Preliminary and Incomplete: PLEASE DO NOT CITE

High-stakes Objective and Subjective Teacher Evaluation Measures and Student Skill Development

Ayman Shakeel*

September 23, 2022

Click here for the latest version

Abstract

Advocates for the use of a multi-measure system of performance evaluation argue that multiple measures may better capture meaningful differences in employee effectiveness and help better align their effort with valued outcomes. This may be particularly important in organizations such as schools that produce multiple outcomes. In this paper, I estimate the relative contributions of the subjective (supervisor observations and student surveys) versus objective (value-added) evaluation measures in capturing teacher effectiveness at increasing cognitive and non-cognitive skills. I use data from a large urban public school district where teacher compensation is tied to the evaluation measures. Estimates reveal that the subjective measures provide information about teacher effects on short-run achievement and behaviors and long-run achievement in ways that value-added does not. However, in the multiple measure setting, value-added remains the most significantly related to high-stakes achievement and low-stakes long-term achievement and absences. Understanding how the evaluation measures perform in a multi-measure high-stakes framework is important from the policy standpoint.

JEL Classification: H75, I21, I28, J24, J45

*Ph.D. candidate, Department of Economics, University of Illinois at Chicago. I am very grateful to my supervisors, Steven Rivkin, Ben Ost, Javaeria Qureshi, Ben Feigenberg, and Kirabo Jackson for their guidance and support. I also thank participants at UIC Economics seminars, American Education Finance Policy (AEFP), Midwest Economic Association (MEA) and Illinois Economic Association (IEA) for their insightful comments. The findings and and interpretations expressed in this paper are my own and do not necessarily represent the views of the school district under study. All errors are my own.

1 Introduction

Measurement of worker productivity has become a more central part of human resource management in many organizations. There has been a growing interest in the implementation of objective measures to evaluate workers, in part because of concerns that supervisor evaluations are prone to biases. However, this development has not been without its detractors. Many are concerned that objective measures are limited, particularly in situations where worker output is multi-dimensional or subject to strategic behaviors (Holmstrom and Milgrom 1991).[1] These concerns range from the public sector, not-for-profit organizations, and the private sector, to corporate leadership. To mitigate incentive distortions, many organizations resort to a combination of objective and subjective measures. The education sector has been at the forefront of efforts to incorporate these measures into evaluation systems, and I use that context to assess the performance of these measures.

Proliferation of state longitudinal data over the past two decades has resulted in efforts by policy makers to measure teacher performance more accurately and comprehensively. Historically, most schools used subjective evaluations, typically based on supervisor observations, but evidence suggested that principals are often lenient in their observations (Weisberg et al. 2009).[2] This prompted a push for the use of objective measures of teacher effectiveness, including test score value-added. However, limitations of value-added, including perverse incentives for strategic behavior and a disproportionate focus on a single type of outcome, led to concerns about using it as a standalone measure. One solution has been the use of a combination of test score value-added and supervisor observations to blunt the adverse effects of each, yet this combination may fail to address the respective deficiencies of the two. For example, supervisors may be hesitant to give negative evaluations and highlight shortcomings of teachers who engage in strategic behaviors that raise achievement without

[1] Multi-tasking theory reveals that performance measures that reward production of the observed output induce agents to substitute effort toward the measurable output at the expense of the unobserved task (Holmstrom and Milgrom 1991).

[2] Supervisor observation refers to the observation of teaching by the administrators, usually the principal, for the evaluation of the classroom instruction.

fostering deeper learning. A potential pathway to improve the process is the introduction of student surveys as an additional subjective measure.

Although the combination of the objective and subjective measures has become a widely used structure of teacher evaluation[3], it remains unclear how these measures interact in a multi-measure framework, especially when high-stakes are attached. The key question that policy makers are interested in is whether the evaluation structure supports and incentivizes instruction that raises valuable skills.

In this paper, I investigate the relationships between cognitive and non-cognitive skill acquisition on the one hand and objective (test score value-added) and subjective measures (supervisor observation and student surveys) of teacher effectiveness on the other. I use administrative data from a large urban public school district that attaches high-stakes to achievement gains, supervisor observations, and student evaluations of teachers, starting in the 2014-2015 academic year. Test score value-added, supervisor observations, and student surveys are high-stakes measures, as they are used to evaluate teacher performance and with some qualification, determine compensation for the subsequent year. This paper focuses on grades 3-7 math and reading teachers.

Although the evaluation system attaches high-stakes to contemporaneous achievement, a fundamental question is whether the system contributes to the development of valuable skills. To answer this, I investigate whether teachers with high value-added or subjective scores raise contemporaneous (year t) achievement *and* valued lower-stakes cognitive and non-cognitive outcomes. I use math and reading test scores as a proxy for cognitive skills, and behaviors, which include absences and receipt of a suspension, serve as a proxy for non-cognitive skills.[4] The lower-stakes outcomes are contemporaneous behaviors and subsequent

[3]As of 2019, 34 states require objective evaluations and 37 mandate supervisor observations for at least some teachers to be a part of the evaluation process. 7 states require student surveys and 24 allow it but do not explicitly make it compulsory. With the exception of New York that prohibits the use of surveys to evaluate teachers, other states do not have any explicit policy regarding the use of this measure to evaluate teachers (Ross and Walsh 2019).

[4]Although most accountability systems focus on achievement, I also include some non-cognitive outcomes in my analysis as research shows that they are key determinants of future success (Heckman and Rubinstein 2001; Lindqvist and Vestman 2011). The choice of the behaviors to serve as a proxy for non-cognitive skills

year achievement and behaviors. For a limited sample, I also study whether the evaluation measures are related to some high school outcomes.[5] I include the long-term outcomes in the analysis to capture teacher effects that persist into the future and raise skills, as opposed to effects on current outcomes that may come in part from strategic behavior.

Despite the fact that non-test score value-added is not a part of the evaluation systems, research shows that it plays an important role in capturing teacher effects, especially on non-cognitive skills (Jackson 2018; Petek and Pope 2016; Liu and Loeb 2019). Therefore, I include absence and suspension value-added in my analysis to compare the different measures. I also compare the incremental contributions of the subjective measures with value-added for both high-stakes achievement and the lower-stakes cognitive and non-cognitive skill acquisition. This analysis aims to illuminate the additional variation explained by each effectiveness measure.

Identification of the relationship between the evaluation measures and teacher effectiveness, as measured by her contribution to student outcomes, requires that the specification accounts for potential confounding factors related to both the measures and the outcomes. Following an approach similar to that outlined by Chetty, Friedman, and Rockoff (2014a), I use leave-year-out measures of value-added, supervisor observation, and student survey scores. I predict a teacher's value-added and scores in a given school year based on the value-added and scores that she receives in all other years. To account for sorting, I use a rich set of controls in specifications that link the performance measures to student outcomes.

The results provide new insights on the relationships between the evaluation measures and student skills. Under the multi-measure system, test score value-added has highly significant relationships with achievement and absences in the current and subsequent year. Furthermore, teachers with high supervisor observation scores increase contemporaneous and subsequent year achievement. Not only are the supervisor observations related to achieve-

is motivated by previous studies (Jackson 2018; Petek and Pope 2021; Kautz and Wladimir 2014).

[5]I also add high school outcomes in my analysis as they are related to college attendance, labor market outcomes and crime. These outcomes include high-school tests related outcomes, absences, receipt of a suspension, being held back and attending a public high school in the district.

ment, they explain variation on top of test score value-added for both the contemporaneous test scores and subsequent year reading. Interestingly, student survey scores also capture current and subsequent year achievement, and they remain significantly related to contemporaneous absences even after controlling for the other performance measures. In fact, of the three high-stakes measures, teachers with high student survey scores are the most effective in reducing contemporaneous absences and increasing the probability of students remaining in the public school district for high school. The pattern of results is similar between subsequent year and high school outcomes, with the exception that supervisor observations are not significantly related to any of the high school outcomes when conditioned on test score value-added and survey score, while student surveys provide some additional information for Algebra I test scores.

In sum, subjective measures provide information about teacher effects on short-run achievement and behaviors as well as long-run achievement in ways that value-added does not. However, in the multi-measure system, high value-added teachers have the strongest impact on high-stakes achievement and valued low-stakes long-run outcomes.

This paper contributes to two important strands of literature on the measurement of employee productivity. First, I extend the research on the effectiveness of performance measures at capturing human capital production in a high-stakes regime. A rich theoretical literature outlines the optimal performance measure contract in a setting where the principal cannot contract on multi-dimensional agent output (Holmstrom and Milgrom 1991; Baker 1992). I evaluate the performance measures included within a given contract in the education sector. Recent work highlights that teachers contribute to the development of both cognitive and non-cognitive skills. However, this non-cognitive skill production is less captured by test score value-added (Jackson 2018; Petek and Pope 2016; Liu and Loeb 2019). Although teachers that raise non-test score value-added contribute to non-cognitive skills, from the policy perspective, it is not wise to attach financial incentives to non-test score value-added, owing to concerns regarding strategic behavior (Jackson 2018). I extend the literature by

assessing how well a system with both test score value-added and the subjective measures reflects teacher effects on student skills.[6]

Furthermore, most of the existing work examining the relationships between different performance measures and student outcomes is based on a low-stakes setting. Recent literature highlights the importance of the context in which teachers are evaluated (Neal 2011). For example, attaching high-stakes to supervisor observation scores changes principal leniency (Grissom and Loeb 2017), and changing teacher stake in the test outcome alters the decay of her effects on future achievement (Corcoran, Jennings and Beveridge 2012). Thus, understanding how the teacher evaluation measures perform in a multi-measure high-stakes framework is important from the policy perspective.

Second, this paper makes an important contribution to the subjective performance assessment literature. Jacob and Lefgren (2008) and Kraft, Papay and Chi (2020) show that supervisor ratings are significantly related to contemporaneous student achievement. I extend their finding by presenting new evidence that teachers with high observation scores also raise subsequent year achievement. To the best of my knowledge, this is also the first paper to examine the relationship between student surveys and skill development among younger students. There is some literature at the post-secondary level where student evaluations are more common. For example, Carrel and West (2010) show that post secondary student evaluations have a positive association with contemporaneous professor value-added but are negatively related to future student achievement. While some educators are concerned that young students may not be sophisticated enough to evaluate good teaching, the findings here suggest otherwise. Student surveys add information on the type of teaching that augments cognitive and non-cognitive skills. More importantly, these surveys can partially fill in the gaps left by value-added and supervisor observations in evaluating the quality of instruction.

[6]Kraft (2019) attempts to explore this by examining the correlations between value-added, supervisor observations and student surveys on the one hand and teacher effects on test scores and socio-emotional measures on the other, in a low-stakes setting. Although his findings of weak correlations give some insights on the performance of the measures, the study does not analyze the long-term outcomes or the relative contributions of subjective versus objective measures.

The findings in this paper also speak to the broader literature on the use of subjective measures to evaluate employees and have implications for professions other than teaching. Despite research documenting biases in supervisor observations (Weisberg et al. 2009; Murphy 1992; Frederiksen, Lange and Kriechel 2017), the findings here suggest that at least, to a certain extent, supervisors are successful in recognizing employees that raise current and future output. Furthermore, the idea of having students evaluate teachers is analogous to the incorporation of customer or client feedback to evaluate product quality and performance.

The remainder of this paper proceeds as follows. Section 2 provides the institutional background. Sections 3 and 4 present the data and empirical strategy, respectively. Section 5 describes the evaluation measures, and Section 6 discusses the results. Finally, Section 7 concludes the paper.

2 Background

This section describes the teacher evaluation framework and its transformation over the years in United States. I then outline the policy environment under study.

