regressing students' test scores on prior scores[2], demographic characteristics, tuition charged (*copago*), teacher assignment history, and teacher fixed effects, as following:

$$A_{it}^{*} = \beta X_{it} + \mu_j + \varepsilon_{it} \tag{3.1}$$

Where A_{it}^{*} corresponds to the raw score test for student i in period t, X_{it} includes student's prior test scores, demographic characteristics, *copago*, and teacher assignment history. μ_{jt} is a teacher fixed effect. We then compute the residuals plus the absorbed teacher effects from that equation:

$$A_{it} = A_{it}^{*} - \widehat{\beta} X_{it} \tag{3.2}$$

And proceed to average them to the classroom-year level for each teacher, to obtain a raw measure of teacher quality for that year. To account for the precision in potentially noisy estimates (as class and student's information vary across different teachers), we compute precision-weighted averages of classroom-average scores within a teacher-year:

$$\overline{A_{jt}} = \sum_{c \in j(c)=j} h_{ct} \overline{A_{ct}} \tag{3.3}$$

Where the weight for classroom c in year t is:

[1] We use Chetty et al. (2014c) and borrow on code from Rothstein (2017) to implement Chetty et al. (2014a) analysis.

[2] A limitation of our estimations we acknowledge is the fact that prior scores are either from two years ago for the tightest gap, or from four years ago. There is no availability of previous year scores.

$$h_{ct} = \frac{1}{\widehat{\sigma}_{\theta}^{2} + \frac{\widehat{\sigma}_{\varepsilon}^{2}}{n_{ct}}} \tag{3.4}$$

h_{ct} corresponds to the precision for the classes taught by a teacher in year t, such as $\widehat{\sigma}_{\theta}^{2}$ is an estimate of the class-level variance,[3] n_{ct} is the number of students in the classroom, and $\widehat{\sigma}_{\varepsilon}^{2}$ is the individual-level variance of residual test scores $Var(\varepsilon_{it})$.

We then estimate the covariances among average scores across years for a same teacher, allowing a different covariance for each possible time lag. Each teacher-year measure is weighted by the number of students taught. The interest for estimating the covariances between lags, is to allow for drift in value-added measures, by allowing different weights on the different teacher-year mean scores. This permits us the flexibility to estimate a teacher effect that is not fixed across years by construction.

We finally estimate teacher's j value-added measure in t, μ_{jt} as the best linear predictor of A_{jt} based on scores from all years except t. Considering that $\vec{A}_{j}^{-t}$ is a vector containing the mean scores at the teacher-year level used to estimate μ_{jt}, the best linear predictor corresponds to:

$$\widehat{\mu}_{jt} = \psi \vec{A}_{j}^{-t} \tag{3.5}$$

Where $\psi = \sum_{A_{jt}}{}^{-1} \gamma_{jt}$, such as $\sum_{A_{jt}}$ corresponds to the precision-adjusted variance-covariance matrix of $\vec{A}_{j}^{-t}$, and γ_{jt} is the vector of auto-covariances between the mean test scores taught in t by teacher j and the mean test scores of the same given teacher j in all other periods but t. By using the data from all other years than t, precision is increased compared to using only data from previous periods. Both items in ψ vary across j and t, that's why we construct them across each teacher-year data.

Notice that this measure is a leave-one-out (jackknife) estimator of teacher year effect. We estimate the teacher's j effect in the classes taught in year t by using a weighted combination of the precision-weighted mean residuals of a teacher from all the years but t, thus allowing for drift in teacher quality measure, and reducing attenuation in the estimates.

This measure is the best linear predictor of actual scores, $\overline{A_{jt}}$, of a teacher based on score data of his students in all the other years, $\vec{A}_{j}^{-t}$. Though $\widehat{\mu}_{jt}$ is an unbiased prediction of $\overline{A_{jt}}$, it is not clear that $\widehat{\mu}_{jt}$ is an unbiased estimate of μ_{jt}, the teacher true causal effect (Rothstein (2017)). This paper adds to the current empirical evidence to argue whether $\widehat{\mu}_{jt}$ is an unbiased estimator of μ_{jt}, and if biased, to what extent. We describe in the next section the data used in these calculations.

[3] This class-level variance corresponds to $\widehat{\sigma}_{\theta}^{2} = Var(A_{it}) - \widehat{\sigma}_{\varepsilon}^{2} - \widehat{\sigma}_{A0}$, where $\widehat{\sigma}_{A0}$ is the within-teacher-year between-class covariance in average scores.

Chapter 4

Data

We use for our implementation data records available upon request from the Ministry of Education of Chile. Test scores correspond to the SIMCE test results in Reading (*Lenguaje y Comunicación*) and Math (*Matemáticas*). SIMCE is a battery of standardized tests administered to all students in some of the following grades: 2nd , 4th , 6th , 8th and 10th grade, depending on the year the test is taken.[1] For purposes of this estimation, we only consider middle school students: the score of the 6th or 8th grade student as actual score, and the student's score in 4th grade as previous score. SIMCE tests are administered alongside a set of questionnaires asked to the student's teacher in that subject, to his parents, and in recent years to the student himself. We use information of the teachers' and on the parents' questionnaire, where socioeconomic and demographic information is available. We also make use of the Ministry of Education teacher assignment database, to link each school-classroom-subject-year to a teacher. This database doesn't uniquely identify one teacher per school-classroom-subject-year, so we apply a recursive algorithm where teachers are classified according to their subjects' code to one of two groups: Reading teacher, or Math teacher. As there may two or more teachers assigned to a same school-classroom-subject-year, we sort the different subjects' codes in term of absolute frequencies, where the most common are placed first. We then assign to a school-classroom-subject-year the first teacher match in the list of subjects' codes of each subject group, and continue throughout each subject code in the list, until all teachers through each subject code in each subject group are assigned to a school-classroom-subject-year.

We use SIMCE scores from years 2009, 2011, 2013, 2014 and 2015, depending on the sample used. Test scores are standardized, to have a mean of 0 and a standard deviation of 1, by subject-year.

[1] For Math and Reading tests.

Our main analytical sample consists of the 6th grade students, from all schools for which we have data on the tuition charged to parents (*copago*), except for private non-subsidized schools. We also present results for a combined sample of 6th and 8th grades.[2] Years available (and considered) for calculations on 6th grade only samples are 2013, 2014 and 2015, and for our combined (6th and 8th grades) sample, years available are 2009, 2011, 2013, 2014 and 2015.

TABLE 1: SUMMARY STATISTICS FOR SAMPLE USED TO ESTIMATE VALUE-ADDED MODEL, 6TH GRADE ONLY, NO PRIVATE SCHOOLS, ALL YEARS

Variable	Mean	SD	Observations
	(1)	(2)	(3)
Class size (not student-weighted)	26.7	7.9	15822
Number of subject-school years per student	1.9	0.3	221472
Test score (SD)	0.24	0.9	423233
Age (years)	11.2	0.4	423233
Female	51.5%		423233
Repeating grade	2.6%		423233
Special education	0.01%		423233
Household income	589,106	477,953	361522
Household education	13	3	342641
Copayment	21,317	20,764	423233

Notes: All statistics reported are for the sample used in estimating the baseline value-added model. This sample includes only students who have non- missing lagged test scores and other requisite controls to estimate the VA model. Number of observations is number of classrooms in the first row, number of students in the second row, and number of student-subject-year observations in all other rows. Student data are from the administrative records of Chile. Test score is based on standardized scale scores. Parent income is the yearly household income. For parents who do not file, household income is defined as zero. Household education is the average parents' education in the household.

In our main sample, mean class size has 26.7 students, with a standard deviation of 7.9; students have on average 1.9 subject-schools years; mean score is 0.24 and its SD is 0.9; students have on average 11.2 years, and 51.5% of them are females.[3] Repeating students represent 2.6% of our sample, and only 0.01% are special education students. Average household income is of 589,106 *chilean pesos* (CLP),

[2] We also implement value-added estimations and quantify their bias in the next - not shown - specifications: 8th grade only, no private schools, without *copago* in the VA estimation; 6th and 8th grades combined, no private schools, no *copago* in the VA estimation; 6th grade only, no *copago*, all schools; 6th grade, no private schools, VA estimation with school fixed effects; 6th grade, no private schools, VA estimation with school dependency type fixed effect (*tipo de dependencia*). All of these value-added estimates are biased, some considerably higher than the others, but none with less than 40% of bias. These results are available upon request.

[3] Standarized scores are at subject-year level, across all schools in Chile, not just the ones considered for our estimation purposes. In our main estimates we focus on schools with data on *copago* available, which consists mainly of subsidized private schools. When considering all schools, mean score is close to 0, as expected.

with a standard deviation of 477,953 CLP. Descriptive statistics for the 6th grade full sample is available in Table 8 in Appendix A,[4] as are descriptive statistics for 8th grade and other samples, from Table 8 through Table 11 (Appendix A).

[4] Sample that consider all schools, and is not restricted to availability of the different variables.

Chapter 5

Value-added estimates

To implement our estimates, following Chetty et al. (2014a) and depending on availability of data, we use as controls a cubic polynomial in twice lagged score (and lagged four times scores in some specifications) in the same and in the other subject, and interact them with grade; student's gender; age; absences percentage; indicators for special education; and grade repetition. We consider class and school-grade means in all the covariates, and additionally cubic polynomials of prior scores in both subjects, interacted (separately) with grade; class size; and grade, year, and rurality of the school indicators. As previous studies have shown substantial socioeconomic stratification in the chilean voucher system (*e.g.* Mizala & Torche (2012)), we also include *copago* (tuition charged) in the control vector interacted with year dummies. Some ways that tuition charged may impact a student score would be by reflecting the availability of school resources or a higher overall disposition to spend on education from the parents.

Given that our previous score measure is not from the prior year but from two years ago (or four years ago for our 8th grade students), we consider the history of teacher assignment between the previous score, and the current year score. We do so by incorporating a teacher assignment indicator for these in-between years for each student.

We proceed next to estimate the auto-covariances of the mean test score residuals from a teacher in a class taught in a year t, and in all years other than t, and compute the respective autocorrelations. These estimates can be seen in Table 2 for our main sample (6th grade, no private schools, all years sample).

