Should Value-Added Models Weight All Students Equally?

Case Tatro[*]

October 23, 2025

[Click here for the latest version.]

Abstract

Conventional value-added (VA) models estimate teacher quality as a simple average of the difference between students' actual and predicted standardized test scores. These models therefore implicitly assume it is just as important to raise test scores of lower-achieving students as it is to raise test scores of higher-achieving students. I consider whether a weighted average of residuals might be more useful. Using data from North Carolina, I find that teacher VA measures become more predictive of teachers' long-run impacts when the highest-achieving students are weighted more than the median student. Strikingly, even impacts on *low-achieving* students' long-run outcomes are best predicted by increasing the weight on impacts on *high-achieving* students' short-run outcomes. These differences in weights may reflect that either (i) in small samples some students are more informative about a teacher's overall VA than others or (ii) test-score effects for different students capture different aspects of teaching, and some of these aspects are more informative than others. I find empirical evidence supporting both explanations. In particular, the large weights for high-achieving students are partially but not completely explained by the fact that their residuals are less noisy.

JEL Classification Codes: I21, I28

[*] Binghamton University (SUNY Binghamton) Email: ctatro1@binghamton.edu. I am grateful to my advisor David Slichter for his comments and feedback on this paper. Thanks to the North Carolina Education Research Data Center for allowing me to use the data for this project. I am grateful to the Binghamton University Economics Department for sponsoring the fees associated with accessing the data. I am also grateful to the Binghamton University Applied Labor Group for helpful comments and feedback. The views expressed are those of the author and do not necessarily reflect the positions or policies of the North Carolina Department of Public Instruction.

1 Introduction

School systems commonly evaluate teachers using statistical models designed to estimate their impact on student outcomes. These are called value-added (VA) models. Typical construction of VA measures follows a three step process. First, the researcher uses a regression model to predict students' outcomes using their predetermined characteristics, the most important of which is a student's lagged outcome. In step 2, the researcher calculates an average of student residuals (i.e., the gap between actual and predicted outcome) within a class. This is referred to as an unadjusted VA measure. The final step is a shrinkage correction, in which the unadjusted VA is "shrunk" or adjusted towards an average to account for statistical noise in the residuals using an Empirical Bayes framework. These VA measures are reported in units of standard deviations (SD) of student achievement. For instance, a test-score VA of 0.1 SD indicates a teacher is estimated to increase students' test scores on average above predicted levels by 0.1.

Value-added models are popular with policymakers and researchers because they have two main statistical properties which make them useful for evaluating teacher quality. First, there is a large body of evidence that VA models are approximately forecast-unbiased (e.g., Kane and Staiger, 2008; Konstantopoulos, 2009; Hanushek and Rivkin, 2010; Bacher-Hicks, Kane, and Staiger, 2014; Chetty, Friedman, and Rockoff, 2014a; Koedel, Mihaly, and Rockoff, 2015). This means that VA measures are fair in the sense that they do not systematically reward teachers for having better or worse students.

Second, teachers who are estimated to raise test scores also tend to promote student success on long-run measures such as educational attainment, wages as an adult, and teen pregnancy (e.g., Chetty, Friedman, and Rockoff, 2014b; Gilraine and Pope, 2021; Backes et al., 2023; Petek and Pope, 2023; Lavy and Megalokonomou, 2024). If the ideal criteria for evaluating teachers is their ability to raise test scores, conventional test-score VA models by construction estimate this aspect of teacher quality. If, on the other hand, the ideal criteria is based on a teacher's impact on longer-run outcomes, test-score VA proxies for these impacts as well. It is not practical to wait to evaluate a teacher until her students' long-run outcomes are realized. Test scores and therefore a teacher's impacts on student test scores are available immediately.

It is important to note that while conventional VA models proxy for a teacher's ability to affect long-run outcomes, there is evidence that this proxy only captures a minority of a teacher's long-run impacts (e.g., Deming, 2009; Chetty et al., 2011; Chamberlain, 2013). Adding VA measures using non-cognitive outcomes (such as absences and disciplinary behavior) to predict teachers' long-run impact closes some of this gap (e.g., Jackson, 2012;

Chetty, Friedman, and Rockoff, 2014b; Blazar and Kraft, 2017; Jackson, 2018; Mulhern and Opper, 2023). These results imply that (i) a teacher's VA is multidimensional (teachers impact multiple outcomes simultaneously) and (ii) conventional test-score VA measures may miss or leave out important information about teacher quality.