2.1 Teacher evaluation framework

Over the years, the evaluation of teaching in the United States has evolved to achieve convergence between teacher effort and valued learning outcomes. Historically, the evaluations were primarily based on supervisor observations, however, recent policy changes call for a multiple measures system of teacher evaluation. Although there is variation in the processes implemented by states and districts, most use supervisor observations and value-added to inform decisions about teacher quality. Nevertheless, in the presence of test based accountability both measures combined may still fail to provide a comprehensive measure of teacher effectiveness, resulting in the growing interest in the student survey measure.

Despite changes in the evaluation systems, supervisor observations are typically the ma-

jor component. Supervisor observations can be conducted for most teachers and help provide them with constructive feedback. However, recent studies also highlight that these observations are lenient, only able to differentiate between teachers in the tails of the quality distribution and can be compromised by the relationship between the observer and teacher (Jacob and Lefgren 2005; Jacob and Lefgren 2008; Weisberg et al. 2009).

Consequently, Race to the Top and the No Child Left Behind waivers incentivized states to include test score value-added in the evaluation system. Based on a low-stakes regime, Chetty, Friedman, and Rockoff (2014b) show that teachers with higher value-added have a significant positive impact on the student long-run outcomes, highlighting the importance of test score value-added. However, as a standalone measure it has certain deficiencies. First, value-added cannot be estimated for all teachers as test scores are only available for some subjects and grades (Whitehurst, Chingos and Lindquist 2014). Second, extensive focus on this measure can narrow the curriculum and induce teachers to teach to the test. If the tests are not related to long-term success, this may be detrimental to the students. More importantly, research shows that teacher quality is multi-dimensional and test score value-added does not capture the full impact of teachers on human capital (Jackson 2018; Petek and Pope 2016; Liu and Loeb 2019).

Several states and districts use student surveys as an additional measure to evaluate teachers. It was mainly after the Measures of Effective Teaching (MET) Project that light was shed on the proposal of giving student voice more importance in the primary and secondary schools.[7] Despite growing interest, student surveys are a relatively less explored measure.[8] As student evaluations are an aggregate of the perceptions of several students

[7]In 2009, the Bill and Melinda Gates Foundation initiated the MET project to test new ways to assess teacher effectiveness. The project mainly focused on the use of student achievement, classroom observations and student surveys to evaluate teachers.

[8]Most of the existing work relates to the MET project and finds a positive association between survey results and achievement gains (Kane and Cantrell 2010; Kane and Staiger 2012; Raudenbush and Jean 2014). A recent study by Bacher-Hicks et al. (2019) investigates whether survey responses under natural assignment are correctly able to predict teacher performance after random assignment. Due to small sample size, the estimates are less precise for student surveys. Nevertheless, the authors suggest that one should be cautious about this measure as it does not seem to contain substantial information about teacher performance.

who interact the most with the teachers, the combined responses have the potential to be related to teacher performance. If the survey is written in a way that elicits evaluations of teacher effectiveness at fostering deeper learning and the development of lasting skills, it can help convey information on the quality of instruction that is not captured by supervisor observations or value-added. Their presence can also alter the relationships between the other measures and teacher effectiveness at raising skills by helping avoid classroom structures where skill development is not supported.

While there is reason to believe that student surveys may be a promising measure of teacher quality, it is contingent on the assumption that students are able to recognize good teaching. Administrators argue that students may not be sophisticated enough to evaluate distinct aspects of instruction (Marsh and Roche 1997). One concern rises from the issue that including this measure in the evaluation process may induce a popularity contest. In light of the work at the post secondary level, there is also a concern that ratings could be a function of race, ethnicity, gender, or sexual orientation in addition to the quality of instruction (Braga, Paccagnella and Pellizzari 2014; Boring 2017; Wagner, Rieger and Voorvelt 2016). Thus, there is some skepticism towards the use of this measure to assess teachers.

Whether these surveys capture differences in teacher effectiveness at fostering valued skills not explained by the other two measures, and more importantly, whether they help mitigate incentive distortions are key questions for school policy. This is especially crucial if high-stakes are attached to the surveys, as unreliable measures can lead to ill-informed pedagogical choices.

2.2 Institutional Details

Starting in the 2012-2013 academic year, a large urban public school district initiated a principal, assistant principal and teacher compensation and evaluation reform with the aim of improving the quality of instruction. The district first undertook the principal reform and in the 2014-2015 academic year introduced a similar reform for assistant principals and

teachers. Unlike the prior compensation system that was largely determined by education and experience, the reform ties the compensation of educators to their evaluations. The evaluations of teachers are based primarily on three components: achievement, supervisor observations and student survey responses.[9]

On the basis of the grade and subject taught, the district uses four different methods to determine evaluation scores. Hence, teachers are divided into four categories: A, B, C and D (Figure A.1). Teachers are allotted points out of 100 for each component that contributes to their evaluation, and then evaluation scores are calculated from combining the components using category-specific weights. The district uses the two year average of the evaluation scores to assign seven score cutoffs, creating eight evaluation ratings, with a fixed proportion of teachers within each rating. In September of the next academic year, teachers learn about their evaluation score and rating for the previous academic year. With some qualifications, the evaluation rating determines the teacher compensation.[10] Compensation ranges from around $45,000 for teachers with the lowest evaluation rating to around $95,000 for those with the highest evaluation rating.

In this paper, I mainly focus on the subgroup of teachers that are evaluated on the basis of all three performance measures, that is Category A teachers (Figure A.1). Supervisor observations have the largest weight in the evaluation scores. The supervisors, typically principals, conduct observations between 4 to 8 times during the year. They follow a specific rubric to evaluate the teachers and assign them points out of 100. These are weighted based on the teacher category to produce the 'observation scores' for the supervisor observation component of the evaluation.

It is important to highlight the context under which the teachers receive observation scores by the principals. To increase the reliability of the observation scores, principals

[9]The achievement component is mainly based on standardized assessments and school performance.

[10]Evaluation ratings do not strictly determine the salary as the district allows for some exceptions in the rules. First, if the teacher salary prior to the reform is higher than the salary she should receive based on her evaluation rating, then she continues to receive her pre-reform salary. Second, for the years under study, a teacher cannot be paid a salary lower than what she received in the prior year.

receive annual training in observation scoring. Furthermore, part of their evaluation scores depend on the alignment of teacher contribution to student achievement, providing a strong incentive to rate teachers accurately. Thus, under the system one would expect the supervisor observations to be a more accurate depiction of teacher quality and help alleviate some concerns about this measure.

Student surveys are also an important component of evaluation for the Category A and C teachers. The surveys are conducted in the second week of April. Most students in grades 3 to 12 complete two surveys, one online and one on paper. The 'on paper' survey is administered in classrooms. Teachers are shuffled across classrooms for the administration of the surveys to ensure that they do not have any direct influence on the responses.[11] Teachers with fewer than 10 survey responses are not eligible to receive survey scores.

The survey questions are grouped into five components, with each component designed to measure a different aspect of the teaching quality. Table A.1 summarizes these five components and the information they contain. For each teacher, a score is calculated out of 100 for each of the five components. The average of the five component scores determines the 'raw survey score'. As student surveys carry a weight of 15% in the total evaluation score, the elementary, middle, and high school-levels separately use a target distribution to calculate points out of 15 (shown in Figure A.2) Separate school-level target distributions are used because younger students tend to evaluate more leniently.

3 Data

This section describes the data that I use for this study. I also present some descriptive statistics for the teachers and students in the analytic sample.

[11]The district outlines strict guidelines for the teachers in charge of administering the surveys. They are not allowed to interact with the students or answer any of their questions. Once the students complete the surveys, the results are sealed in an envelope.

3.1 Data Description

This paper uses student and teacher level data provided by the district and the state administrative system for the 2013-2014 to 2018-2019 academic years.[12] I link the students to their teachers and the panel nature of the data provides information on all students and teachers for the years that they are a part of the district public school system.

The district-provided data include the supervisor observations and student survey responses, allowing me to observe the yearly survey scores out of 15, as well as the raw and weighted observation scores.[13] For the empirical analysis, I standardize the survey scores out of 15 to have mean of zero and variance equal to one within a grade and year. I also standardize the weighted observation scores by teacher category, grade and year.

The state administrative system provides data on student and teacher characteristics. Student data includes information on gender, race, free and reduced price lunch status, limited English proficiency, special education status, absences and suspensions. Teacher data on race, gender, experience, education and salary is also available. I use end of year standardized math and reading test scores from grades 3-7 to measure student achievement. I standardize these scores to have mean of zero and variance equal to one for each subject, grade and year within the district. To assess whether student surveys can capture persistence of teacher quality, I also use the grade 9 (high school) data. For the high school test score outcomes, I focus on the Algebra I and Reading I tests. The district requires students to pass these tests in order to graduate high school.

It is important to highlight some data restrictions. First, as I observe math and reading test scores for students from grades 3-7, my empirical analysis is restricted to math and reading teachers who teach these grades. Second, when analyzing whether teacher evaluation measures are related to high school outcomes, I only use data for students from grades 6 and

[12] Even though the district introduced the student surveys in 2014-2015 academic year, I also use the 2013-2014 academic year data to account for lag test scores.

[13] Complete information on the raw survey score out of 100 is only available from the 2016-2017 to 2018-2019 academic years.

7 (middle school).[14] Due to the limited number of years used in this study, I do not have enough information to observe the elementary school students reaching high school. Thus, I restrict that analysis to middle school students.

3.2 Teacher and Student Characteristics

Table 1 presents information on teacher characteristics for grades 3-7 math and reading teachers. For the sample, the average total observation score is around 75 points out of 100. and the average student survey score is around 8 points out of 15. The teachers are more likely to be female and have an average professional experience of 8.7 years. Over 70% of the teachers are White or Black.

Table 2 shows the summary statistics for the students from grades 3-7. The majority are Hispanic (71%) or Black (22%), and almost 90% receive free or reduced price lunch. The average number of absences is is 5.3 days. While 8.5% of the students get suspended, it should be noted that most of these occur in middle school: 16% of the students in middle school are suspended as opposed to 4.5% in elementary school.

Approximately 78% students attend a public high school within the district. Note that this does not imply that the others do not go to high school. Some may be attending high school outside the district or be enroll in a private school, however, I cannot observe that. While the majority of students take the Algebra I and Reading I test in grade 9, it is important to highlight that in my sample 27.5% of students take the Algebra I test in grade 8. Therefore, I also assess whether teachers have any influence on when the students take the Algebra I test.

[14] I further restrict this analysis to only those cohorts who can reach high school by the 2018-2019 academic year.

4 Empirical Strategy

This section describes the empirical method employed to estimate the relationships between the performance measures and teacher effects on test scores and behaviors. I first describe the procedure used to estimate the measures of teacher effectiveness and then outline the model that relates them to student outcomes.

4.1 Estimation of the Teacher Effectiveness Measures

Estimates of test score, absence and suspension value-added are generated from the following specification:

$$Y_{ijgst} = f(A_{ijt-1}) + \beta X_{igst} + \eta C_{jgst} + \gamma_{st} + \eta_{gt} + \epsilon_{ijgst} \tag{1}$$

Let Y_{ijgst} be the standardized test score, indicator of suspension or number of absences for student i in year t, grade g and school s with teacher j. The equation controls for a cubic polynomial of prior achievement, lagged absences and suspensions ($f(A_{ijt-1})$); a vector of student controls (X_{igst}); classroom-level controls (C_{jgst}); and grade-by-year (η_{gt}) and school-by-year (γ_{st}) fixed effects. The student controls include race, gender, and indicators for free-reduced price lunch, limited-English proficiency and special education. C_{jgst} includes class size and classroom averages of lagged test scores, absences, suspensions and student controls.