There is a decay between the first and second lag. This continues as we go further apart: Figure 5 in Appendix A presents the plotted autocorrelation vector for the combined sample (6th and 8th grades), where we have four lags.[1] We corroborate a decay in autocorrelations, and for both figures the autocorre-

[1] Contrary to our main sample, where we only have two lags.

TABLE 2: TEACHER VALUE-ADDED MODEL PARAMETER ESTIMATES, 6TH GRADE, SCHOOLS WITH *copago* ONLY, ALL YEARS

Sample	(1) Reading	(2) Math
Lag 1	0.025	0.050
	(0.002)	(0.002)
	[0.365]	[0.637]
Lag 2	0.013	0.045
	(0.002)	(0.003)
	[0.235]	[0.567]
Total variance	0.476	0.383
Individual-level variance	0.418	0.310
Class variance	0.022	0.007
Teacher variance	0.036	0.065

Notes: Table 2 reports the estimated autocovariance, the standard error of that covariance estimate clustered at the teacher level (in parentheses), and the autocorrelation (in brackets) of average test score residuals between classrooms taught by the same teacher. We measure these statistics at time lags ranging from one (i.e., two classrooms taught one year apart) to two lags (i.e., two classrooms taught two years apart), weighting by the sum of the relevant pair of class sizes. Row 7 reports the raw variance of test score residuals and decomposes this variation into components driven by idiosyncratic student-level variation, classroom shocks, and teacher-level variation. We estimate the variance of teacher effects as the covariance of mean score residuals across a random pair of classrooms within the same year. Years considered are 2013, 2014 and 2015.

lation seems to be converging, replicating the possibility of a transitory and permanent component noted by CFR. We are nonetheless unable to approach the permanent component in our data due to limited lags availability. Though we don't have as many lags available as previous literature (Chetty et al. (2014a), Rothstein (2017), Bacher-Hicks et al. (2014)), these results are notwithstanding in accordance with their findings, although with higher autocorrelation coefficients for Math and Reading for our main sample (Table 2) as well as the 6th and 8th grades sample combined (Table 12, Appendix A), compared to Chetty et al. (2014a).

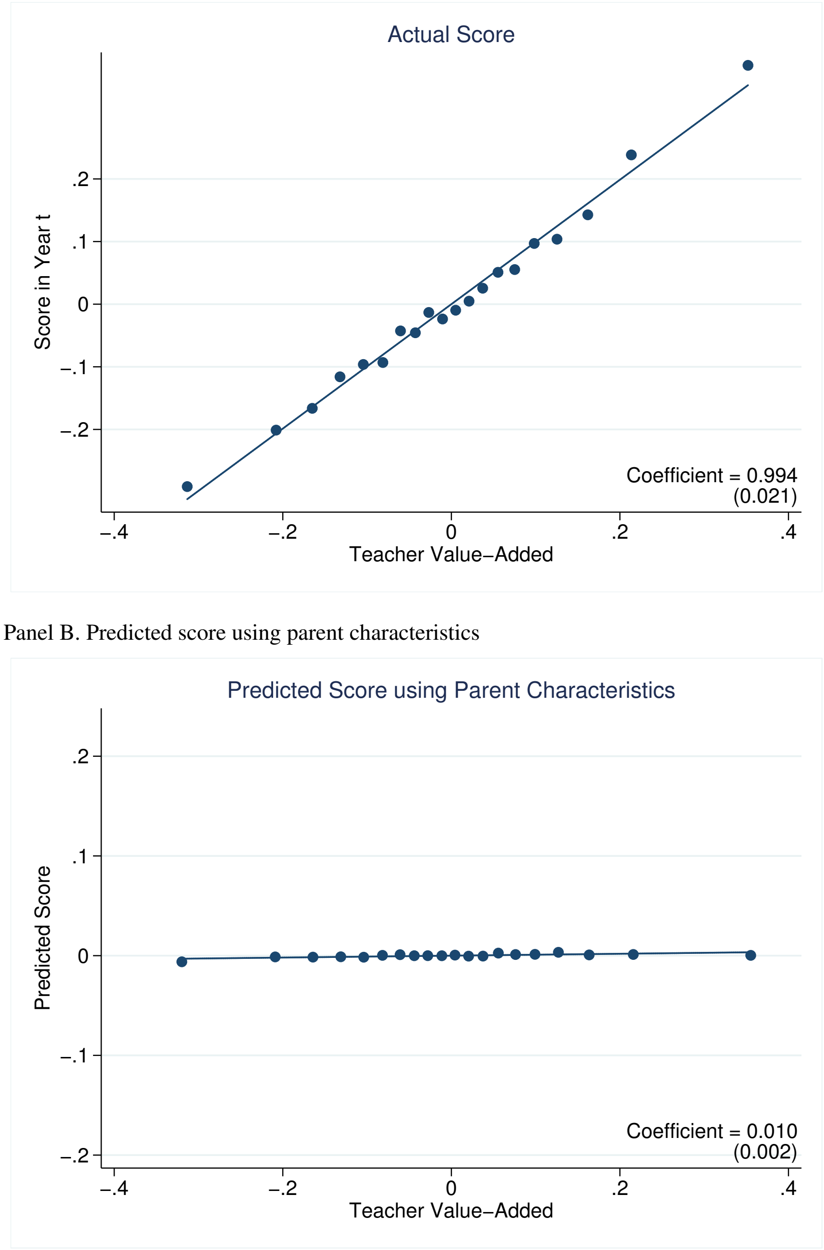
As our calculations are based on middle school data, the availability of multiple classes per teacher per year allows us to compute the teacher's within-year covariance of mean test score residuals σ_{A0}, which corresponds to the variance of teacher effects σ_{μ}^2, if class and student level shocks are independent and identically distributed. Our variances of teacher effects are considerably larger than those found by CFR (0.036 for Reading versus 0.01 for CFR; 0.065 for Math versus 0.018 for CFR), as are the class level variances. Unlike CFR, we find that teacher variance is more important in explaining the scores than class level variance for both subjects. Individual variance is in our results the main component of variance, explaining a lower share of total variance than in CFR. Total variance is also higher than in CFR.

We compute conditional variances to explore differences among groups of measures traditionally associated with teacher quality, such as teacher experience or qualifications. These results can be seen from Table 13 through Table 20 in Appendix A. We can see that the variance of teacher effects explains a higher share of total variance for teachers with master or PhD vis-a-vis teachers with no such qualification. We cannot say the same for any kind of post-secondary education beyond college[2], where the share of total variance explained by the teacher variance is roughly the same between no post-secondary education, and with some post-secondary education, given the college degree. When conditioning on teacher experience quartiles, we see that this share has a non linear relationship on experience: this relation seems concave for Math scores, with an increasing share until the third quartile (13.8%, 14.6%, 14.7% respectively), but considerably lower for the fourth one (12.78%). In the case of Reading scores, this relationship is highly non linear, with the lower teacher variance shares of total variance for the first (8.1%) and third (7.49%) quartile, slightly higher than these ones for the fourth (7.66%), and the highest for the second quartile (9.27%).

Finally, along with the test score residuals and the autocovariances, we proceed to compute the value-

[2] This category master and PhD, but also continuing education, skill enhancing programs, and in general any kind of certified learning activities.

Figure 1: Effects of Teacher Value-Added on Actual and Predicted Scores, 6TH grade, schools with *copago* only, all years



Panel A. Actual score

Panel B. Predicted score using parent characteristics

Notes: These figures pool all grades and subjects and are constructed using the sample used to estimate the VA model, which has one observation per student-subject-school year. The two panels are binned scatter plots of actual scores and predicted scores based on parent characteristics versus teacher VA. These plots correspond to the regressions in columns 1 and 2 of Table 3 and use the same sample restrictions and variable definitions. To construct these binned scatter plots, we first residualize the y-axis variable with respect to the baseline control vector separately within each subject by school- level cell, using within-teacher variation to estimate the coefficients on the controls. We then divide the VA estimates into 20 equal-sized groups (vingtiles) and plot the means of the y-variable residuals within each bin against the mean value of the VA estimates within each bin. The solid line shows the best linear fit estimated on the underlying micro data using OLS. The coefficients show the estimated slope of the best- fit line, with standard errors clustered at the school-cohort level reported in parentheses.

TABLE 3: ESTIMATES OF FORECAST BIAS USING PARENT CHARACTERISTICS, 6^{TH} GRADE ONLY, SCHOOLS WITH *copago* ONLY, ALL YEARS

Sample	(1) Score in year *t*	(2) Pred. score using parent chars.	(3) Score in year *t*	(4) Pred. score using parent chars.	(5) Pred. score using parent chars.
Teacher VA	0.994	0.010	0.988	0.005	0.007
	(0.021)	(0.002)	(0.021)	(0.002)	(0.002)
Parent chars. controls			X		
School Fixed Effects				X	
Copago					X
Obsevations	289200	231784	231784	231784	231784

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort 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-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 using the baseline control vector, which includes: a cubic in lagged own- and cross-subject scores, interacted with the student's grade level; student-level characteristics including gender, age, lagged absences, and indicators for grade repetition, special education; class size and class-type indicators; cubics in class and school-grade means of lagged own- and cross-subject scores, interacted with grade level; class and school-year means of all the student-level characteristics; *copago* and grade and year dummies. When prior test scores in the other subject are missing, we set the other subject prior score to zero and include an indicator for missing data in the other subject interacted with the controls for prior own-subject test scores. In columns 1 and 3, the dependent variable is the student's test score in a given year and subject. In columns 2, 4 and 5, the dependent variable is the predicted value generated from a regression of test score on indicators of household income and parents' education, after residualizing all variables with respect to the baseline control vector.

added estimates. Distributions of VA estimates can be seen in Appendix A (Figure 4, Appendix A). Notice that the standard deviation of VA estimate is considerably smaller in Reading than in Math (0.108 in Reading, 0.183 in Math, Figure 4, Appendix A), slightly larger than the ones found in Bacher-Hicks et al. (2014) for our main sample, though smaller for Math in the 6th and 8th combined sample, and about twice the dispersion found in Chetty et al. (2014a). Both are smaller than true teacher effects (Table 2),[3] due to the shrinkage in the estimates.