Teachers appear to differ in their impacts of different types of students (e.g., Lavy, Paserman, and Schlosser, 2012; Condie, Lefgren, and Sims, 2014; Fox, 2016; Delgado, 2020; Aucejo et al., 2022; Gershenson et al., 2022; Graham et al., 2023). The literature also finds that teachers differ in their ability to increase test scores for students with a higher versus lower baseline level of achievement (e.g., Biasi, Fu, and Stromme, 2021; Eastmond et al., 2024). Conventional VA measures focus on average impacts across all students, and do not attempt to exploit information that may be contained in a teacher's heterogeneous impact across students.

Test-score VA models equate raising test scores for the lowest-achieving students with raising test scores by the same amount for the highest-achieving students. It is not obvious that a one-point increase in test scores is equally valuable or as important no matter what the baseline achievement is. Nielsen (2019) finds that getting easier questions right on the Armed Forces Qualification Test (AFQT) is more predictive of students' long-run outcomes than getting harder questions correct. These results suggest there may be room for improvement in using test-score VA measures to evaluate teachers. In particular, weighting students equally in a conventional VA model might be unwise if a teacher has different impacts on different groups of students.

As an example, consider a policymaker whose goal is to select teachers with the highest long-term impact, where the long-run outcome of interest is high school graduation. The policymaker could use a conventional test-score VA measure to proxy for a teacher's impact on high school graduation. But not all students are equally at risk of not graduating high school. Students with high lagged test scores are likely to graduate high school no matter how good or bad a given teacher is. Under such a scenario, a more informative short-run measure of teacher quality might be a VA measure that gives a higher weight to a teacher's impact on lower-achieving students since those are the students most at risk.

In this paper, I ask the following research questions: suppose I construct a new test-score value-added model as a weighted (rather than unweighted) average of student residuals. If I choose the weights with the objective of maximizing the predictive power of my new VA measure on a teacher's long-run impact, such as high school graduation VA, how do these estimated weights compare to a conventional VA model in which all students are weighted equally? Furthermore, how much better is this weighted VA at predicting teachers' long-run impacts?

In this paper, I use data on North Carolina students and teachers in grades 3-5 from the North Carolina Education Research Data Center (NCERDC). I first document that teachers have different underlying impacts on math and reading scores for higher and lower achieving students. I classify students into 5 bins, then estimate weights on these bins to maximize the predictive power of my weighted VA measure to predict teachers' high school graduation VA. I compare the predictive power of this new VA measure on a teacher's long-run impact to the predictive power of a conventional VA model and investigate the potential mechanisms driving differences in weights for different bins.

I find that the estimated weights for my new VA measure are not equal across all students, rejecting that a conventional VA model is the most predictive measure of teacher's impact on longer-run outcomes. Specifically I find that, for both math and reading, the lowest-achieving students should receive lower weight than the median student, and the highest achieving students should receive the highest weight.

There are three possible explanations for this result. Explanation 1 is that these weights reflect differences in the noisiness of residuals. The students with the highest signal to noise ratio receive the highest weights. Explanation 2 is that a teacher's impact on test scores in particular bins may be more informative about a teacher's impacts on other bins. This is an explanation based on efficient use of a finite sample rather than a true-effects explanation. A teacher may differ in their true VA for students in different bins, but the bins that receive the highest weight are bins that reveal the most information about a teacher's impact not only for students in that particular bin but also about a teacher's impact for students in other bins. Finally, explanation 3 is that these weights represent true effects. Bin-specific value-added measures likely reflect a mix of factors such as pacing, clarity of instruction, classroom management, and inspiring curiosity. If some aspects are more important than others for promoting high school graduation, than the the students in bins which reflect the most important factors receive the highest weight.