Residualizing the error term and taking its average by year for each teacher produces the teacher-year value-added, $\bar{\hat{\epsilon}}_{j,t}$. Even though this method yields an unbiased estimate of value-added, there is a mechanical correlation between the student outcomes and estimated value-added. To avoid it, I use the weighted average of residuals in all other years, except for year t to estimate the leave-year-out value-added (denoted by μ_{jt}). Specifically, I use the

following to estimate the best linear predictor of μ_{jt}:

$$\hat{\mu}_{jt} = \hat{\rho}_{-t}\bar{\hat{\epsilon}}_{j,-t} \tag{2}$$

$\hat{\rho}_{-t}$ are the weights on the value-added estimates and they vary by the number of years before or after year t. The coefficients $\hat{\rho}_{-t}$ are obtained from an OLS regression of $\bar{\hat{\epsilon}}_{j,t}$ on a vector of $\bar{\hat{\epsilon}}_{j,-t}$. The above process yields leave-year-out value-added estimates that allow for teacher quality to vary over time. These estimates are shrunk to the mean through Bayesian shrinkage to account for noise (Chetty, Friedman, and Rockoff 2014a).

Parallel to the value-added estimation, one of the problems in determining the relationship between subjective evaluation measures and student outcomes arises from the issue of correlated errors (Kane and Staiger 2008; Chetty, Friedman, and Rockoff 2014a). The issue is exacerbated for student surveys. Theoretically, unobservable characteristics that may cause students to give teachers better (worse) evaluations may also cause them to have better (worse) outcomes. To avoid the problem of correlated errors, I use an approach similar to that used for the value-added estimation. I predict a teacher's standardized observation and survey score in year t using her scores in all others years except year t to avoid biasing the estimates (Jacob, Lefgren, and Sims 2010). Specifically, I estimate $\hat{s}_{jt} = \hat{\rho}_{-t}s_{j,-t}$. Here $s_{j,-t}$ is the standardized observation or survey score for the teacher in all other years except year t and $\hat{s}_{jt}$ is the leave-year-out supervisor observation or survey score for each teacher-year that account for drift.

4.2 Relationship Between Teacher Effectiveness Measures and Student Outcomes

After the construction of the leave-year-out estimates, I estimate whether teachers with high value-added or subjective scores improve student outcomes. While the value-added estimates include a rich set of controls, simply regressing the student outcomes on supervisor

observation and survey scores may not address concerns about sorting.

One issue is that there may be non-random sorting of students to teachers within and across schools. Parental socioeconomic status plays a significant role in sorting of students into schools. To address this concern, I include school-by-year fixed effects and a rich set of student controls in my model. Students are also likely to sort across teachers within schools, creating a source of potential bias. While the leave-year-out estimators address estimation issues caused by random variation across classrooms, they may not solve problems caused by persistent sorting. Persistent sorting patterns may occur if for example, teachers particularly good at classroom management are assigned students with a certain set of characteristics year after year. The baseline model controls for observed characteristics to account for such persistent sorting and I also conduct specification tests to test for selection on unobservables. I include lagged test score, absences and suspension to address sorting of students to teachers on the basis of cognitive and non-cognitive ability. Classroom level controls account for sorting at the group level. (Protic et al. 2013).

Combining the above mentioned strategies, I run the following regression to estimate the relationship between teacher effectiveness measures and achievement and behaviors:[15]

$$Y_{ijgst} = \alpha M_{jt} + f(A_{ijt-1}) + \beta X_{igst} + \eta C_{jgst} + \gamma_{st} + \eta_{gt} + \epsilon_{ijgst} \tag{3}$$

Here M_{jt} denotes the respective standardized leave-year-out performance measures ($\hat{\mu}_{jt}$ or $\hat{s}_{jt}$) and the controls are the same as those mentioned in Equation (1). I standardize the leave-year-out performance measures to have a mean of zero and standard deviation one so that α is comparable across different performance measures. I also run versions of Equation (3) that control for all the performance measures together to assess the relative strength of the subjective versus the objective measures. The main outcomes I focus on are math and

[15]Some students are taught math and reading by multiple teachers during the academic year, so they appear more than once in the sample each year. I weight the observations on the basis of the number of teachers that teach the student and the fraction of the year that they spend with the student (Hock and Isenberg, 2017).

reading test scores, number of absences and receipt of a suspension. When analyzing the absence and suspension outcomes, I pool the sample across the math and reading teachers. The standard errors are clustered two ways: at the student and teacher level (Jackson 2018; Cameron, Gelbach, and Miller 2011).

The key identifying assumption is that after conditioning on the control variables, students are not sorted to teachers based on unobservables. To explore this assumption, I conduct placebo tests in Section 6.4 where I test for sorting based on the twice-lagged outcomes and also estimate the results after removing sorting among classrooms by aggregating the measures to the school-by-grade-by-year level.

To estimate whether better performing teachers raise long-term skills, I modify Equation (3) by using the subsequent year or high school outcomes as the dependent variables. The analysis with high school outcomes is restricted to a limited sample (grades 6 and 7). In Section 6, when I refer to value-added, observation score and survey score, I am referring to the leave-year-out estimates.

Table A.2 shows the test score and non-test score value-added estimates obtained from Equation (2). The estimates show that the standard deviation of teacher effects is 0.109 in math and 0.043 in English. Magnitudes of value-added to absences and suspension is 0.107 standard deviations and 0.009 standard deviations, respectively. I will henceforth refer to them combined as 'behaviors value-added'.

5 Description of the Evaluation Measures

I start off by exploring the distribution of the raw observation and survey scores out of 100 for each school-level in Figure 1. While the observation scores follow a relatively similar pattern for all three school-levels (Figure 1a), there is a stark difference in the distribution of the survey scores for the elementary school versus the middle and high school, as elementary school students seem to be more generous in their evaluations (Figure 1b). The average

survey score received by elementary school teachers is approximately 85 points, while the average for middle and high school is 70. Although the distribution of survey scores is different for the 3 school-levels, the district accounts for it in the evaluation score calculation by separately ranking teachers for each school-level.

Next, I assess the stability of the observation and survey scores a teacher receives across years. One of the concerns of the student survey measure is that the students may not be sophisticated enough to evaluate teachers. The lack of student evaluation skills may be reflected in large fluctuations from year-to-year for the same teacher and rankings that appear not to be related to teacher effectiveness as conveyed by the other measures. However, there are two important considerations. First, a high year-to-year correlation between the measures for a teacher may not necessarily reflect her productivity. For example, prejudice or a popularity contest could be driving correlations over time in survey scores. Second, one should be mindful that a lack of correlation between the distinct evaluation measures may also suggest that they each pick up a different set of information, which may still reflect teacher effectiveness at producing important skills. Nevertheless, it is informative to explore the stability of the measures and their similarity to each other.

I conduct two analyses to investigate the stability of the subjective evaluation measures. First, I generate the score transition matrix to assess how the scores change from one year to the next. Less teacher movement from low (high) to high (low) scores will suggest that the measures are stable. Second, I estimate the teacher year-to-year correlation of different evaluation measures and examine how the subjective measures perform relative to value-added. I also assess the similarity between the different measures. To do so, I display the raw correlations between the different measures.

Tables 3a and 3b report the transition matrix where the cells show the percentage of teachers receiving the respective scores in years t and $t+1$. Each row sums up to 100%. Table 3a presents the quintile transition matrices for observation scores. Around 77% of the teachers in the bottom quintile stay in the bottom two quintiles in the subsequent year and

87% in the highest quintile remain in the top two quintiles in year $t+1$. Similarly, Table 3b reveals that approximately 70% of those who receive a survey score of 0 in year t end up with a score of 0 or 3 (top 2 scores) the next year and 68% with the highest survey score obtain either of the top 2 scores in the subsequent year. This indicates that different cohorts of students are responding consistently.

For further evidence of stability, Table 4 outlines the year-to-year correlations of the different measures of teacher quality. Observation and survey scores have a year-to-year correlation of 0.669 and 0.519, respectively. The year-to-year correlation of the subjective evaluation measures is slightly higher when compared to the test score value-added which has a correlation of 0.301. In general, Tables 3 and 4 suggest the observation and survey scores are not very noisy across years and show stability between two consecutive years.

Next, I explore the similarity between the different measures of teacher performance by estimating the correlations between them. As these measures may have some measurement error, the real association between them can differ from that represented by raw correlations.[16] To account for this, I use a split sample approach and estimate disattenuated correlations along with the raw ones (Tables 5a and 5b).[17] Overall, the the raw and disattenuated correlations follow a similar pattern.

As suggested by previous studies, I find some positive correlation between observation scores, survey scores and test score value-added (Raudenbush and Jean 2014). This shows that the subjective measures pick up some information that is also conveyed by test-score value-added. Nevertheless, these low correlations suggest that the performance measures may

[16]If variables are measured with error, the correlation between them is the product of the true correlation and the square root of the product of the reliability of each outcome (Spearman, 1904). Theoretically, the raw correlations may be biased downwards if there is noise in the estimates, or may be biased upwards if factors like unobserved classroom level shocks or student characteristics cause a correlation between the estimation errors across measures in the same year.

[17]To calculate the disattenuated correlations I follow the steps outlined in Jackson et al. (2022). Specifically, I divide the sample into even and odd years and estimate the correlation by using the following formula: $r_{12} = (\hat{\rho}_{12}^{even-odd}) / \sqrt{(\hat{\rho}_{11}^{even-odd} \hat{\rho}_{22}^{even-odd})}$. Here $\hat{\rho}_{12}^{even-odd}$ is the cross outcome correlation between even and odd year estimates; and $\hat{\rho}_{11}^{even-odd}$ and $\hat{\rho}_{22}^{even-odd}$ are the correlations between even and odd year estimates for the same outcomes. Refer to Jackson et al. (2022) for more details.

be capturing different aspects of teaching. In line with the literature on non-test score value-added, I observe that behaviors value-added have a low association with test score value-added (Jackson 2018; Petek and Pope 2016; Liu and Loeb 2019). The relationship between the subjective measures and behaviors value-added are also small. The low correlations between the different measures is consistent with evidence that teacher quality is multi-dimensional.

6 Results

I now present the results for the relationship between the evaluation measures and teacher effects on the development of cognitive and non-cognitive skills. Specifically, I answer whether each evaluation measure is 1) significantly related to contemporaneous student outcomes, 2) conveys information above and beyond the other measures, and 3) captures effects on the accumulation of persistent skills as reflected in future outcomes. I also test for heterogeneity based on the prior performance of the students. Finally, I present some falsification tests and check for robustness of the results.

6.1 Relationship Between the Evaluation Measures and Contemporaneous Outcomes

Table 6 presents the relationship between the evaluation measures and contemporaneous achievement. The top and bottom panels report the estimates for math and reading, respectively. The first three columns show the results for the high-stakes measures: test score value-added, supervisor observations and student surveys. The estimates reveal that teachers with high test score value-added or subjective scores raise both math and reading test scores. More specifically, on average a one standard deviation increase in test score value-added increases math and reading achievement by 0.110 and 0.039 standard deviations, respectively; a one standard deviation increase in observation score increases math and reading achieve-

ment by 0.095 and 0.044 standard deviations, respectively; and a one standard deviation increase in survey score increases math and reading achievement by 0.036 and 0.023 standard deviations, respectively. To put the magnitude of the coefficients in perspective, the estimates on test score value-added and supervisor observation score for math are comparable to the effects of reducing class size by approximately 5 students (Schanzenbach 2006). On the other hand, behaviors value-added have little relationship with contemporaneous achievement, with only absence value-added being related to reading test scores (column 4).