Figure 1, Panel A, shows the relation between students' score, A_{it}, and teacher VA estimates, $\widehat{\mu}_{jt}$, controlling by subject fixed effects. As the VA estimate of teacher j corresponds to the best linear prediction of mean students' score in year t of teacher j, the teacher VA estimate should yield a coefficient of 1, *i.e.* a one-higher standard deviation teacher VA increases a student's test score by one standard deviation. The coefficient obtained from our preferred sample is below 1, but within the 5% confidence interval ($0.994 \pm 1.96 \times 0.021$). Coefficient obtained from this regression is closer to 1 in the combined (and larger) sample (coefficient of 1.002, column 1 of Table 21, Appendix A).

We further explore if there is bias on excluded observable characteristics, such as parents characteristics. Bias is identified under this approach if students are sorted to teacher based only on these excluded observables. We regress the predicted - from the vector of unexplained by the baseline control X_{it} and teacher fixed effects, share of parent characteristics, residual scores (A^{p}_{it}) - on the value-added estimates $\widehat{\mu}^{-t}_{jt}$ (where the superscript denotes the year excluded in the estimate calculation), including subject fixed effects. The coefficient obtained from this regression corresponds to:

$$B_p = \frac{cov(A^{p}_{it}, \widehat{\mu}^{-t}_{jt})}{Var(\widehat{\mu}^{-t}_{jt})} \tag{5.1}$$

Where B_p is the coefficient from the feasible estimation. CFR find a degree of bias from selection on parent characteristics of 0.2 percent. In our case the degree of bias is considerably higher, of 1% (Table 3, Column 2). The result is presented graphically in Figure 1, Panel B. This coefficient is however not significantly different from 0 in the combined sample, where it is very precisely estimated (Table 21, Column 2, Appendix A). When exploring bias within schools (Table 3, Column 4), bias is reduced at half (0.5%). Bias increases to 0.7% when considering *copago* fixed effects (Table 3, Column 5), but is also significant only at 5%, not at 1%. Similar bias calculations are computed in Table 21 (Appendix A) for the combined sample. Within schools, there is no bias on parent characteristics (Column 4), but there is a small bias of 0.1% within *copago* levels (Column 5). These results show us that for value-added

[3] Teacher effects in Table 2 are in terms of variance.

estimates that control for our full baseline vector, bias on parent characteristics is higher in Chile than in other settings as in New York City, and this parent characteristics-school sorting is not fully captured by the amount of tuition charged, in a setting where school-family matches is importantly determined by a comprehensive socioeconomic sorting, beyond the tuition charged by the school.

We next proceed to explore whether the relationship between A_{it} and $\widehat{\mu}_{jt}^{-t}$ is causal or biased by sorting on unobservables.

Chapter 6

Quasi-experimental estimates of bias

Before introducing the quasi-experiment, we first define bias in value-added estimates. Following Chetty et al. (2014a), we define the following regression on observational data:

$$A_{it} = \alpha_t + \lambda A_{jt} + \chi_{it} \tag{6.1}$$

Where we consider momentarily $E[A_{it}/1, \widehat{\mu}_{jt}] = \alpha_t + \lambda A_{jt}$. Notice that $A_{it} = \mu_{jt} + \varepsilon_{it}$. As a regression of A_{it} on $\widehat{\mu}_{jt}$ yields a coefficient of 1 by construction,

$$\lambda = \frac{cov(A_{it}, \widehat{\mu}_{jt})}{Var(\widehat{\mu}_{jt})} = \frac{cov(\mu_{jt}, \widehat{\mu}_{jt}) + cov(\varepsilon_{it}, \widehat{\mu}_{jt})}{Var(\widehat{\mu}_{jt})} = 1 \tag{6.2}$$

Then the bias corresponds to the extent of which teachers are sorted to students due to unobservables, ε_{it}. This corresponds to

$$B(\widehat{\mu}_{jt}) = 1 - \lambda = \frac{cov(\varepsilon_{it}, \widehat{\mu}_{jt})}{Var(\widehat{\mu}_{jt})} \tag{6.3}$$

The key problem is to evaluate this parameter, $B(\widehat{\mu}_{jt})$. Chetty et al. (2014a) devise a quasi-experiment to evaluate this bias, based on changes in teaching staff. Teacher turnover at the school-grade-subject level provides a plausibly exogenous mean value-added change of the teaching staff across cohorts within a school-grade, given that the change in these teachers' VA is orthogonal to changes in other determinants of students' scores. More formally, let $\Delta Q_{sgt} = Q_{sgt} - Q_{sg,t-1}$ be the change in mean teacher value-added from year $t-1$ to t in grade g in school s, where this time VA estimates are constructed leaving out not only the current year t, but also $t-1$. This applies both for the VA estimates used to compute Q_{sgt} as for the ones used to compute $Q_{sg,t-1}$, *i.e.*, we use $\widehat{\mu}_{jt}^{-t,t-1}$, and $\widehat{\mu}_{j,t-1}^{t,t-1}$, where the superscript denotes the years excluded in the estimate calculation. The reason for this is to avoid a systematic correlation between the dependent variable ΔA_{sgt}, and ΔQ_{sgt}, in the next regression:

$$\Delta A_{sgt} = a + b\Delta Q_{sgt} + \Delta\chi_{sgt} \tag{6.4}$$

(9)

Where $\Delta A_{sgt} = A_{sgt} - A_{sg,t-1}$ corresponds to the change in mean residual scores, and A_{sgt} to the average of A_{it} within a school s, grade g, year t cell. If year t and year $t-1$ were not excluded in the construction of VA estimates used in the right-hand variables, we would have a simultaneity problem. This way, we ensure that the variation of value-added is driven by changes in the teaching staff. The key assumption to identify bias in value-added measures is:

$$cov(\Delta Q_{sgt}, \Delta\chi_{sgt}) = 0 \tag{6.5}$$

In this case $b = \lambda = 1 - B(\widehat{\mu}_{jt}^{-t,t-1})$, where B corresponds to the magnitude of the bias in the VA measure. Following Chetty et al. (2014a), we implement some tests to evaluate the validity of the key assumption.

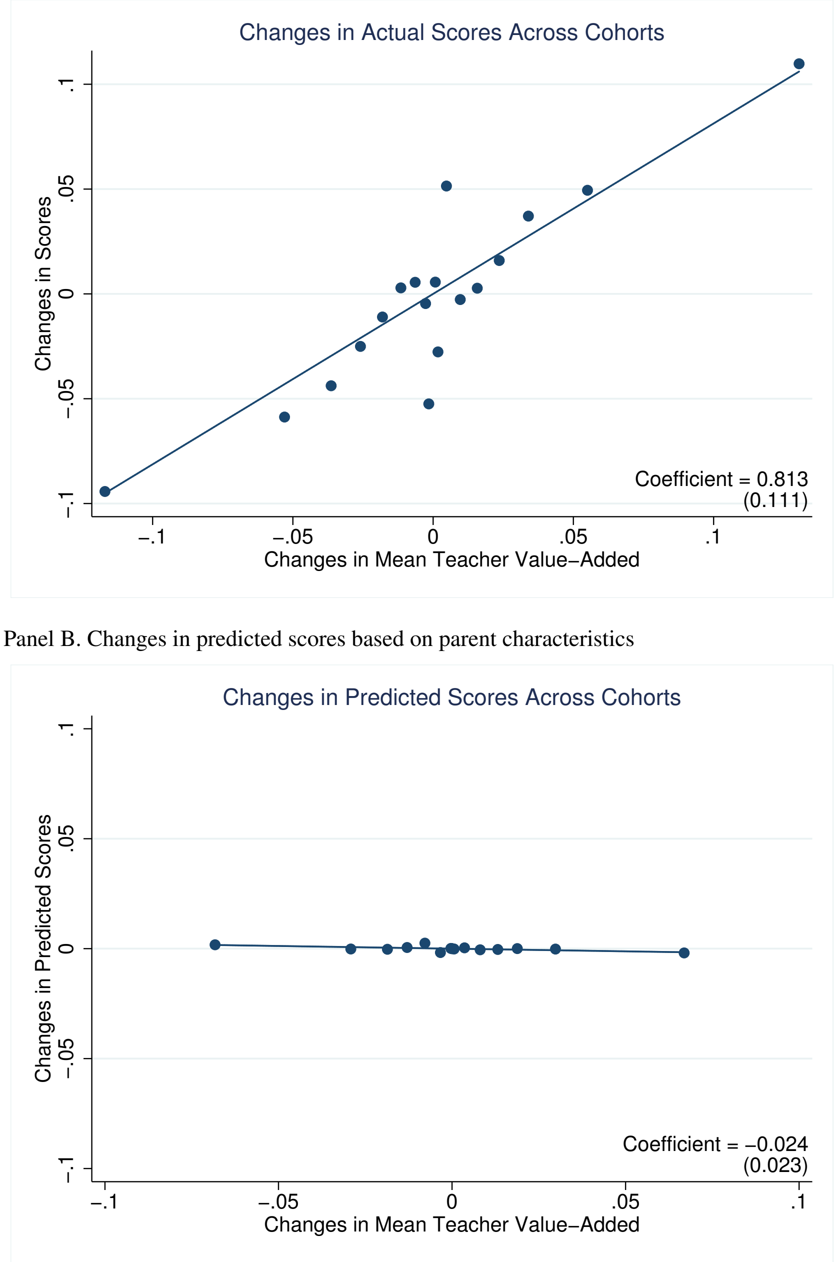
As a first approach, we look to evaluate whether there is a linear relationship between changes in mean value-added estimates with changes in mean parent characteristics. In particular, we use predicted raw scores based on parent characteristics, by regressing A_{it} on P_{it}^{*}, obtaining the prediction from this regression, $\widehat{A}_{it}^{*P}$. We then run the regression $\Delta\widehat{A}_{sgt}^{*P} = a^{*} + b^{*}\Delta Q_{sgt} + \Delta\chi_{sgt}^{*}$. Results are displayed in column 3 of Table 4, and plotted nonparametrically in Figure 2, Panel B. We cannot reject that $\widehat{b}^{*}$ is equal to 0, thus providing evidence that changes in mean value-added are uncorrelated with changes in student characteristics.