I find that the observed differences in weights for lower-achieving students most likely reflect an efficient use of a finite sample (explanations 1 and 2) than true differences in teacher quality (explanation 3). For higher-achieving students, a small-sample efficiency story cannot fully explain the large weights relative to the median students. My weighted VA does a marginally better job of predicting a teacher's high school graduation VA compared to a conventional VA model. A teacher's impacts on non-cognitive skills are more important than a teacher's test-score impacts for predicting high school graduation VA regardless of whether test-score VA is weighted or unweighted.

I make three contributions to the value-added literature. First, I derive optimal weights for use as an alternative to a conventional VA measure. Second, my results are suggestive that a high test-score VA for the highest-achieving students might reflect teachers with underlying characteristics which are especially important for promoting students' long-term outcomes. Third, I provide additional evidence (and confirm the results from Biasi, Fu, and Stromme, 2021 and Eastmond et al., 2024) that teacher's VA for math and reading scores differs across students of different baseline achievement.

The rest of the paper proceeds as follows. In Section 2, I describe the North Carolina data I use in my analyses and provide student-level summary statistics. I provide additional details regarding conventional VA models in Section 3. I also document summary statistics regarding my estimated teacher VA and evidence that teachers differ in their VA for lower vs. higher achieving students. I then discuss the details of my procedure for estimating this new VA measure and my results in Section 4. I then conclude in Section 5.

2 Data

My data comes from the North Carolina Education Research Center (NCERDC) and contains information regarding all North Carolina public school students. I restrict the data to students in grades 3-5 from 1997 to 2011. The data include standardized test scores in math and reading, demographic information, and an identifier of the teacher who administered the math and reading test. The data also include information on attendance (available starting in 2006), disciplinary behavior (available starting in 2001) (such as number of days suspended within a given school year), whether each student dropped out in a given year, withdrew, and whether or not each student graduated high school (available starting in 2002).

I further restrict the North Carolina data to students for whom I have a non-missing student identifier and for whom I can map to a particular teacher. I also limit my data to students in reasonably-sized classrooms. I define reasonably-sized classes as those in which the number of students is between 10 and 35.

I construct standardized measures of cognitive and non-cognitive outcomes for use in my analyses. I normalize test scores in math and reading to be mean zero with variance 1 for each testing year and grade. I re-define absences as the negative natural log of days a student is absent plus 1 (to avoid undefined values) following Mulhern and Opper (2023). I replace missing values for absences (as well as days suspended) with values of 0s for years in which absence or disciplinary data is available. I also create a behavioral index for students based on a principal component analysis of suspensions and other disciplinary infractions such that worse behavior receives a lower (more negative) value, which I describe further in Appendix B.

My restricted sample contains approximately 2.5 million student-year observations 1998 to 2011 for students in fourth and fifth grade.[1] My sample contains slightly more males (51%) than females (49%). Most students (74%) are defined as economically disadvantaged, and 80% of my sample graduated high school. The average logged-absences is approximately 8. The average class size in my restricted sample is 23 students. I report summary statistics in Table 1.

3 Conventional Value-Added Models

I begin my discussion of conventional VA models by providing additional details regarding the four-step construction process. The first step is to use a regression model to predict a student's outcome using their predetermined characteristics. The most important characteristic is a student's lagged outcome. Test-score VA models use a student's standardized test scores in math and reading as the outcome, while other VA models (such as non-cognitive VA) use student outcomes such as attendance, disciplinary behavior, high-school graduation, and other longer-term outcomes. Regardless, the construction of VA models is the same for all outcomes. The typical way of estimating these residuals is to estimate models of the form

$$\tilde{Y}_{i,j,s,t} = \alpha + \gamma \tilde{Y}_{i,s,t-1} + \tilde{\mathbf{X}}_i \beta + \epsilon_{i,j,s,t} \tag{1}$$

where $Y_{i,j,s,t}$ represents outcome s for student i in teacher j's class in year t. $Y_{i,j,s,t-1}$ is a function of lagged outcomes, and $\mathbf{X}_i$ represents a vector of student-level demographics.[2] Tilde ($\sim$) indicates a variable demeaned at the classroom level.[3]