While the above mentioned results denote that subjective measures capture the causal effect of teachers on test scores, it is also crucial to determine whether they explain variability in the test scores once conditioned on test score value-added. To the extent that there is measurement error in test score value-added, the subjective measures may be related to contemporaneous achievement even after controlling for it. The last two columns of the table compare the incremental contributions of the subjective measures to the contributions of the other measures of teacher effectiveness. Column 5 puts together the three high-stakes measures and and column 6 also adds behaviors value-added.

Combining the high-stakes measures decreases the coefficient on observation score to almost half of the original effect for math and to almost three-fourths of the original effect for reading. Nevertheless, even after controlling for the other measures, the observation score is strongly connected to contemporaneous achievement. On the other hand, the coefficient on survey score remains only marginally significant for reading achievement. Given that behaviors value-added are weakly related to achievement, it is not surprising that the point estimates on the high-stakes measures remain largely unchanged once they are included in the specification.

Table 7 presents similar results for absences and suspensions. The estimates reveal that both the objective and subjective performance measures are related to contemporaneous behaviors. On average, having a teacher with a standard deviation higher test score value-added and survey score decreases absences by 0.03 and 0.06 days, respectively. Unlike the

estimates for achievement, controlling for test score value-added or supervisor observation does not reduce the coefficient on survey score for this outcome (column 5). If anything, conditional on the other measures, teachers with a high survey score are the most effective in reducing student absences. Although the magnitude of the teacher effect on absences as measured by absence value-added is small, it is not surprising. Absences in the elementary and middle schools are more likely driven by parental decisions, so students may have less discretion over them. Hence, one may expect that teachers are less likely to have a major impact on attendance. It is important to note though, that the coefficient on student survey in column 3 is almost three-fourths the size of the coefficient on absence value-added in column 4. This highlights that the student surveys are a fruitful measure of the causal effect of teachers on year t absences.

Test score value-added, supervisor observations, and survey scores are at least marginally related to contemporaneous suspensions. However, only the coefficient on test score value-added is highly significant with an effect size of -0.25 percentage points and retains its significance in the specification that puts the three measures together. As expected, suspension value-added is strongly related to receipt of a suspension.

Although I show the estimates for contemporaneous absences and suspensions, one should be cautious in their interpretation as they may not necessarily measure non-cognitive skills. The concern with using contemporaneous behaviors as an outcome is that the teachers may have a direct control over them, especially for suspensions. For example, a lenient teacher may be less likely to refer students for suspension and at the same time receive better supervisor and student evaluations across all years.[18] To avoid mechanical relations and manipulation, it is better to examine whether teachers with high value-added or subjective scores improve subsequent year behaviors, which I do below.

Overall, the results for contemporaneous outcomes provide some important information on the teacher evaluation measures. First, the estimates show that teachers with high sub-

[18]Mostly, the decision to suspend a student is in the hand of the administration, however, teachers may refer students for suspension if there is any misconduct.

jective scores improve skills. This finding is particularly important for student surveys. Considering the fact that students are not trained evaluators, they do a good job at evaluating the teachers as they explain variation in both contemporaneous achievement and behaviors. Second, even after conditioning on test score value-added, the subjective measures not only provide incremental information for the behaviors, but also test scores. This suggests that there is some noise in test score value-added and combining the different measures allows one to better capture the teacher effects on contemporaneous test scores.

6.2 Relationship between the evaluations measures and long-run outcomes

One of the concerns about using contemporaneous outcomes to measure teacher quality is that teachers may raise them through strategic behavior and other actions that do not lead to the acquisition of cognitive and non-cognitive skills. If supervisors and students give high ratings to the types of teaching that produces lasting skills, then one would expect these ratings to be related to both the current and future outcomes. Whether students assigned to high value-added or subjective score teachers perform better in future, provides important information on the potential usefulness of these measures. I estimate the relationship between the evaluation measures and subsequent year outcomes to capture teacher effects that persist into the future. Table 8 displays the results for subsequent year math and reading achievement.

The table shows that all three high-stakes measures are significantly related to subsequent year achievement. Compared to the results for the relationship between the high-stakes measures and contemporaneous achievement (columns 1, 2 and 3 in Table 6), the coefficients on test score value-added and observation score are reduced to almost one-fifth for math and half for reading, however the decay is lesser for the survey score estimates. One possible explanation for a greater reduction in math coefficients relative to reading is this there is more overlap in content for reading between years t and $t+1$. While all the measures are

significantly related to subsequent year math, once I put the high-stakes measures together, only the coefficient on test score value-added remains statistically significant. On average, a one standard deviation increase in test score value-added increases the subsequent year math achievement by 0.015 standard deviations (column 5). It is also important to note that suspension value-added is strongly related to subsequent year math scores, with an effect size of -0.026 standard deviations, and remains significant and similar in magnitude (-0.024) even after conditioning on the other measures.

Unlike the estimates for math, supervisor observations entail information for subsequent year reading not revealed by test score value-added. Specifically, column 5 reports that on average a one standard deviation increase in test score value-added increases the reading achievement in year $t+1$ by 0.012 standard deviations and a one standard deviation increase in the observation score increases the subsequent year reading by 0.011 standard deviations. Although student surveys also provide some additional information for subsequent year reading, the estimate is only marginally significant. These results demonstrate that the subjective measures capture aspects of teaching that are related to persistent reading skills and are not measured by test score value-added.

Table 9 presents similar results for the subsequent year behaviors. Teachers with high test score value-added or low suspension value-added are effective in reducing subsequent year absences, with suspension value-added being the most strongly related to this outcome. On average, a one standard deviation increase in test score value-added reduces subsequent year absences by 0.03 days and a one standard deviation increase in suspension value-added increases subsequent year absences by 0.05 days. On the other hand, none of the measures seem to be strongly related to subsequent year suspensions.

Given the findings for year $t+1$, I also explore whether the effects documented in the previous tables persist into high school. This analysis is restricted to middle school students (grades 6 and 7). Table 10 outlines the results. The top panel shows the high school test related outcomes. As mentioned before, while the majority of the students take the Algebra

I test in grade 9, a few take it in grade 8. Hence, I also measure if teachers have any influence on when the students decide to enroll in Algebra I. To put this outcome into perspective, one may think of it as reflecting the teacher effect on student readiness to take the class early. Estimates reveal that test score value-added is strongly related to when the students take Algebra I. Conditional on the high-stakes measures, a one standard deviation increase in test score value-added increases the probability that the students take the test in grade 8 by around 1.4 percentage points. Columns 1 and 2 also show a weak relationship between survey score and this outcome, with an effect size of 1 percentage point. Consistent with the results for subsequent achievement, the test score value-added is significantly related to Algebra I and Reading I test scores, conditional on the other measures. However, student surveys are also strongly related to Algebra I performance. On average, conditional on the high-stakes measures, a one standard deviation increase in survey score increases the Algebra I score by 0.015 standard deviations. This provides further evidence that teachers with high subjective scores improve cognitive skills.

I also present estimates for four high school non-test score outcomes: attending a public high school within the district, being held back, number of absences and probability of suspension in grade 9. With the exception of the first outcome, the rest are commonly used in the literature to measure teacher effects. Districts often care about maintaining their enrollment, so it is intuitive to test whether the evaluation measures capture teacher ability to attract students to remain within the public school district education system. Students assigned to teachers who are generally better liked are more likely to remain in the district for high school. In the specification that puts all the performance measures together, on average, a one standard deviation increase in teacher quality, as measured by the survey score, increases the probability that the student attends a public high school in the district by 0.4 percentage points. This suggests that teachers with better student survey scores possibly make the students feel more positive about remaining in the district.

Similar to the estimates for subsequent year absences, test score value added is signifi-

cantly related to high school absences, with the effect size being more pronounced (-0.13 in column 5). Only absence value-added is significantly related to the probability of being held back (0.2 percentage points). In line with the findings for subsequent year suspensions, none of the high-stakes measures are related to high school suspensions.

Broadly, the results in this section suggest that teachers have an impact on the long-term outcomes and these effects are explained by both the objective and subjective measures. To to a certain extent the subjective measures are significantly related to teacher effects on long-run achievement and provide information not captured by test score value-added. However, in this multiple measure setting, teachers with high test score value-added are the most successful in improving lower-stakes long-term achievement and absences.

6.3 Heterogeneity by prior outcomes

Schools often care about the performance of the at-risk students and want to implement strategies to boost their outcomes. Hence, I also examine the heterogeneity of the teacher effects by year $t-1$ test scores and absences. The rationale behind this analysis is to assess the relative strength of the objective versus the subjective measures in capturing achievement and behaviors of the high versus low performing students. We may expect differences in the estimates for the two groups if for example, low achievers are worse at evaluating teacher quality or principals only base their ratings on the performance of the high achieving students. I divide students based on their year $t-1$ test scores and absences being above or below the median, and run separate regressions for the two groups. Tables 11 and 12 present the results for the contemporaneous and subsequent year outcomes, respectively.

Table 11 shows some interesting differences in the results for the two groups. Conditional on the measures of teacher quality, test score value-added is equally informative of the achievement of the high and low performing students. On the other hand, survey score seems to be more related to the test scores of the low achieving students, once conditioned on the other measures. The magnitude of the effect is almost three times larger for the students

with higher prior absences, compared to the magnitude for those with lower absences. This shows that effective teachers make it more attractive for students with higher prior absences to attend school. Overall, these findings regarding the student surveys support the idea that a considerate and caring teacher may particularly matter more for the at-risk students.

The estimates for observation score are slightly mixed. The observation score seems to be strongly connected to the math achievement of the students with prior math test scores above the median, but is almost equally related to the reading achievement of the two groups. The general pattern of the findings is similar for the subsequent year outcomes (Table 12). One major exception is that teachers with high test score value-added reduce the subsequent year absences of the students with high prior absences, and seem to do nothing for those with low prior absences.

6.4 Falsification tests and Robustness checks

I test the identifying assumptions and check for the robustness of the estimates to alternate specifications. The results in this paper rely on the key assumption that after conditioning on the control variables, students are not sorted to teachers based on unobservables. The leave-year-out estimators address issues caused by mechanical correlation between the treatment and outcome, however, one may be concerned about persistent sorting. Following Chetty, Friedman, and Rockoff (2014a), I perform two analyses to test for bias and the results broadly suggest a lack of bias in the estimates.

I conduct a placebo test for selection on the observables by examining whether students assigned to better teachers are expected to have higher outcomes based on their characteristics (first 4 columns in Table A.3). I test for sorting on the twice-lagged outcomes. First, I predict residualized student achievement and behaviors using two year lags of the outcomes.[19] Next, I regress these predicted outcomes on the evaluation measures, conditioning on the controls used in Equation (3). The sample is restricted to students with data on both

[19] I residualize achievement and behaviors using the controls mentioned in Equation (1)

lagged and twice-lagged outcomes. The results in all four columns of Table A.3 reveal that none of the subjective measures are strongly related to the predicted outcomes. Although, there are some statistically significant relationships, the magnitude of the effects is small.

I also test for selection on unobservables by aggregating the performance measures to the school-by-grade-by-year level. Aggregating the treatment avoids bias resulting from the sorting of the students to the teachers within the school. If the reduced form estimates are similar to the estimates obtained from the empirical strategy that I use in this paper, it will strengthen confidence in the results. The last four columns in Table A.3 present the results. The magnitudes of the effects are consistent with the results presented in Tables 6 and 7. These estimates help rule out the concern that sorting of students to teachers may be driving the main results.