Second, we evaluate if there is a correlation with changes in mean raw scores in the other subject (Table 4, Column 4). By regressing changes of mean scores in the other subject on changes in mean value-added estimates, we obtain a coefficient non statistically different from 0, that can be seen nonparametrically in Figure 3. This result would support the validity of the quasi-experiment. In Appendix B we run some robustness estimates for this test.

Considering that our evidence either supports the validity of the key assumption or doesn't invalidate it, we proceed to estimate the bias in value-added estimates with the quasi-experiment. In this case, violations of the key assumption should be driven by unobserved determinants unrelated to parent

Figure 2: Effects of Changes in Teaching Staff on Scores across Cohorts, 6th grade only, schools with *copago* only, all years

Panel A. Changes in actual scores



Panel B. Changes in predicted scores based on parent characteristics

Notes: This figure plots changes in average test scores across cohorts versus changes in average teacher VA across cohorts. Panel A is a binned scatterplot of changes in actual scores versus changes in mean VA, corresponding to the regression in column 1 of Table 4. Panel B is a binned scatterplot of changes in predicted scores based on parent characteristics versus changes in mean VA, corresponding to the regression in column 3 of Table 4. See notes to Table 4 for details on variable definitions and sample restrictions. Both panels are plotted using the core sample collapsed to school-grade-subject-year means. To construct this binned scatterplot, we first demean both the x- and y-axis variables by school year to eliminate any secular time trends. We then divide the observations into 20 equal-size groups (vingtiles) based on their change in mean VA and plot the means of the y variable within each bin against the mean change in VA within each bin, weighting by the number of students in each school- grade-subject-year cell. The solid line shows the best linear fit estimated on the underlying microdata using a weighted OLS regression as in Table 4. The coefficients show the estimated slope of the best-fit line, with standard errors clustered at the school- cohort level reported in parentheses.

Table 4: Quasi-Experimental Estimates of Forecast Bias, 6th grade only, schools with *copago* only, all years

	(1)	(2)	(3)	(4)
			Δ predicted	Δ other
	Δ score	Δ score	score	subject score
Changes in mean teacher	0.813	0.507	-0.024	0.121
VA across cohorts	(0.111)	(0.278)	(0.023)	(0.118)
Year fixed effects	X			X
School × year fixed effects		X	X	
Other-subject change in mean teacher VA				X
Grades	6	6	6	6
Number of school × grade × subject × year cells	3984	3984	3933	2690

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on a dataset containing school-grade-subject-year means, excluding classrooms in which we cannot construct the leave-two-year-out VA estimate described below. All regressions are weighted by the number of students in the school-grade-subject-year cell. Changes in mean teacher VA across consecutive cohorts within a school-grade-subject cell as follows: first, we calculate teacher VA for each teacher in a school-grade-subject cell in each adjacent pair of school years using information excluding those two years. We then calculate mean VA across all teachers, weighting by the number of students they teach. Finally, we compute the difference in mean teacher VA (year t minus year t - 1) to obtain the independent variable. The dependent variables are defined by calculating the change in the mean of the dependent variable (year t minus year t - 1) within a school-grade-subject cell. In columns 1–2, the dependent variable is the change in mean test scores within subject (Reading or Math). In column 3, it is the change in the predicted score, constructed based on parental characteristics. In columns 4, the dependent variable is the change in the score in the other subject (e.g., Math scores for Reading teachers). Column 1 includes only year fixed effects and no other controls. Columns 2 and 3 include school-by-year fixed effects. Columns 4 control for the change in mean teacher VA in the other subject as well as year fixed effects.

characteristics, and change differentially between school-grade cells at annual frequency.

Results are presented in Table 4. All specifications in Table 4 pool three sources of variation of teachers' value-added across school-grade-years: entries to school, exits of school, and switches between grades within schools.[1] In Column 1, we estimate the regression in equation (6.4) including year fixed effects. This relation can be seen nonparametrically in Figure 2, Panel A. We obtain a non significant 18.7% of bias in this estimation. Though unbiased, the corresponding higher bound of the confidence interval is considerably high, with about 40.5% bias. Column 2 runs a specification with school-year fixed effects, to address the possibility that changes in teacher quality are related to improvements in a school that affect test scores, and thus biasing our biases estimates if not considered. This time, bias augments considerably and is significantly different from 0, with a point estimate of $(1-0.507) = 49.3\%$, but considerably less precisely estimated, with a confidence interval of roughly $49.3\% \pm 54.5\%$, making it potentially completely biased. Identifying variation is reduced compared to year fixed effects only, which can be the source of the bias increase in this measure. These bias are considerably more important than those found in Chetty et al. (2014a), whose confidence interval (smaller than the one found in our study) includes the value of the coefficient equal to 1 (no bias). Despite this level of bias (or unbiasedness), when estimating without considering *copago* in the value-added estimation, bias augments to a significantly different from zero 38.7% (Table 22, Appendix A). Bias is less when we consider *copago* in the estimation (Table 23), but is only insignificantly different from 0 when we restrict the sample to the schools where data on tuition is available only.

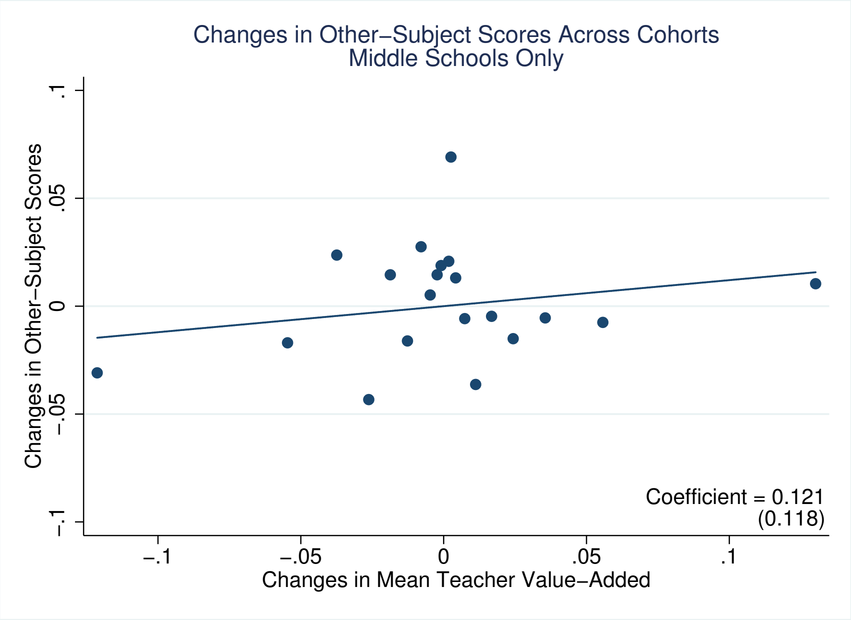
We run some robustness checks following Chetty et al. (2014a) in Table 5. All estimations in Table 5 include year fixed effects. In Table 5, Column 1, we only consider variation in teachers' value-added that stems from school exits. This variation is less likely to be correlated with changes in unobserved students' characteristics that influence scores. We isolate this source of variation by instrumenting ΔQ_{sgt} with the fraction of students taught by school-leaving teachers in the previous cohort, multiplied by the mean VA of these school-leavers. Point estimate of bias in this case is 0.436 with a large confidence interval of $0.436 \pm 1.96 \times 0.229$.

The next columns evaluate to what extent there is selection bias due to sample selection. Previous estimations discarded classrooms whose teachers had missing VA estimates, but this procedure excludes a nonrandom subset of classrooms.

CFR implement different approaches to tackle this. Column 2 provides an imputation of the uncon-

[1] In Table 4, as well as in Tables 5, 6 and 28, and also their equivalent with the combined sample, all scores used in the dependent variable are raw scores instead of residual scores, unless stated otherwise.

FIGURE 3: EFFECTS OF CHANGES IN TEACHING STAFF ON SCORES IN OTHER SUBJECT, 6TH GRADE ONLY, SCHOOL WITH *copago* ONLY, ALL YEARS



Notes: This figure plots changes in average test scores in the other subject across cohorts versus changes in average teacher VA, controlling for changes in other-subject VA, corresponding to the regression in column 4 of Table 4. See notes to Table 4 for details on variable definitions and sample restrictions. The panel is plotted using the core sample collapsed to school-grade-subject-year means. To construct this binned scatterplot, we first regress both the x- and y-axis variables on changes in mean teacher VA in the other subject as well as year fixed effects and compute residuals, weighting by the number of students in each school- grade-subject-year cell. We then divide the x residuals into 20 equal-size groups (vingtiles) and plot the means of the y residuals within each bin against the mean of the x residuals within each bin, again weighting by the number of students in each school-grade-subject-year cell. The solid line shows the best linear fit estimated on the underlying microdata using a weighted OLS regression as in Table 4. The coefficient shows the estimated slope of the best-fit line, with standard errors clustered at the school-cohort level reported in parentheses.

TABLE 5: QUASI-EXPERIMENTAL ESTIMATES OF FORECAST BIAS: ROBUSTNESS CHECKS, 6TH GRADE ONLY, SCHOOLS WITH *copago* ONLY, ALL YEARS

	(1) Teacher exit only Δ score	(2) Full sample Δ score	(3) <25 percent imputed VA Δ score	(4) 0 percent imputed VA Δ score
Changes in mean teacher	0.564	0.575	0.900	0.892
VA across cohorts	(0.229)	(0.053)	(0.137)	(0.146)
Year fixed effects	X	X	X	X
Grades	6	6	6	6
Number of school × grade × subject × year cells	3984	17098	3183	3148

Notes: Each column reports coefficients from a regression with standard errors clustered by school-cohort in parentheses. The regressions are estimated on a dataset containing school-grade-subject-year means. The dependent variable in all specifications is the change in the mean test scores (year t minus year t - 1) within a school-grade-subject cell. The independent variable is the change in mean teacher VA across consecutive cohorts within a school-grade-subject cell; see notes to Table 4 for details on the construction of this variable. All regressions are weighted by the number of students in the school-grade-subject-year cell and include year fixed effects. In column 1, we report 2SLS (two-stage least-squares) estimates, instrumenting for changes in mean teacher VA with the fraction of students in the prior cohort taught by teachers who leave the school multiplied by the mean- VA among these school-leavers. Columns 2–4 replicate the specification in column 1 of Table 4, varying the way in which we handle classrooms with missing teacher VA. Column 2 includes all classrooms, imputing the sample mean VA to classrooms with missing teacher VA. Column 3 replicates column 2, excluding entire school- grade-subject-year cells in which more than 25 percent of student observations have missing teacher VA. Column 4 restricts to entire school-grade-subject-year cells with no missing teacher VA.

ditional grand mean to VA measures, whenever $\widehat{\mu}_{jt}^{-t,t-1}$ is missing. Chetty et al. (2017) demonstrate the bias in the estimation of λ potentially caused by this procedure: unless VA among teachers in a same school is uncorrelated, this imputation introduces measurement error, and will lead to attenuation bias. The coefficient in this regression is considerably smaller than the one in column 1 of Table 4, in line with a positive correlation in VA among teachers in a same school. Bias is therefore more important than our main estimate (Table 4, Column 1).

We implement in columns 3 and 4 of Table 5 specifications that restrict the sample to school-grade-subject-year cells: where less than 25% of the teachers have their VA estimates imputed by the methodology above in both the current and previous years for Column 3, and we restrict our sample to only to school-grade-subject-year cells where there is no missing teacher-year estimates in Column 4. In both specifications, complete cohorts are excluded, not individual classrooms for some cohorts, avoiding the sample selection bias problem. Restricting to school-grade-subject-year cells where no classes are missing should eliminate sample selection bias. Indeed, bias is reduced importantly as we restrict the sample to less imputed cells, as in Chetty et al. (2014a): Column 3 shows a point bias of 10%, and Column 4 a bias of 10.8%. Even though the bias sample with no imputed VA score is 0.8% higher than the one with less than 25% imputed, both estimates are insignificantly different from 0%, and considerably smaller than the fully imputed sample.

We further analyse the bias in these specifications, when the dependent variable is the mean changes in residual 6th scores. This is the correct dependent variable if the key assumption *doesn't* hold. These results can be seen in Table 25 (Appendix A). Compared to the preferred sample estimates, these results present more bias, through specifications 2 and 3.

Specification 4 in Table 25 shows an important result. Once we restrict the sample to entire school-grade-subject-year cells with no missing teacher VA, the coefficient obtained in this specification is roughly the same point estimate than the one obtained from our preferred sample, as is the confidence interval: when we impose this sample restriction, the quasi-experiment estimation gives about the same results whether we regress the change in raw test scores ΔA_{sgt}^{*} on ΔQ_{sgt}, or by regressing the change in mean residual scores, ΔA_{sgt}, on ΔQ_{sgt}.[2] When we restrict the sample to classrooms with non-missing VA estimates for teachers, we don't use the entire school-grade-subject-year cell for identification. In that case, if assignment to classrooms is not as good as randomly assigned, the identifying assumption is violated. When the key identifying assumption holds, *i.e.*, when observable characteristics X_{it} are also

[2] This coefficient is the exact same coefficient for the raw score and for the residual score estimation for specifications without teacher history, not reported.

orthogonal to changes in teacher quality across cohorts, estimation from the former or the latter should give the same result. We see that the coefficient from the regression of ΔA^*_{sgt} on ΔQ_{sgt} is about the same with the one obtained of regressing A_{sgt}, on ΔQ_{sgt} *only* for the specification where the sample is restricted to sample to school-grade-subject-year cells with no missing VA in the current and preceding year (see Tables 5 and 25). This result provides further evidence to support the validity of the quasi-experiment devised by Chetty et al. (2014a).

We proceed next to replicate Chetty et al. (2014a) analysis to apprehend what controls are most relevant to the unbiasedness of value-added estimates. We re-estimate value-added measures by implementing different specifications for equation (3.1). Results can be seen in Table 6. All specifications have grade and year indicators. The first, as a benchmark, includes all covariates previously used; the second, the baseline vector of covariates with no teacher fixed effect; the third controls only on all prior scores - school, grade, and student-level - considered in the baseline estimates; the fourth, student-level prior scores in both subjects; the fifth, student-level prior scores of the same subject; the sixth, all controls except the scores, *copago*, teacher assignment history, and rurality; the seventh and last, no controls whatsoever (except grade and year fixed effects: thus, the score prediction will be the student's mean score by teacher). We can see that controlling only on prior scores for the VA computations gives measures almost as unbiased that the ones computed with the full baseline vector of controls (0.813 coefficient for the baseline versus the prior test scores VA coefficient of 0.732). VA computation without teacher FE has less bias (coefficient of 0.890). This can be caused by the few years of our sample where we have VA measures for teachers. Measures that control for any kind of lagged score have a high correlation with the baseline VA estimate (90% or more), contrary to measures that do not control for previous scores (correlation of 74.7% for non-score controls VA measures with the baseline VA measure, and of 72.3% for the measure without controls). No-controls VA measure has a bias not considerably different to the measure that controls for the baseline vector without the previous score controls or *copago*. We can therefore see that lagged-scores are an important factor for the estimation of unbiased estimates, as noticed in previous literature, as well as *copago*. This pattern of correlation and bias in the different specifications is seen throughout the combined sample as well (Table 27, Appendix A).

Finally, as Chetty et al. (2014a) we estimate the relationship between VA measures and different student characteristics. Results can be seen in the Table 7. We can see in Column 1 that one standard deviation higher lagged test score is significantly associated with being assigned to a teacher with a VA 0.0175 higher. In Column 2 we estimate the assignment with school fixed effects, thus estimating within school differences in assignment. Within schools, students with a one SD higher lagged test score are significantly assigned teachers with 0.00299 higher VA. In Column 3 we see the relationship with special education students. On average, special education students are assigned VA teachers 0.0285 higher than non special students. In Column 5 we explore the assignment on parent income (in 10.000$). We see that students with higher household income get assigned higher VA teachers. This relationship is nonetheless small. We further explore if controlling on parent income changes the assignment on lagged score. In Column 6, conditional on parent income, one standard deviation higher in lagged test score implies being assigned a 0.0160 higher VA teacher. In Column 7 we regress value-added on the average household parent education: we find that an additional year in mean household parent education leads to being assigned a 0.00384 higher VA teacher. When evaluating the teacher-parent education assignment within school (Column 8), we find that there is no significant relationship between parent education and teacher value-added. Column 9 studies the relationship between a school-level demographic, mean school income, and VA estimates. Higher mean income schools have significantly better teachers, as measured by VA. Finally, in Column 10 we evaluate the assignment of teacher quality to *copago* charged by schools: schools that charge (on average) a 10.000$ higher *copago*, have on average teachers with a 0.00702 VA higher.

Together, these results unravel a new aspect of the segregation present in the chilean educational system: more advantaged schools have the ability to attract better teachers, therefore widening the existing socioeconomic differences between students from different socioeconomic backgrounds.

TABLE 6: COMPARISONS OF FORECAST BIAS ACROSS VALUE-ADDED MODELS, 6TH GRADE ONLY, SCHOOLS WITH *copago* ONLY, ALL YEARS

	Correlation with baseline VA estimates (1)	Quasi-experimental estimate of bias (%) (2)
Baseline	1.000	0.187
		(0.111)
Baseline no teacher FE	0.915	0.110
		(0.137)
Prior test scores	0.956	0.268
		(0.111)
Student's prior scores in both subjects	0.925	0.287
		(0.107)
Student's prior score in same subject only	0.912	0.318
		(0.104)
Non-score controls	0.747	0.663
		(0.065)
No controls	0.723	0.693
		(0.061)

Notes: In this table, we estimate seven alternative VA models and report correlations of the resulting VA estimates with the baseline VA 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. These coefficients are estimated using exactly the specification in column 1 of Table 4. All the VA models are estimated on a constant sample of students for whom all the variables in the baseline control vector are non-missing. 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 (3.1); the remaining steps of the procedure used to construct VA estimates are the same for all the models. Model 1 replicates the baseline model as a reference; see notes to Table 3 for definition of the baseline control vector. The estimated forecast bias for this model coincides with that implied by column 1 of Table 4. Model 2 uses all of the baseline controls but omits teacher fixed effects when estimating equation 3.1, so that the coefficients on the controls are identified from both within- and between-teacher variation as in traditional VA specifications. Model 3 includes only student-, class-, and school-level test score controls from the baseline control vector along with grade and year fixed effects. Model 4 includes only cubic polynomials in prior-year scores in Math and Reading along with grade and year fixed effects. Model 5 replicates model 4, dropping the cubic polynomial for prior scores in the other subject. Model 6 removes all controls related to test scores and *copago* from the baseline specification, leaving only non-score controls at the student, class, and school level (e.g., demographics, free lunch participation, etc.). Model 7 drops all controls except grade and year fixed effects.

TABLE 7: DIFFERENCES IN TEACHER QUALITY ACROSS STUDENTS AND SCHOOLS, 6TH GRADE ONLY, SCHOOLS WITH *copago* ONLY, ALL YEARS

Dependent variable:	Teacher Value-Added									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Lagged Test Score	0.0175	0.00299		0.0175		0.0160				
	(0.00190)	(0.000376)		(0.00190)		(0.00209)				
Special Education student			0.0285	0.0483						
			(0.0454)	(0.0443)						
Parent Income (10,000$)					0.000225	0.000176				
					(0.0000345)	(0.0000353)				
Parent education							0.00384	-0.0000532		
							(0.000619)	(0.0000686)		
School mean income (10,000$)									0.000704	
									(0.000102)	
School mean *copago* (10,000$)										0.00702
										(0.00149)
School fixed effects		X						X		
Constant	0.00618	0.00972	0.0104	0.00618	-0.00233	-0.00343	-0.0387	0.0121	-0.0322	-0.00517
	(0.00290)	(0.00181)	(0.00307)	(0.00290)	(0.00394)	(0.00387)	(0.00889)	(0.00209)	(0.00712)	(0.00463)
Observations	289200	289200	289200	289200	247673	247673	234839	234839	289200	289200

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by teacher in parentheses. Teacher VA, which is the dependent variable in all columns, is scaled in units of student test score standard deviations. Teacher VA is estimated using data from classes taught by the same teacher in other years, following the procedure in 5 and using the baseline control vector (see notes to 3 for more details). The regressions are run at the student-subject-year level on the sample used to estimate the baseline VA model. Each specification includes the student-level covariate(s) listed at the left hand side of the table and no additional control variables. See notes to Table 1 for definitions of these independent variables. Columns 2 and 8 include school fixed effects. In Column 9, the independent variable is the school-mean of the independent variable in Columns 5. In column 10, the independent variable is the mean of *copago* across the school. We calculate these means as the unweighted means across all student-subject-year observations with non-missing data for the relevant variable in each school.