To test for the robustness of the results, I exclude the school-by-year fixed effects from the value-added model and rerun the main regressions. Removing school-by-year fixed effect changes the source of variation for the value-added estimation. Furthermore, policy relevant value-added models usually do not control for school-by-year fixed effects. Hence, I test the sensitivity of the results to their exclusion from value-added. Table A.4 establishes that their absence does not affect the basic pattern of the estimates. These findings corroborate the notion that the results are not driven by selection.

I also test the sensitivity of the estimates by residualizing the subjective measures by the classroom averages of the baseline controls and the grade-by-year and school-by-year fixed effects. Although I address concerns about sorting by adjusting for covariates in the estimation of the relationship between subjective measures and student outcomes, I do not residualize the subjective measures in the first stage. One may be concerned that while the value-added estimates explicitly include controls, not incorporating them in the leave-year-out subjective measure estimation may affect the signal value of the observation and survey scores and not facilitate an apples-to-apples comparison between the value-added and subjective measure estimates. I test whether the coefficients change when I residualize the

observation and survey scores when estimating the leave-year-out measures. Table A.5 yields very similar estimates to those presented previously in this section, which reassures that the results represent real effects.

7 Conclusion

This paper investigates the relationships between the objective and subjective evaluation measures and the teacher effects on cognitive and non-cognitive skills, in a system that attaches high-stakes to test score value-added, supervisor observations and student surveys. I use test scores as a proxy for cognitive skills and absences and receipt of a suspension as a proxy for non-cognitive skills.

The findings provide compelling evidence on the effectiveness of the teacher evaluation measures. Both the objective and subjective measures are significantly related to contemporaneous achievement and behaviors. Correlation between the high-stakes measures is positive and suggests that the subjective measures pick up some variation that is also explained by test score value-added. Despite this positive correlation, the subjective measures explain significant variability in teacher effects on test scores and behaviors, even after conditioning on value-added. In fact, of the three measures, student surveys are the most strongly related to contemporaneous absences.

To better evaluate the efficacy of the evaluation measures, I also document whether teachers with high value-added or subjective scores improve long-run outcomes. The long-run outcomes include subsequent year and high school outcomes. Test score value-added has persistent effects on subsequent year and high school achievement and absences. The subjective measures are also related to long run achievement. Even after conditioning on test score value-added, supervisor observations have a significant relationship with subsequent year reading and student surveys provide some additional information on Algebra I performance. Teachers with high survey scores are also effective at retaining the students in the public

school district for high school. It is important to note that in the multi-measure framework value-added remains most strongly related to long-term achievement and absences.

Despite the fact that test score value-added does a better job at capturing student cognitive and non-cognitive skill production, it does not diminish the importance of the other measures. A good performance measure mitigates incentive distortions and provides information on employee productivity in ways that is not explained by the other measures. Although I show the subjective measures explain the causal effect of teachers, one of the limitations of this study is that I am unable to observe the relationship between the measures when only a single measure is incentivized. It is likely that the presence of the subjective measures may be helping better align teacher effort with valued outcomes and the strong relationships between test-score value-added and low-stakes outcomes may not have existed if teachers were only incentivized on the basis of objective measures. Furthermore, while test score value-added is more significantly related to student outcomes, it is important to note that the student survey measure provides more information on the performance of the at-risk students. This information is important for policy makers who aim to retain and improve the outcomes of the students who may be on the margin of dropping out.

Although the results illustrate that all three high-stakes measures have a relationship with both contemporaneous and longer-term outcomes, one should be cautious in drawing broad policy conclusions from these findings. More work is needed to understand how the system as a whole affects learning and whether the policy implementation raises overall skills. Also, this paper does not guide on how the relationship between the measures would change if high-stakes are attached to only one of these measures. It is possible that moving to a system where teacher compensation is only tied to a subset of the measures, or where the measures no longer have incentives attached would change the conclusions. These measures are costly to implement, and further work is needed to fully comprehend the effectiveness of the evaluation system. Nevertheless, the fact that several states and districts implement high-stakes evaluations, it is reassuring to know that both test score value-added and subjective

evaluations are fruitful measures of teacher quality.

References

Bacher-Hicks, Andrew, Mark J Chin, Thomas J Kane, and Douglas O Staiger. 2019. "An Experimental Evaluation of Three Teacher Quality Measures: Value-added, Classroom Observations, and Student Surveys." *Economics of Education Review*

Baker, George P. 1992. "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, 100 (3), 598-614.

Boring, A. 2017. Gender biases in student evaluations of teaching. *Journal of Public Economics* 145, 27–41.

Braga, M., Paccagnella, M., Pellizzari, M. 2014. "Evaluating students evaluations of professors." *Economics of Education Review* 41, 71–88

Cameron, Colin, Jonah Gelbach, and Douglas L. Miller. 2011. "Robust Inference with Multi-way Clustering." *J. Bus. and Econ. Statis.* 21 (2): 238–49.

Carrell, S. E., and West, J. E. (2010). "Does professor quality matter? Evidence from random assignment of students to professors." *Journal of Political Economy*, 118, 409–432.

Chaplin D., Gill B., Thompkins A., and Miller H. 2014. *Professional Practice, Student Surveys, and Value-Added: Multiple Measures of Teacher Effectiveness in the Pittsburgh Public Schools.* Mathematica Policy Research report

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a."Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review*, 104(9): 2593-2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, 104(9): 2633-2679.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2007. "Teacher credentials and student achievement: Longitudinal analysis with student fixed effects." *Economics of Education Review*, 26(6): 673–682

Corcoran, Sean P., Jennifer L. Jennings, and Andrew A. Beveridge. 2012. *"Teacher Effectiveness on High- and Low-Stakes Tests"* Working Paper." New York, NY: New York University

Ewing, A. M. 2012."Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review*, 31(1), 141–154.

Frederiksen, A., Lange, F., Kriechel, B. 2017. "Subjective Performance Evaluations and Employee Careers." *Journal of Economic Behavior and Organization*, 134: 408-429.

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

Heckman, James J., and Yona Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review*. 91 (2): 145–49.

Hock, Heinrich and Isenberg, Eric. 2017. "Methods for Accounting for Co-Teaching in Value-Added Models", *Statistics and Public Policy*

Holmstrom, Bengt and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design," *Journal of Law, Economics and Organization*, 7, 24-52.

Jackson, Clement Kirabo. 2018. "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes." *Journal of Political Economy*, 126(5): 2072-2107

Jackson, C. Kirabo; Porter, Shanette; Easton, John and Kiguel, Sebastián. 2022. *Who Benefits from Attending Effective Schools? Examining Heterogeneity in High School Impacts* Unpublished Manuscript

Jacob, Brian A., and Lars Lefgren. 2005a. "Principals as agents: Subjective performance measurement in education." NBER Working Paper no. 11463, National Bureau of Economic Research, Cambridge, MA.

Jacob, Brian A., and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education" *Journal of Labor Economics*

Jacob, Brian A., Lars Lefgren, and David P. Sims. 2010. "The Persistence of Teacher-Induced Learning." *Journal of Human Resources*, 45(4): 915-943.

Kane, Thomas J., and Cantrell, S. 2010. *Learning about teaching: Initial findings from the measures of effective teaching project.* Seattle: The Bill and Melinda Gates Foundation.

Kane, Thomas J., and Staiger, Douglas O. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." *National Bureau of Economic Research* Working Paper No. 14607.

Kane, T. J., and Staiger, Douglas O. 2012. *Gathering feedback for teaching: Combining*

high-quality observations with student surveys and achievement gains. Seattle: The Bill and Melinda Gates Foundation.

Kautz, Tim, and Wladimir Zanoni. 2014. *Measuring and Fostering Non-cognitive Skills in Adolescence: Evidence from Chicago Public Schools and the OneGoal Program.* Manuscript, Univ. Chicago

Kraft, M.A. 2019. Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. *Journal of Human Resources* 54(1), 1-36.

Krautmann, A. C., and Sander, W. 1999. "Grades and student evaluations of teachers". *Economics of Education Review*, 18(1), 59–63. doi:10.1016/s0272-7757(98)00004-1

Lindqvist, E., and R. Vestman. 2011. "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment." *American Economic Journal: Applied Economics* 3(1): 101–28.

Liu, J., and Loeb, S. 2019. "Engaging Teachers: Measuring the Impact of Teachers on Student Attendance in Secondary School". *Journal of Human Resources*, 1216–8430R3. doi:10.3368/jhr.56.2.1216-8430r3

Marsh, H. W., and Roche, L. A. 1997. "Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility." *American Psychologist*, 52(11), 1187. doi:10.1037/0003-066X.52.11.1187

Michela B., Marco P., Michele P. 2014. "Evaluating students' evaluations of professors", *Economics of Education Review,* Volume 41, Pages 71-88, ISSN 0272-7757

Murphy, K.J. 1992. *Performance measurement and appraisal: motivating managers to identify and reward performance.* In: Burns, W.J. (Ed.), Performance Measurement, Evaluation, and Incentives. Harvard Business School Press, Boston, MA, pp. 37–62.

Neal, Derek. 2011. *The Design of Performance Pay in Education*, Handbook of the Economics of Education, Vol. 4, Elsevier, pp. 495-550.

Petek, Nathan and Pope, Nolan G. 2016. *The Multidimensional Impact of Teachers on Students.* Job Market Paper

Protik, Ali, Elias Walsh, Alexandra Resch, Eric Isenberg, and Emma Kopa. 2013. *"Does Tracking of Students Bias Value-Added Estimates for Teachers?"* Working paper, Mathematica Policy Res., Princeton, NJ.

Raudenbush, Stephen and Jean, Marshall. 2015. *To What Extent Do Student Perceptions of Classroom Quality Predict Teacher Value Added.* 10.1002/9781119210856.ch6.

Redding C. 2019. "A Teacher Like Me: A Review of the Effect of Student–Teacher Racial/Ethnic Matching on Teacher Perceptions of Students and Student Academic and Behavioral Outcomes" *Review of Educational Research*

Ross, E. and Walsh, K. 2019. State of the States 2019: *Teacher and Principal Evaluation Policy.* Washington, DC: National Council on Teacher Quality.

Schanzenbach, Diane Whitmore. 2006. *What Have Researchers Learned from Project STAR?*, Brookings Papers on Education Policy, (9), 205–228. Publisher: Brookings Institution Press

Spearman, C. 1904. The proof and measurement of association between two things. *The American Journal of Psychology,* 15(1), 72–101.

Steinberg, Matthew P., and Rachel Garrett. 2016. "Classroom Composition and Measured Teacher Performance: What Do Teacher Observation Scores Really Measure?" *Educational Evaluation and Policy Analysis*, 38(2): 293-317.

Wagner, N., Rieger, M., Voorvelt, K. 2016. Gender, ethnicity and teaching evaluations: evidence from mixed teaching teams. *Economics of Education Review* 54 (54), 7994.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., and Morgan, K.2009. *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New York, NY: TNTP (The New Teacher Project).

Whitehurst, Grover J., Chingos Matthew M. and Lindquist Katharine M. 2014. *"Evaluating Teachers with Classroom Observations Lessons Learned in Four Districts"* Washington, DC: The Brookings Institution.

Wiswall, Matthew. 2013. "The dynamics of teacher quality." *Journal of Public Economics* 100: 61-78.