Chapter 7

Conclusion

Unbiased value-added estimates can be retrieved in a different setting to the ones used so far, namely schools districts within the United States, such as in the competitive and strongly segregated in socioeconomic terms school system in Chile, even with short panel structures. Our results suggest that in this context, tuition as well as prior scores are fundamental controls for an unbiased value-added estimation, and their omission is likely to cause severely biased estimates. Evidence provided throughout the paper supports the validity of the quasi-experiment devised by Chetty et al. (2014a): it allows us the identification in bias in teacher value-added estimates, whenever experimental evaluations have not been implemented or are not feasible, as is the case of Chile.

When evaluating sorting on demographic characteristics, we find that the teacher quality is sorted positively to students and schools with higher income, to parents with higher education, to better achieving students in previous grades, and to higher tuition charging schools. This sorting on teacher quality contributes to corroborate the segregation in the chilean educational system, stemming not only from sorting among schools, but also as seen in this study, on teacher quality. As such, this sorting is not only likely to replicate socioeconomic differences, but also to increase these already present educational and economic disparities, as a higher quality teacher has a higher positive effect on achievement. However, unbiased identification of teacher quality not only allows to apprehend the current situation in the educational system, but also permits to orient public policy. Unbiased value-added measures of teacher quality might be useful to face these challenges, by implementing a reassignment of better teachers to more disadvantaged students.

Appendix

Appendix A

TABLE 8: SUMMARY STATISTICS FOR FULL SAMPLE: 6TH GRADE ONLY, ALL SCHOOLS, ALL YEARS

Variable	Mean (1)	SD (2)	Observations (3)
Class size (not student-weighted)	24.0	8.5	44009
Number of subject-school years per student	1.9	0.3	548793
Test score (SD)	0.05	1.0	1057724
Age (years)	11.2	0.5	1055997
Female	50.6%		1057724
Repeating grade	2.6%		1049391
Special education	0.03%		1056007
Household income	554,929	572,926	893607
Household education	12	3	840647
Copayment	21,185	20,722	453675

Notes: See Notes to Table 1. Years considered are 2013, 2014, 2015. Does not include schools from code of dependency *Corporación de Administración Delegada*, a small fraction of the overall population school type.

TABLE 9: SUMMARY STATISTICS FOR SAMPLE USED TO ESTIMATE VALUE-ADDED MODEL, 6TH GRADE ONLY, NO PRIVATE SCHOOLS, ALL YEARS

Variable	Mean (1)	SD (2)	Observations (3)
Class size (not student-weighted)	22.6	8.6	39963
Number of subject-school years per student	1.9	0.3	473563
Test score (SD)	0.01	1.0	902575
Age (years)	11.2	0.5	902575
Female	51.0%		902575
Repeating grade	2.6%		902575
Special education	0.03%		902575
Household income	444,978	401,189	773115
Household education	12	3	725079
Copayment	21,290	20,748	430657

Notes: See Notes to Table 1. Years considered are 2013, 2014, 2015.

TABLE 10: SUMMARY STATISTICS FOR SAMPLE USED TO ESTIMATE VALUE-ADDED MODEL, 8TH GRADE ONLY, NO PRIVATE SCHOOLS, ALL YEARS

Variable	Mean (1)	SD (2)	Observations (3)
Class size (not student-weighted)	21.7	8.2	62845
Number of subject-school years per student	1.9	0.3	707728
Test score (SD)	0.03	1.0	1363630
Age (years)	13.2	0.4	1363630
Female	51.5%		1363630
Repeating grade	1.9%		1363630
Special education	0.05%		1363630
Household income	412,674	385,647	1199904
Household education	11	3	1117700
Copayment	19,921	18,945	630094

Notes: See Notes to Table 1. Years considered are 2009, 2011, 2013, 2014, 2015.

TABLE 11: SUMMARY STATISTICS FOR SAMPLE USED TO ESTIMATE VALUE-ADDED MODEL, 6TH AND 8TH GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS

Variable	Mean (1)	SD (2)	Observations (3)
Class size (not student-weighted)	26.2	7.8	40423
Number of subject-school years per student	2.2	0.7	490740
Test score (SD)	0.25	0.9	1059083
Age (years)	12.3	1.0	1059083
Female	51.7%		1059083
Repeating grade	2.5%		1059083
Special education	0.02%		1059083
Household income	566,492	467,801	916401
Household education	13	3	864902
Copayment	20,489	19,717	1059083

Notes: See Notes to Table 1. Years considered are 2009, 2011, 2013, 2014, 2015.

TABLE 12: TEACHER VALUE-ADDED MODEL PARAMETER ESTIMATES, 6TH AND 8TH GRADES, WITH *copago* ONLY, ALL YEARS

	(1)	(2)
Sample	Reading	Math
Lag 1	0.024	0.042
	(0.001)	(0.001)
	[0.369]	[0.592]
Lag 2	0.019	0.036
	(0.001)	(0.001)
	[0.314]	[0.510]
Lag 3	0.015	0.028
	(0.002)	(0.002)
	[0.219]	[0.378]
Lag 4	0.013	0.027
	(0.003)	(0.003)
	[0.210]	[0.382]
Total variance	0.503	0.405
Individual-level variance	0.438	0.336
Class variance	0.027	0.011
Teacher variance	0.038	0.057

Notes: See Notes to Table 2. Years considered are 2009, 2011, 2013, 2014, 2015.

TABLE 13: TEACHER VALUE-ADDED MODEL PARAMETER ESTIMATES, 6TH AND 8TH GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS, NO MASTER OR PHD

Sample	(1) Reading	(2) Math
Lag 1	0.024	0.043
	(0.001)	(0.001)
	[0.354]	[0.592]
Lag 2	0.022	0.037
	(0.001)	(0.002)
	[0.341]	[0.520]
Lag 3	0.015	0.028
	(0.003)	(0.003)
	[0.220]	[0.372]
Lag 4	0.015	0.031
	(0.004)	(0.004)
	[0.236]	[0.409]
Total variance	0.499	0.401
Individual-level variance	0.434	0.333
Class variance	0.024	0.011
Teacher variance	0.040	0.057

Notes: See Notes to Table 2. Years considered are 2009, 2011, 2013, 2014, 2015.

Table 14: Teacher Value-Added Model Parameter Estimates, 6th and 8th grades, schools with *copago* only, all years, with master or PhD

	(1)	(2)
Sample	Reading	Math
Lag 1	0.027	0.045
	(0.004)	(0.004)
	[0.378]	[0.572]
Lag 2	0.016	0.024
	(0.005)	(0.005)
	[0.251]	[0.360]
Lag 3	0.014	0.026
	(0.006)	(0.008)
	[0.191]	[0.334]
Lag 4	0.010	0.005
	(0.015)	(0.010)
	[0.151]	[0.085]
Total variance	0.526	0.428
Individual-level variance	0.451	0.352
Class variance	0.029	0.008
Teacher variance	0.046	0.069

Notes: See Notes to Table 2. Years considered are 2009, 2011, 2013, 2014, 2015.

TABLE 15: TEACHER VALUE-ADDED MODEL PARAMETER ESTIMATES, 6TH AND 8TH GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS, NO POST-SECONDARY OR GRADUATE EDUCATION BEYOND COLLEGE

	(1)	(2)
Sample	Reading	Math
Lag 1	0.031	0.045
	(0.004)	(0.002)
	[0.415]	[0.596]
Lag 2	0.024	0.039
	(0.003)	(0.003)
	[0.336]	[0.533]
Lag 3	0.011	0.028
	(0.004)	(0.004)
	[0.181]	[0.421]
Lag 4	0.009	0.028
	(0.006)	(0.007)
	[0.157]	[0.373]
Total variance	0.504	0.397
Individual-level variance	0.440	0.331
Class variance	0.022	0.009
Teacher variance	0.043	0.057

Notes: See Notes to Table 2. Years considered are 2009, 2011, 2013, 2014, 2015.

Table 16: Teacher Value-Added Model Parameter Estimates, 6th and 8th grades, schools with *copago* only, all years, with post-secondary or graduate education beyond college

Sample	(1) Reading	(2) Math
Lag 1	0.023	0.041
	(0.002)	(0.002)
	[0.331]	[0.560]
Lag 2	0.021	0.036
	(0.002)	(0.002)
	[0.330]	[0.494]
Lag 3	0.018	0.029
	(0.003)	(0.003)
	[0.252]	[0.369]
Lag 4	0.016	0.027
	(0.005)	(0.004)
	[0.238]	[0.395]
Total variance	0.502	0.409
Individual-level variance	0.436	0.340
Class variance	0.025	0.011
Teacher variance	0.041	0.058

Notes: See Notes to Table 2. Years considered are 2009, 2011, 2013, 2014, 2015.

TABLE 17: TEACHER VALUE-ADDED MODEL PARAMETER ESTIMATES, 6TH AND 8TH GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS, QUARTILE 1 OF TEACHER EXPERIENCE

	(1)	(2)
Sample	Reading	Math
Lag 1	0.027	0.042
	(0.003)	(0.002)
	[0.391]	[0.587]
Lag 2	0.024	0.037
	(0.003)	(0.003)
	[0.378]	[0.530]
Lag 3	0.007	0.010
	(0.008)	(0.004)
	[0.110]	[0.186]
Lag 4	0.020	0.012
	(0.008)	(0.010)
	[0.389]	[0.208]
Total variance	0.520	0.406
Individual-level variance	0.450	0.339
Class variance	0.028	0.011
Teacher variance	0.042	0.056

Notes: See Notes to Table 2. Years considered are 2009, 2011, 2013, 2014, 2015.

TABLE 18: TEACHER VALUE-ADDED MODEL PARAMETER ESTIMATES, 6TH AND 8TH GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS, QUARTILE 2 OF TEACHER EXPERIENCE

Sample	(1) Reading	(2) Math
Lag 1	0.032	0.044
	(0.003)	(0.002)
	[0.434]	[0.616]
Lag 2	0.025	0.037
	(0.003)	(0.002)
	[0.364]	[0.506]
Lag 3	0.020	0.033
	(0.006)	(0.004)
	[0.259]	[0.426]
Lag 4	0.008	0.019
	(0.009)	(0.005)
	[0.098]	[0.315]
Total variance	0.507	0.404
Individual-level variance	0.437	0.333
Class variance	0.024	0.011
Teacher variance	0.047	0.059

Notes: See Notes to Table 2. Years considered are 2009, 2011, 2013, 2014, 2015.

Table 19: Teacher Value-Added Model Parameter Estimates, 6th and 8th grades, schools with *copago* only, all years, quartile 3 of teacher experience

	(1)	(2)
Sample	Reading	Math
Lag 1	0.021	0.048
	(0.002)	(0.002)
	[0.336]	[0.632]
Lag 2	0.017	0.043
	(0.002)	(0.003)
	[0.286]	[0.571]
Lag 3	0.015	0.036
	(0.003)	(0.004)
	[0.254]	[0.471]
Lag 4	0.018	0.044
	(0.005)	(0.005)
	[0.285]	[0.543]
Total variance	0.494	0.414
Individual-level variance	0.433	0.341
Class variance	0.024	0.013
Teacher variance	0.037	0.061

Notes: See Notes to Table 2. Years considered are 2009, 2011, 2013, 2014, 2015.

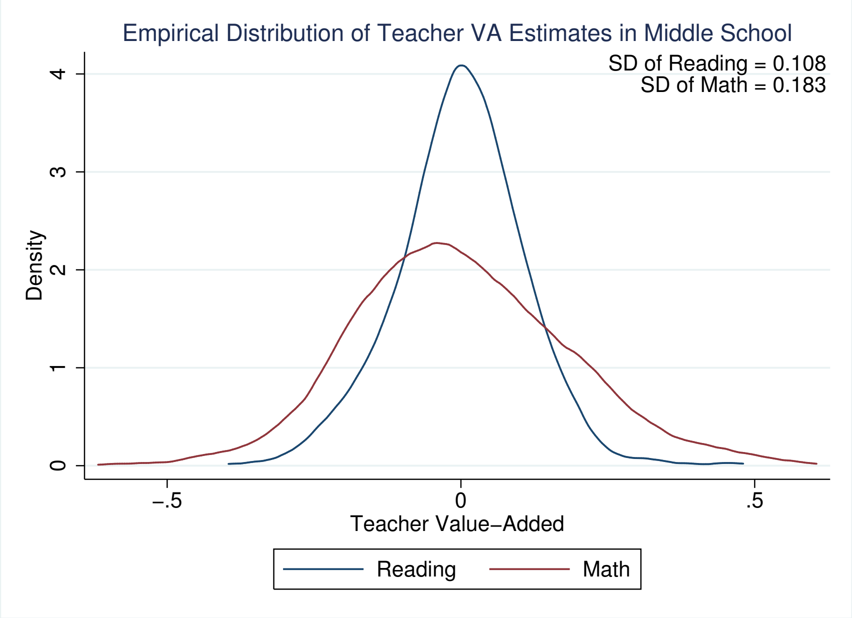
TABLE 20: TEACHER VALUE-ADDED MODEL PARAMETER ESTIMATES, 6TH AND 8TH GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS, QUARTILE 4 OF TEACHER EXPERIENCE

	(1)	(2)
Sample	Reading	Math
Lag 1	0.031	0.042
	(0.002)	(0.002)
	[0.454]	[0.626]
Lag 2	0.025	0.039
	(0.002)	(0.003)
	[0.406]	[0.567]
Lag 3	0.026	0.035
	(0.004)	(0.005)
	[0.359]	[0.451]
Lag 4	0.021	0.031
	(0.005)	(0.006)
	[0.338]	[0.425]
Total variance	0.496	0.399
Individual-level variance	0.428	0.331
Class variance	0.030	0.017
Teacher variance	0.038	0.051

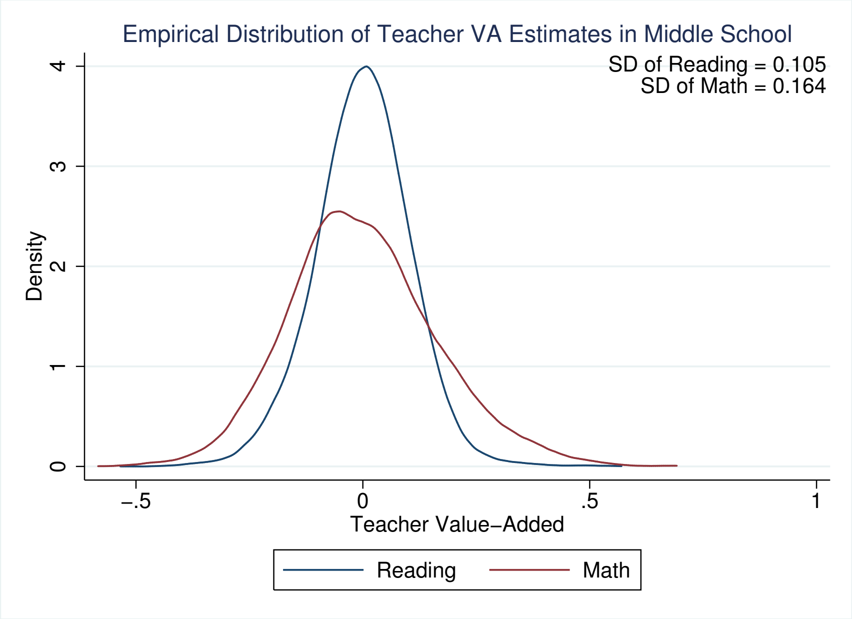
Notes: See Notes to Table 2. Years considered are 2009, 2011, 2013, 2014, 2015.

FIGURE 4: EMPIRICAL DISTRIBUTIONS OF TEACHER VA ESTIMATES

6^{TH} GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS

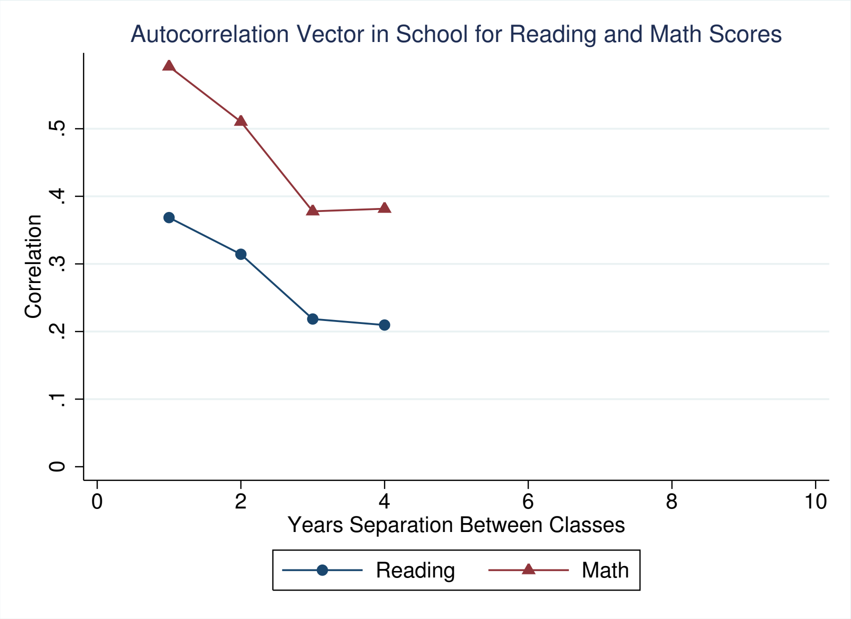


6^{TH} AND 8^{TH} GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS



Notes: These figures plot kernel densities of the empirical distribution of teacher VA estimates $\hat{\mu}_{jt}$ for each subject (Math and Reading). The densities are weighted by the number of student test score observations used to construct the teacher VA estimate and are estimated using a bandwidth of 0.025. We also report the standard deviations of these empirical distributions of VA estimates. Note that these standard deviations are smaller than the standard deviation of true teacher effects reported in Table 2 because VA estimates are shrunk toward the mean to account for noise and obtain unbiased forecasts.

FIGURE 5: DRIFT IN TEACHER VALUE-ADDED ACROSS YEARS, 6TH AND 8TH GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS



Notes: These figures show the correlation between mean test-score residuals across classes taught by the same teacher in different years. To calculate these vectors, we first residualize test scores using within-teacher variation with respect to our baseline control vector. We then calculate a (precision-weighted) mean test score residual across classrooms for each teacher-year. Finally, we calculate the autocorrelation coefficients as the correlation across years for a given teacher, weighting by the sum of students taught in the two years.

TABLE 21: ESTIMATES OF FORECAST BIAS USING PARENT CHARACTERISTICS, 6TH AND 8TH GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS

Sample	(1) Score in year *t*	(2) Pred. score using parent chars.	(3) Score in year *t*	(4) Pred. score using parent chars.	(5) Pred. score using parent chars.
Teacher VA	1.002	-0.001	0.992	0.000	0.001
	(0.015)	(0.001)	(0.015)	(0.000)	(0.000)
Parent chars. controls			X		
School Fixed Effects				X	
Copago					X
Obsevations	819614	663410	663410	663410	663410

Notes: See Notes to Table 3.

Table 22: Quasi-Experimental Estimates of Forecast Bias, 6th grade only, no private schools, *copago* not considered, all years

	(1)	(2)	(3)	(4)
			Δ predicted	Δ other
	Δ score	Δ score	score	subject score
Changes in mean teacher	0.613	0.371	0.000	0.204
VA across cohorts	(0.071)	(0.174)	(0.018)	(0.091)
Year fixed effects	X			X
School × year fixed effects		X	X	
Other-subject change in mean teacher VA				X
Grades	6	6	6	6
Number of school × grade × subject × year cells	10396	10396	10219	6908

Notes: See Notes to Table 4.