Figures

Figure 1a: Distribution of raw observation scores by school-level

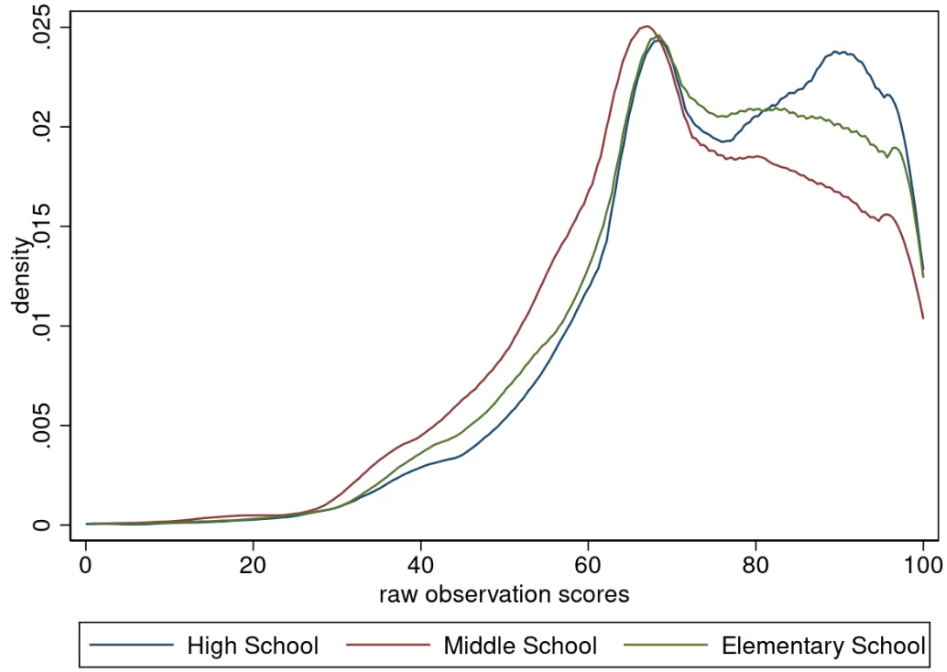
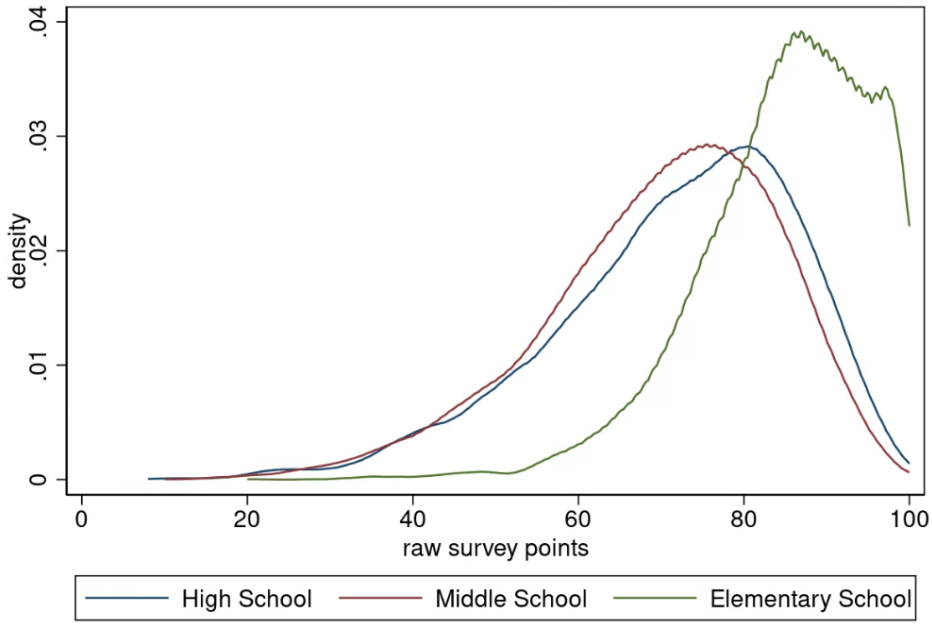


Figure 1b: Distribution of raw survey scores by school-level



Notes: Figure 1b is plotted using data from 2016-2017 to 2018-2019 academic years.

Tables

Table 1: Summary Statistics for teachers

	Grades 3-7 Math and Reading teachers	
	Mean	SD
Class size	20.391	6.282
Male	0.216	0.412
Professional experience years	8.718	8.949
District experience years	6.730	7.812
Bachelors	0.726	0.446
Masters	0.253	0.435
White	0.341	0.474
Black	0.359	0.480
Hispanic	0.203	0.402
Survey points (out of 15)	8.281	3.502
Supervisor observation points (out of 100)	75.419	16.451
Teacher-year Observations	10,660	

Notes: The table presents summary statistics for all grades 3-7 math and reading teachers. The last two rows specify the scores that contribute to the evaluation score calculation. *Professional experience years* refers to the years of experience that a teacher has within the education sector.

Table 2: Summary statistics for student characteristics

	(1) Mean	(2) SD	(3) N
Standardized Math Score	0.027	0.999	284,499
Standardized Reading Score	0.011	0.996	281,279
Retained	0.009	0.092	286,036
Total Absences	5.342	6.306	285,977
Suspended	0.085	0.279	286,036
Male	0.515	0.500	286,036
White	0.046	0.209	286,036
Black	0.222	0.416	286,036
Hispanic	0.710	0.454	286,036
Asian	0.012	0.109	286,036
Receive Free Reduced Price Lunch	0.902	0.298	286,036
Limited English Proficiency	0.511	0.500	286,036
Receive Special Education	0.073	0.261	284,701
High School Variables			
Attend HS in the public school district	0.778	0.416	61,949
Algebra I score	0.125	0.971	44,997
English I score	0.099	0.926	44,179
Take Algebra I test in grade 8	0.275	0.447	44,997
Retained in grade 9	0.039	0.195	48,389
Total Absences in grade 9	10.586	15.917	48,389
Suspended in grade 9	0.117	0.322	48,389

Notes: The unit of observation is a student-year. The sample in the top panel includes all students in grades 3-7 who are matched to a classroom. The high school variables in the bottom panel are presented for cohorts of grades 6 and 7 students, who can reach high school by the 2018-2019 academic year. All high school variables, except for the first one, are based on the sample of students who attend high school within the district.

Table 3a: Stability of observation scores across years

Quintile	Quintile of observation score in $t+1$						
year t	1	2	3	4	5	Total(%)	Row Total
1	49.27	28.28	14.27	5.99	2.19	100.00	5,657
2	19.57	34.86	27.24	13.05	5.27	100.00	6,221
3	8.65	21.68	34.04	24.30	11.33	100.00	6,861
4	3.56	9.22	21.03	39.70	26.49	100.00	7,209
5	1.10	3.62	7.66	23.73	63.89	100.00	7,523
Total	15.15	18.98	20.93	21.99	22.95	100.00	
Column Total	4,967	6,221	6,861	7,209	7,523		32,781

Notes: The unit of observation is a teacher-year. The table is constructed using all Category A, B, C and D teachers (that is those who receive observation scores). Refer to Figure A.1 to see the four teacher categories.

Table 3b: Stability of survey scores across years

	Survey points $t+1$							
Survey points t	0	3	6	9	12	15	Total(%)	Row Total
0	30.22	39.11	18.89	10.22	1.11	0.44	100.00	450
3	8.55	33.43	35.58	19.26	2.15	1.03	100.00	2,046
6	2.08	16.24	38.59	36.20	4.94	1.95	100.00	4,716
9	0.49	4.91	22.35	51.98	13.64	6.63	100.00	7,713
12	0.30	1.62	8.64	45.28	24.50	19.67	100.00	2,339
15	0.18	0.61	3.78	27.54	24.68	43.21	100.00	1,641
Total (%)	2.42	10.86	24.44	40.56	12.23	9.49	100.00	
Column Total	457	2,053	4,621	7,667	2,312	1,795		18,905

Notes: The unit of observation is a teacher-year. The table is constructed using all Category A and C teachers (that is those who receive survey scores). Refer to Figure A.1 to see the four teacher categories.

Table 4: Teachers' year-to-year correlation of different performance measures

	Correlation
Test score VA in year one and year two	0.301
Standardized observation score in year one and year two	0.669
Standardized survey score in year one and year two	0.519
Absence VA in year one and year two	0.197
Suspension VA in year one and year two	0.431

Notes: The unit of observation is a teacher-year. The year-to-year correlation is the teachers' year t and $t+1$ correlation of the different measures of performance.

Table 5a: Cross-correlation of measures of teacher performance

Variables	Test score VA	Observation Score	Survey Score	Absence VA	Suspension VA
Test score VA	1.000				
Observation score	0.270	1.000			
Survey score	0.167	0.318	1.000		
Absence VA	-0.085	-0.044	-0.060	1.000	
Suspension VA	-0.077	-0.028	-0.017	0.275	1.000

Notes: The table reports the raw correlations between the different measures of teacher performance. Test score value-added is based on math and reading value-added. All measures are standardized to have a mean of 0 and standard deviation 1.

Table 5b: Disattenuated Cross-correlation of measures of teacher performance

Variables	Test score VA	Observation Score	Survey Score	Absence VA	Suspension VA
Test score VA	1.000				
Observation score	0.361	1.000			
Survey score	0.195	0.366	1.000		
Absence VA	-0.143	-0.036	-0.084	1.000	
Suspension VA	-0.101	-0.056	-0.003	0.285	1.000

Notes: The table reports the disattenuated correlations between the different measures of teacher performance. (Refer to section 6.1 for details on the estimation process.) Test score value-added is based on math and reading value-added. Survey scores and observation scores are standardized.

Table 6: Relationship between performance measures and contemporaneous achievement

	Math Test Scores					
	(1)	(2)	(3)	(4)	(5)	(6)
Test score VA	0.1098***				0.0960***	0.0967***
	(0.0036)				(0.0040)	(0.0040)
Standardized observation score		0.0947***			0.0440***	0.0437***
		(0.0061)			(0.0044)	(0.0044)
Standardized survey score			0.0358***		-0.0009	-0.0005
			(0.0064)		(0.0038)	(0.0038)
Absence VA				-0.0080		0.0031
				(0.0068)		(0.0033)
Suspension VA				-0.0095		0.0017
				(0.0062)		(0.0031)
Observations	121,056	121,056	121,056	121,056	121,056	121,056
Adjusted R-squared	0.6711	0.6654	0.6616	0.6610	0.6719	0.6719
	Reading Test Scores					
	(1)	(2)	(3)	(4)	(5)	(6)
Test score VA	0.0388***				0.0315***	0.0315***
	(0.0031)				(0.0032)	(0.0032)
Standardized observation score		0.0439***			0.0303***	0.0304***
		(0.0040)			(0.0036)	(0.0036)
Standardized survey score			0.0229***		0.0058*	0.0057*
			(0.0037)		(0.0031)	(0.0031)
Absence VA				-0.0063**		-0.0022
				(0.0032)		(0.0024)
Suspension VA				0.0018		0.0035
				(0.0031)		(0.0025)
Observations	133,706	133,706	133,706	133,706	133,706	133,706
Adjusted R-squared	0.6308	0.6305	0.6298	0.6295	0.6313	0.6313

Notes: The table reports coefficients from the regression of the indicated outcomes on the leave-year-out measures of teacher quality. All columns control for lagged $(t-1)$ test score, absences and indicator for suspensions, student controls, classroom controls, grade-by-year FE and school-by-year FE. Standard errors in the parenthesis are clustered two-ways: at student and teacher level.
* p<0.1, ** p<0.05, *** p<0.01

Table 7: Relationship between performance measures and contemporaneous behaviors

	Absences					
	(1)	(2)	(3)	(4)	(5)	(6)
Test score VA	-0.0272**				-0.0216	-0.0126
	(0.0135)				(0.0140)	(0.0119)
Standardized observation score		-0.0167			0.0124	0.0119
		(0.0144)			(0.0152)	(0.0139)
Standardized survey score			-0.0572***		-0.0564***	-0.0488***
			(0.0132)		(0.0139)	(0.0125)
Absence VA				0.0896***		0.0859***
				(0.0176)		(0.0173)
Suspension VA				0.0234*		0.0211
				(0.0140)		(0.0139)
Observations	254,760	254,760	254,760	254,760	254,760	254,760
Adjusted R-squared	0.3812	0.3812	0.3812	0.3814	0.3812	0.3814
	Suspended					
	(1)	(2)	(3)	(4)	(5)	(6)
Test score VA	-0.0025***				-0.0022**	-0.0012
	(0.0009)				(0.0009)	(0.0008)
Standardized observation score		-0.0018*			-0.0005	-0.0009
		(0.0010)			(0.0011)	(0.0009)
Standardized survey score			-0.0017*		-0.0011	-0.0007
			(0.0009)		(0.0009)	(0.0008)
Absence VA				0.0008		0.0006
				(0.0009)		(0.0008)
Suspension VA				0.0088***		0.0087***
				(0.0011)		(0.0011)
Observations	254,762	254,762	254,762	254,762	254,762	254,762
Adjusted R-squared	0.1925	0.1925	0.1925	0.1933	0.1925	0.1933

Notes: The table reports coefficients from the regression of the indicated outcomes on the leave-year-out measures of teacher quality. The estimates are based on a pooled sample across math and reading teachers. All columns control for lagged $(t-1)$ test score, absences and indicator for suspensions, student controls, classroom controls, grade-by-year FE and school-by-year FE. Standard errors in the parenthesis are clustered two-ways: at student and teacher level.
* p<0.1, ** p<0.05, *** p<0.01

Table 8: Relationship between performance measures and subsequent year achievement

	Math Test Scores in $t+1$					
	(1)	(2)	(3)	(4)	(5)	(6)
Test score VA	0.0186***				0.0152**	0.0112*
	(0.0060)				(0.0068)	(0.0064)
Standardized observation score		0.0177***			0.0077	0.0104
		(0.0058)			(0.0070)	(0.0068)
Standardized survey score			0.0107*		0.0040	0.0033
			(0.0059)		(0.0067)	(0.0068)
Absence VA				-0.0047		-0.0043
				(0.0059)		(0.0059)
Suspension VA				-0.0264***		-0.0237***
				(0.0068)		(0.0065)
Observations	80,241	80,241	80,241	80,241	80,241	80,241
Adjusted R-squared	0.6059	0.6058	0.6057	0.6060	0.6060	0.6062
	Reading Test Scores in $t+1$					
	(1)	(2)	(3)	(4)	(5)	(6)
Test score VA	0.0159***				0.0117**	0.0118**
	(0.0044)				(0.0047)	(0.0047)
Standardized observation score		0.0188***			0.0109**	0.0109**
		(0.0048)			(0.0049)	(0.0049)
Standardized survey score			0.0156***		0.0083*	0.0084*
			(0.0045)		(0.0046)	(0.0046)
Absence VA				-0.0001		0.0023
				(0.0045)		(0.0044)
Suspension VA				-0.0040		-0.0015
				(0.0036)		(0.0035)
Observations	86,232	86,232	86,232	86,232	86,232	86,232
Adjusted R-squared	0.5718	0.5717	0.5717	0.5716	0.5719	0.5719

Notes: The table reports coefficients from the regression of the indicated outcomes on the leave-year-out measures of teacher quality. All columns control for lagged $(t-1)$ test score, absences and indicator for suspensions, student controls, classroom controls, grade-by-year FE and school-by-year FE. Standard errors in the parenthesis are clustered two-ways: at student and teacher level.
* p<0.1, ** p<0.05, *** p<0.01

Table 9: Relationship between performance measures and subsequent year behaviors

	Absences *t+1*					
	(1)	(2)	(3)	(4)	(5)	(6)
Test score VA	-0.0335**				-0.0343**	-0.0299*
	(0.0156)				(0.0168)	(0.0163)
Standardized observation score		-0.0131			-0.0001	-0.0024
		(0.0164)			(0.0183)	(0.0183)
Standardized survey score			-0.0033		0.0049	0.0052
			(0.0162)		(0.0173)	(0.0174)
Absence VA				-0.0209		-0.0217
				(0.0214)		(0.0214)
Suspension VA				0.0564**		0.0508**
				(0.0248)		(0.0241)
Observations	164,826	164,826	164,826	164,826	164,826	164,826
Adjusted R-squared	0.2896	0.2896	0.2896	0.2896	0.2896	0.2896
	Suspended *t+1*					
	(1)	(2)	(3)	(4)	(5)	(6)
Test score VA	0.0004				0.00003	0.00001
	(0.0009)				(0.00093)	(0.00094)
Standardized observation score		0.0016			0.0020*	0.0020*
		(0.0010)			(0.0011)	(0.0011)
Standardized survey score			-0.0005		-0.0012	-0.0012
			(0.0009)		(0.0010)	(0.0010)
Absence VA				-0.0004		-0.0004
				(0.0011)		(0.0011)
Suspension VA				-0.00002		-0.0001
				(0.00125)		(0.0013)
Observations	167,762	167,762	167,762	167,762	167,762	167,762
Adjusted R-squared	0.1207	0.1207	0.1207	0.1207	0.1207	0.1207

Notes: The table reports coefficients from the regression of the indicated outcomes on the leave-year-out measures of teacher quality. The estimates are based on a pooled sample across math and reading teachers. All columns control for lagged $(t-1)$ test score, absences and indicator for suspensions, student controls, classroom controls, grade-by-year FE and school-by-year FE. Standard errors in the parenthesis are clustered two-ways: at student and teacher level

* p<0.1, ** p<0.05, *** p<0.01

Table 10: Relationship between performance measures and high school outcomes

	Take Algebra I in grade 8		Algebra I Test score		Reading I Test score	
	(1)	(2)	(3)	(4)	(5)	(6)
Test score VA	0.0141***	0.0129***	0.0190***	0.0191***	0.0108**	0.0099**
	(0.0039)	(0.0039)	(0.0059)	(0.0062)	(0.0049)	(0.0050)
Standardized observation score	-0.0008	-0.0001	-0.0042	-0.0045	0.0028	0.0031
	(0.0059)	(0.0058)	(0.0086)	(0.0086)	(0.0059)	(0.0060)
Standardized survey score	0.0102*	0.0101*	0.0151**	0.0145*	-0.0051	-0.0053
	(0.0057)	(0.0056)	(0.0076)	(0.0077)	(0.0042)	(0.0043)
Absence VA		0.0004		-0.0039		-0.0052
		(0.0034)		(0.0056)		(0.0034)
Suspension VA		-0.0042		0.0029		0.0033
		(0.0026)		(0.0050)		(0.0033)
Observations	25,021	25,021	25,021	25,021	27,652	27,652
Adjusted R-squared	0.6655	0.6656	0.4596	0.4596	0.6014	0.6015

	Attend HS in the district		Held back in grade 9		Absences in grade 9		Suspended in grade 9	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test score VA	-0.0004	-0.0002	-0.0003	-0.0004	-0.1340**	-0.1193**	-0.0001	0.00001
	(0.0013)	(0.0013)	(0.0009)	(0.0009)	(0.0619)	(0.0581)	(0.0012)	(0.00123)
Standardized observation score	-0.0003	-0.0004	0.0001	0.0003	-0.0143	-0.0135	0.0006	0.0006
	(0.0015)	(0.0015)	(0.0010)	(0.0010)	(0.0581)	(0.0576)	(0.0013)	(0.0013)
Standardized survey score	0.0038***	0.0038***	-0.0004	-0.0003	0.0570	0.0669	0.0008	0.0009
	(0.0014)	(0.0014)	(0.0008)	(0.0008)	(0.0504)	(0.0512)	(0.0012)	(0.0012)
Absence VA		-0.0008		0.0017**		0.0770		0.0014
		(0.0012)		(0.0008)		(0.0474)		(0.0010)
Suspension VA		0.0017		-0.0007		0.0237		-0.0004
		(0.0010)		(0.0007)		(0.0434)		(0.0010)
Observations	67,624	67,624	56,324	56,324	56,324	56,324	56,324	56,324
Adjusted R-squared	0.0490	0.0490	0.0807	0.0809	0.1727	0.1728	0.1150	0.1150

Notes: The table reports coefficients from the regression of the indicated outcomes on the leave-year-out measures of teacher quality. The estimates in the bottom panel are based on a pooled sample across math and reading teachers. All columns control for lagged $(t-1)$ test score, absences and indicator for suspensions, student controls, classroom controls, grade-by-year FE and school-by-year FE. Standard errors in the parenthesis are clustered two-ways: at student and teacher level.
* p<0.1, ** p<0.05, *** p<0.01

Table 11: Relationship between performance measures and contemporaneous outcomes - By prior achievement and absence

	Math Test Score		Reading Test Score		Absences	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: t-1 outcome below median						
Test score VA	0.0897***	0.0909***	0.0345***	0.0346***	-0.0211*	-0.0154
	(0.0052)	(0.0053)	(0.0041)	(0.0041)	(0.0116)	(0.0105)
Standardized observation score	0.0316***	0.0308***	0.0346***	0.0347***	0.0065	0.0063
	(0.0055)	(0.0056)	(0.0044)	(0.0044)	(0.0124)	(0.0118)
Standardized survey score	0.0064	0.0077*	0.0085**	0.0086**	-0.0278**	-0.0238**
	(0.0047)	(0.0047)	(0.0041)	(0.0041)	(0.0115)	(0.0110)
Absence VA		0.0059		-0.0012		0.0592***
		(0.0036)		(0.0033)		(0.0129)
Suspension VA		0.0047		0.0038		0.0007
		(0.0034)		(0.0034)		(0.0133)
Panel B: t-1 outcome above median						
Test score VA	0.1010***	0.1008***	0.0285***	0.0283***	-0.0267	-0.0158
	(0.0048)	(0.0048)	(0.0035)	(0.0035)	(0.0217)	(0.0197)
Standardized observation score	0.0560***	0.0561***	0.0280***	0.0281***	0.0188	0.0176
	(0.0057)	(0.0057)	(0.0045)	(0.0045)	(0.0244)	(0.0228)
Standardized survey score	-0.0054	-0.0054	0.0037	0.0035	-0.0896***	-0.0769***
	(0.0051)	(0.0051)	(0.0037)	(0.0038)	(0.0222)	(0.0204)
Absence VA		0.0002		-0.0026		0.1110***
		(0.0044)		(0.0031)		(0.0274)
Suspension VA		-0.0012		0.0014		0.0478**
		(0.0040)		(0.0031)		(0.0239)

Notes: The table reports coefficients from the regression of the indicated outcomes on the leave-year-out measures of teacher quality. For the outcome *Absences*, the estimates are based on a pooled sample across math and reading teachers. All columns control for lagged $(t-1)$ test score, absences and indicator for suspensions, student controls, classroom controls, grade-by-year FE and school-by-year FE. Standard errors in the parenthesis are clustered two-ways: at student and teacher level.