Table 23: Quasi-Experimental Estimates of Forecast Bias, 6th grade only, no private schools, all years

	(1)	(2)	(3)	(4)
			Δ predicted	Δ other
	Δ score	Δ score	score	subject score
Changes in mean teacher	0.666	0.403	-0.000	0.212
VA across cohorts	(0.073)	(0.185)	(0.019)	(0.095)
Year fixed effects	X			X
School × year fixed effects		X	X	
Other-subject change in mean teacher VA				X
Grades	6	6	6	6
Number of school × grade × subject × year cells	10396	10396	10219	6908

Notes: See Notes to Table 4.

TABLE 24: QUASI-EXPERIMENTAL ESTIMATES OF FORECAST BIAS, 6TH AND 8TH GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS

	(1) Δ score	(2) Δ score	(3) Δ score	(4) Δ predicted score	(5) Δ other subject score
Changes in mean teacher VA across cohorts	0.543 (0.044)	0.442 (0.065)	0.364 (0.195)	0.005 (0.009)	0.171 (0.050)
Year fixed effects	X				X
School × year fixed effects		X	X	X	
Lagged score controls			X		
Lead and lag changes in teacher VA			X		
Other-subject change in mean teacher VA					X
Grades	6/8	6/8	6/8	6/8	6/8
Number of school × grade × subject × year cells	13011	13011	2624	12922	9264

Notes: See Notes to Table 4.

TABLE 25: QUASI-EXPERIMENTAL ESTIMATES OF FORECAST BIAS: ROBUSTNESS CHECKS, 6TH GRADE ONLY, SCHOOLS WITH *copago* ONLY, ALL YEARS, RESIDUAL SCORES

	(1) Teacher exit only Δ score	(2) Full sample Δ score	(3) <25 percent imputed VA Δ score	(4) 0 percent imputed VA Δ score
Changes in mean teacher VA across cohorts	0.647 (0.252)	0.547 (0.056)	0.890 (0.137)	0.890 (0.147)
Year fixed effects	X	X	X	X
Grades	6	6	6	6
Number of school × grade × subject × year cells	3885	17112	3190	3148

Notes: See Notes to Table 5.

TABLE 26: QUASI-EXPERIMENTAL ESTIMATES OF FORECAST BIAS: ROBUSTNESS CHECKS, 6TH AND 8TH GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS

	(1) Teacher exit only Δ score	(2) Full sample Δ score	(3) <25 percent imputed VA Δ score	(4) 0 percent imputed VA Δ score
Changes in mean teacher VA across cohorts	0.398 (0.091)	0.431 (0.031)	0.547 (0.049)	0.532 (0.049)
Year fixed effects	X	X	X	X
Grades	6/8	6/8	6/8	6/8
Number of school × grade × subject × year cells	13011	46904	11032	10870

Notes: See Notes to Table 5.

TABLE 27: COMPARISONS OF FORECAST BIAS ACROSS VALUE-ADDED MODELS, 6TH AND 8TH GRADES, SCHOOLS WITH *copago* ONLY, ALL YEARS

	Correlation with baseline VA estimates (1)	Quasi-experimental estimate of bias (%) (2)
Baseline	1.000	0.457 (0.044)
Baseline no teacher FE	0.961	0.427 (0.048)
Prior test scores	0.973	0.486 (0.043)
Student’s lagged scores in both subjects	0.929	0.565 (0.039)
Student’s lagged score in same subject only	0.916	0.595 (0.037)
Non-score controls	0.709	0.789 (0.023)
No controls	0.688	0.810 (0.021)

Notes: See Notes to Table 6.

Appendix B

In this section, we explore further the placebo test seen in Column 4, Table 4. Through Tables 22-24, this test shows a significant positive correlation. While these results don't support the plausibility of the key assumption, they do not invalidate the quasi-experiment if the teacher quality in one subject affects positively the scores in the other subjects. This is what would be expected for example if a Reading (Math) teacher quality affects positively (negatively) Math (Reading) scores. We study these relations by restricting the sample to scores in one or the other subject. Results can be seen in columns 1 and 2 of Table 28. Though statistically insignificant, the effect varies across the different subjects. It is more important for math teacher quality on Reading than reading teacher quality on Math for our main sample. We consider if there is a different effect when considering test score achievement below or above the median in Columns 3, 4, 5 and 6 of Table 28. We can see that the positive effect of math teacher quality on reading scores is strong for the lower performing group, *i.e.* when scores are below the median, and the opposite is true when scores are from the high achievement group (above the median). Reading teacher quality impacting more importantly Math scores than viceversa, is also seen in Table 29. Intuitively, better reading skills can improve the comprehension in other subjects such as math. Reading teacher quality impacting math scores more importantly than viceversa is seen throughout the different sample restrictions in Table 29.

TABLE 28: QUASI-EXPERIMENTAL ESTIMATES OF FORECAST BIAS: CHANGES IN THE OTHER SUBJECT: DETAILS, 6TH GRADE ONLY, SCHOOLS WITH *copago* ONLY, ALL YEARS

	Math on reading	Reading on math	Math on reading below median mean score	Reading on math below median mean score	Math on reading above median mean score	Reading on math above median mean score
	(1)	(2)	(3)	(4)	(5)	(6)
Changes in mean VA	0.137 (0.155)	0.048 (0.184)	0.263 (0.248)	-0.033 (0.234)	-0.014 (0.193)	0.111 (0.248)
Observations	1345	1345	595	592	750	753

Notes: Each column replicates the estimation in Table 4, Column 4, restricting further the sample. In Column 1 we restrict to Reading scores as the dependent variable observations only; in column 2 to Math scores as a dependent variable only; in column 3 we restrict further to Reading scores as a dependent variable only, where we restrict to below median mean scores; in column 4 we rescrict to Math scores as a dependent variable only, and restrict to below median mean scores; in column 5 we restrict to Reading scores as a dependent variable only, considering above median mean scores; and in column 6 we restrict to Math scores as a dependent variable only, considering above median mean scores.

Table 29: Quasi-Experimental Estimates of Forecast Bias: Changes in the Other Subject: Details, 6$^{\text{th}}$ and 8$^{\text{th}}$ grades, schools with *copago* only, all years

	Math on reading	Reading on math	Math on reading below median mean score	Reading on math below median mean score	Math on reading above median mean score	Reading on math above median mean score
	(1)	(2)	(3)	(4)	(5)	(6)
Changes in mean VA	0.105 (0.065)	0.306 (0.090)	0.115 (0.103)	0.177 (0.131)	0.095 (0.078)	0.355 (0.110)
Observations	4632	4632	1984	1925	2648	2707

Notes: See notes to Table 28.

Bibliography

Angrist, J., Hull, P., Pathak, P., & Walters, C. (2016). Interpreting Tests of School VAM Validity. *American Economic Review*, 106(5):388–92.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2017). An Evaluation of Bias in Three Measures of Teacher Quality: Value-Added, Classroom Observations, and Student Surveys. Working Paper 23478, National Bureau of Economic Research.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles. Working Paper 20657, National Bureau of Economic Research.

Canales, A. & Maldonado, L. (2018). Teacher quality and student achievement in Chile: Linking teachers' contribution and observable characteristics. *International Journal of Educational Development*, 60:33 – 50.

Chetty, R., Friedman, J. N., & Rockoff, J. (2016). Using lagged outcomes to evaluate bias in value-added models. *American Economic Review*, 106(5):393–399.

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

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

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014c). Stata Code for Implementing Teaching-Staff Validation Technique. *Unpublished manuscript.*, 21:2014.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2017). Measuring the Impacts of Teachers: Reply. *American Economic Review*, 107(6):1685–1717.

Corcoran, S. P. (2010). Can Teachers Be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Education Policy for Action Series. *Annenberg Institute for School Reform at Brown University*.

Glazerman, S. & Protik, A. (2015). Validating Value-Added Measures of Teacher Performance. *Unpublished manuscript*.

Glazerman, S., Protik, A., Teh, B.-r., Bruch, J., & Max, J. (2013). Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment. NCEE 2014-4004. *National Center for Education Evaluation and Regional Assistance*.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, 30(3):466–479.

Hanushek, E. A. & Rivkin, S. G. (2010). Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review*, 100(2):267–71.

Hanushek, E. A. & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics*, 4(1):131–157.

Hanushek, E. A. & Woessmann, L. (2011). How much do educational outcomes matter in OECD countries? *Economic Policy*, 26(67):427–491.

Hsieh, C.-T. & Urquiola, M. (2006). The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program. *Journal of public Economics*, 90(8):1477–1503.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. In *Research Paper. MET Project. Bill & Melinda Gates Foundation*.

Kane, T. J. & Staiger, D. O. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. Working Paper No. 14607, National Bureau of Economic Research.

Koedel, C. & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*, 6(1):18–42.

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

Mizala, A. & Torche, F. (2012). Bringing the schools back in: the stratification of educational achievement in the Chilean voucher system. *International Journal of Educational Development*, 32(1):132–144.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, 4(4):537–571.

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

Rothstein, J. (2014). Revisiting the impacts of teachers. *Unpublished*.

Rothstein, J. (2017). Measuring the Impacts of Teachers: Comment. *American Economic Review*, 107(6):1656–84.

Santelices, M. V., Galleguillos, P., González, J., & Taut, S. (2015). Un estudio sobre la calidad docente en chile: El rol del contexto en donde enseña el profesor y medidas de valor agregado. *Psykhe (Santiago)*, 24(1):1–14.

Staiger, D. O. & Kane, T. J. (2014). Making decisions with imprecise performance measures: The relationship between annual student achievement gains and a teacher's career value-added. *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*.

Taut, S., Valencia, E., Palacios, D., Santelices, M. V., Jimenez, D., & Manzi, J. (2016). Teacher performance and student learning: linking evidence from two national assessment programmes. *Assessment in Education: Principles, Policy & Practice*, 23(1):53–74.