* p<0.1, ** p<0.05, *** p<0.01

Table 12: Relationship between performance measures and subsequent year outcomes - By prior achievement and absence

	Math Test Score $t+1$		Reading Test Score $t+1$		Absences $t+1$	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: t-1 outcome below median						
Test score VA	0.0110	0.0097	0.0077	0.0079	-0.0001	0.0009
	(0.0070)	(0.0072)	(0.0057)	(0.0057)	(0.0154)	(0.0151)
Standardized observation score	0.0021	0.0032	0.0145**	0.0146**	-0.0037	-0.0041
	(0.0074)	(0.0075)	(0.0062)	(0.0062)	(0.0170)	(0.0170)
Standardized survey score	0.0068	0.0065	0.0130**	0.0134**	0.0084	0.0088
	(0.0071)	(0.0071)	(0.0061)	(0.0061)	(0.0163)	(0.0164)
Absence VA		-0.0002		0.0015		0.0009
		(0.0056)		(0.0053)		(0.0201)
Suspension VA		-0.0103*		0.0036		0.0092
		(0.0058)		(0.0047)		(0.0251)
Panel B: t-1 outcome above median						
Test score VA	0.0113	0.0064	0.0132***	0.0135***	-0.0744***	-0.0671**
	(0.0076)	(0.0074)	(0.0048)	(0.0049)	(0.0281)	(0.0275)
Standardized observation score	0.0189**	0.0219***	0.0074	0.0074	0.0074	0.0032
	(0.0082)	(0.0080)	(0.0057)	(0.0056)	(0.0310)	(0.0312)
Standardized survey score	0.0016	0.0012	0.0046	0.0049	-0.0033	-0.0031
	(0.0085)	(0.0086)	(0.0049)	(0.0049)	(0.0287)	(0.0288)
Absence VA		-0.0027		0.0064		-0.0393
		(0.0083)		(0.0046)		(0.0331)
Suspension VA		-0.0306***		-0.0035		0.0860**
		(0.0083)		(0.0044)		(0.0382)

Notes: The table reports coefficients from the regression of the indicated outcomes on the leave-year-out measures of teacher quality. For the outcome *Absences* $t+1$, the estimates are based on a pooled sample across math and reading teachers. All columns control for lagged $(t-1)$ test score, absences and indicator for suspensions, student controls, classroom controls, grade-by-year FE and school-by-year FE. Standard errors in the parenthesis are clustered two-ways: at student and teacher level.

* p<0.1, ** p<0.05, *** p<0.01

Appendix

Figure A.1: Teacher categories and break up of evaluation scores

Teacher Category	Supervisor Observations	Student Achievement	Student Surveys
Category A – Most grade 3-12 teachers whose students take standardized tests	**50%**	**35%**	**15%**
Category B – Most grade 1-2 teachers whose students take some assessment	**65%**	**35%**	**-**
Category C – Most grade 3-12 teachers whose students do not take standardized tests	**65%**	**20%**	**15%**
Category D – Teachers whose students neither take standardized tests nor are eligible to complete a student survey	**80%**	**20%**	**-**

Figure A.2: Target distribution for survey scores at Elementary, Middle and High School Level

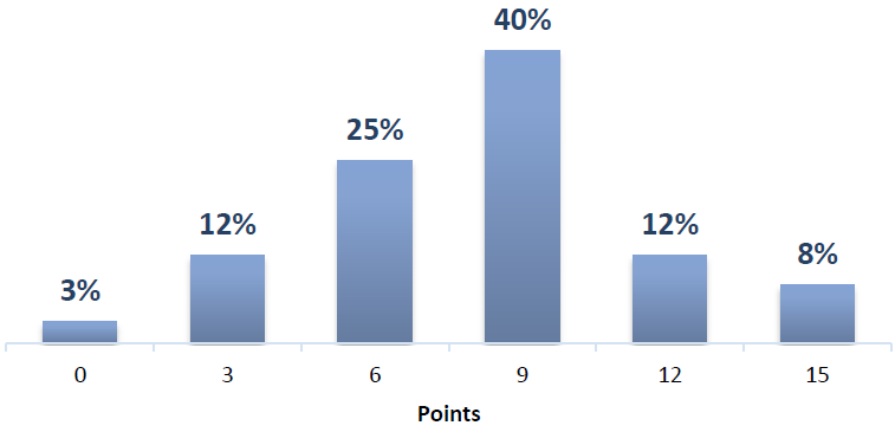


Table A.1: Description of the information included in each survey score component

Component	Information
Class environment	the overall social physical and psychological environment of the classroom
Teacher Expectations	how much the teacher challenges the students to perform their best
Pedagogical Effectiveness	how well the teacher is able to communicate the subject
Student Engagement	behavioral and cognitive investment in the subject and classroom
Supportive Relations	social and academic relation between the student and teacher

Table A.2: Teacher value-added model estimates

	Estimates of teacher SD
Math	0.1094
Reading	0.0430
Absences	0.1065
Suspension	0.0087

Notes: Equation (2) is used to report the estimates.

Table A.3: Testing for bias

	Predicted year t outcomes using year $t-2$ outcomes				Aggregating measures at school-by-grade-by-year level to predict year t outcomes			
	Math	Reading	Absences	Suspensions	Math	Reading	Absences	Suspensions
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test score VA	0.0018***	0.0005	-0.0081	0.0001	0.1090***	0.0467***	-0.0169	0.0001
	(0.0004)	(0.0004)	(0.0063)	(0.0002)	(0.0073)	(0.0053)	(0.0289)	(0.0010)
Standardized observation score	-0.0009*	0.0009**	0.0068	0.0001	0.0405***	0.0268***	0.0751*	0.0004
	(0.0005)	(0.0004)	(0.0075)	(0.0002)	(0.0099)	(0.0072)	(0.0448)	(0.0012)
Standardized survey score	-0.0001	0.00003	-0.0048	0.00004	0.0049	0.0009	-0.1597***	-0.0005
	(0.0004)	(0.00041)	(0.0057)	(0.00016)	(0.0079)	(0.0065)	(0.0424)	(0.0012)
Absence VA	-0.00002	0.00003	0.0014	0.0003	0.0093	0.0054	0.1186***	0.0008
	(0.00031)	(0.00032)	(0.0068)	(0.0002)	(0.0071)	(0.0054)	(0.0338)	(0.0014)
Suspension VA	0.0001	-0.0009***	0.0048	-0.00004	0.0044	0.0027	0.0351	0.0149***
	(0.0003)	(0.0002)	(0.0059)	(0.00016)	(0.0084)	(0.0053)	(0.0340)	(0.0015)
Observations	81,245	90,990	174,773	174,775	121,056	133,706	254,760	254,762
Adjusted R-squared	0.6533	0.6142	0.3653	0.2095	0.6667	0.6328	0.3813	0.1935

Notes: In the first four columns the dependent variable is the predicted year t outcome generated from a regression of the residualized outcomes on twice-lagged outcomes. The reported coefficients are obtained from the regression of the predicted scores on standardized survey scores, controlling for lagged $(t-1)$ test score, absences and indicator for suspensions, student controls, classroom controls, grade-by-year FE and school-by-year FE. The sample is restricted to students with data on both lagged and twice-lagged outcomes. The last four columns regress the respective year t outcomes on the measures that are aggregated at the school-by-grade-by-year level, controlling for lagged $(t-1)$ test score, absences and indicator for suspensions, student controls, classroom controls, grade-by-year FE and school-by-year FE. The estimates for absences and suspensions are based on a pooled sample across math and reading teachers. Standard errors in the parenthesis are clustered two-ways: at student and teacher level.
* p<0.1, ** p<0.05, *** p<0.01

Table A.4: Relationship between performance measures and outcomes - Exluding school-by-year FE in value-added estimation

	Year t				Year $t+1$			
	Math	Reading	Absences	Suspensions	Math	Reading	Absences	Suspensions
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test score VA	0.1165***	0.0409***	-0.0037	-0.0009	0.0165**	0.0146***	-0.0500***	-0.0006
	(0.0053)	(0.0038)	(0.0135)	(0.0009)	(0.0074)	(0.0056)	(0.0171)	(0.0010)
Standardized observation score	0.0412***	0.0301***	0.0069	-0.0008	0.0082	0.0110**	0.0044	0.0022**
	(0.0046)	(0.0036)	(0.0147)	(0.0010)	(0.0068)	(0.0048)	(0.0182)	(0.0011)
Standardized survey score	-0.0029	0.0047	-0.0513***	-0.0007	0.0033	0.0082*	0.0069	-0.0012
	(0.0039)	(0.0031)	(0.0128)	(0.0008)	(0.0067)	(0.0047)	(0.0174)	(0.0010)
Absence VA	0.0032	-0.0033	0.0930***	0.0010	-0.0040	0.0048	-0.0279	-0.0012
	(0.0042)	(0.0035)	(0.0230)	(0.0012)	(0.0070)	(0.0057)	(0.0270)	(0.0014)
Suspension VA	0.0132**	0.0028	-0.0019	0.0096***	-0.0111	-0.0013	0.0459*	-0.0008
	(0.0059)	(0.0035)	(0.0186)	(0.0015)	(0.0091)	(0.0048)	(0.0258)	(0.0016)
Observations	121,056	133,706	254,760	254,762	80,241	86,232	164,826	167,762
Adjusted R-squared	0.6721	0.6315	0.3814	0.1930	0.6060	0.5719	0.2896	0.1207

Notes: The table reports coefficients from the regression of the indicated outcomes on the leave-year-out measures of teacher quality. The leave-year-out value-added estimates do not include school-by-year FE as controls. The estimates for absences and suspensions are based on a pooled sample across math and reading teachers. All columns control for lagged $(t-1)$ test score, absences and indicator for suspensions, student controls, classroom controls, grade-by-year FE and school-by-year FE. Standard errors in the parenthesis are clustered two-ways: at student and teacher level.
* p<0.1, ** p<0.05, *** p<0.01

Table A.5: Relationship between performance measures and outcomes - After residualizing the subjective measures

	Year t				Year $t+1$			
	Math	Reading	Absences	Suspensions	Math	Reading	Absences	Suspensions
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test score VA	0.0979***	0.0359***	-0.0104	-0.0011	0.0120*	0.0118**	-0.0322*	-0.00001
	(0.0041)	(0.0034)	(0.0121)	(0.0008)	(0.0065)	(0.0048)	(0.0167)	(0.00095)
Standardized observation score	0.0341***	0.0252***	0.0081	-0.0011	0.0086	0.0096**	0.0116	0.0020**
	(0.0038)	(0.0030)	(0.0124)	(0.0008)	(0.0061)	(0.0041)	(0.0161)	(0.0009)
Standardized survey score	-0.0016	0.0033	-0.0502***	-0.0005	-0.0003	0.0060	-0.0071	-0.0014
	(0.0035)	(0.0027)	(0.0113)	(0.0007)	(0.0064)	(0.0041)	(0.0158)	(0.0009)
Absence VA	0.0030	-0.0024	0.0854***	0.0006	-0.0045	0.0023	-0.0220	-0.0004
	(0.0033)	(0.0026)	(0.0173)	(0.0008)	(0.0059)	(0.0044)	(0.0214)	(0.0011)
Suspension VA	0.0017	0.0021	0.0212	0.0087***	-0.0238***	-0.0017	0.0498**	-0.0001
	(0.0031)	(0.0025)	(0.0139)	(0.0011)	(0.0065)	(0.0035)	(0.0241)	(0.0013)
Observations	121,056	133,706	254,760	254,762	80,241	86,232	164,826	167,762
Adjusted R-squared	0.6717	0.6340	0.3815	0.1933	0.6061	0.5719	0.2896	0.1207

Notes: The table reports coefficients from the regression of the indicated outcomes on the leave-year-out measures of teacher quality. The observation and survey scores are residualized by classroom averages of lagged test scores, absences, suspensions and student controls; grade-by-year and school-by-year fixed effects. The estimates for absences and suspensions are based on a pooled sample across math and reading teachers. All columns control for lagged $(t-1)$ test score, absences and indicator for suspensions, student controls, classroom controls, grade-by-year FE and school-by-year FE. Standard errors in the parenthesis are clustered two-ways: at student and teacher level.

* p<0.1, ** p<0.05, *** p<0.01