Step 2 of constructing a value-added model is to construct a student residual as the difference between a student's actual and predicted outcome. The predicted outcome is calculated using the coefficients from the regression model from step 1. Each student's residual $\epsilon_{i,j,s,t}$ is calculated as

$$Y_{i,j,s,t} - \hat{\alpha} - \hat{\gamma} Y_{i,s,t-1} - \mathbf{X}\hat{\beta}, \tag{2}$$

[1] I exclude 1997 due to the lack of lagged outcomes available. These lagged outcomes would come from the 1996 data, which is not available. Students do not take a standardized test in either math or reading in grade 2. Therefore the earliest grade in which students have a lagged standardized test score for math and reading is fourth grade.

[2] For test-score VA and high school graduation VA measures, I use a cubic of lagged test scores for both math and reading as well as whether a student is economically disadvantaged, gender, disability status, and english learner status. To estimate non-cognitive VA measures, I use a cubic of the lagged non-cognitive outcome in addition to whether a student is economically disadvantaged, gender, disability status, and english learner status.

[3] Demeaning within classroom is econometrically equivalent to including teacher fixed effects (as noted in Mulhern and Opper, 2023) and necessary given the number of teachers I have in my sample.

where each student's characteristics are multiplied by the corresponding coefficient estimated using the demeaned regression from 1. These residuals are recentered to ensure $\epsilon_{i,j,s,t}$ is mean 0 by subtracting the mean residual across all students and years. I define the recentered residuals as

$$r_{i,j,s,t} = \epsilon_{i,j,s,t} - \bar{\epsilon}_{s,t}, \tag{3}$$

where $\bar{\epsilon}_{s,t}$ represents the average residual across all students and years.

In step 3, a teacher's unadjusted VA is calculated as the average of student residuals within a classroom. This unadjusted VA measure, defined as

$$V\hat{A}_{j,s,t} = \sum_{i \in j,t} r_{i,j,s,t}, \tag{4}$$

contains a mixture of teacher j's true VA and statistical noise. In step 4, this unadjusted VA is adjusted, or "shrunk", (commonly using a Bayesian framework) to account for sampling error. [4].

Value-added models are a popular way to evaluate teacher quality and have desirable statistical properties. In 2023, 30 states used VA measures as part of teacher evaluations, down from 43 states in 2015 (National Council on Teacher Quality, 2024). Perhaps the most desirable statistical property of VA models is that they are forecast unbiased. The average true VA of teachers with the same estimated VA is in fact, their estimated VA. This allows VA measures to have a causal interpretation. The literature has confirmed this property using quasi-experiments involving teachers who switch schools, comparing student outcomes before and after the switch to show that estimated teacher value-added predicts future student achievement without systematic bias (e.g. Chetty, Friedman, and Rockoff, 2014a; Rivkin, Hanushek, and Kain, 2005. This result also holds using random assignment of teachers to students (e.g. Kane and Staiger, 2008; Bacher-Hicks, Kane, and Staiger, 2014).

Another desirable property of VA models is that test-score VA measures in particular have been found to be predictive of a teacher's impacts on long-run outcomes. This enables test-score VA measures to act as a short-run proxy for teacher quality on long-run impacts. (Chetty, Friedman, and Rockoff, 2014b) find that teachers with higher test-score VA are also teachers who improve students' long-term outcomes such as earnings, college attendance, and reduce teen pregnancy among their students.

Test-score VA measures, however, do not perfectly predict a teacher's impacts on these long-run outcomes. For example, Chetty et al. (2011) analyze data from Project STAR and finds that the actual variation in teacher impacts on long-run outcomes is larger than the

[4] I adjust my VA estimates using the methodology from Mulhern and Opper, 2023. I include summary statistics of my VA estimates in Appendix A

variation implied by test-score VA measures. Deming, 2009 uses NLSY data to show that teacher's impacts on test-scores fade-out over time but a teacher's impacts on longer-run outcomes persist. Further, Chamberlain (2013) finds much smaller impacts on a teacher's impact on adult earnings than the impact implied by the same teacher's impact on test scores.

The literature finds that including a teacher's impact on other non-cognitive outcomes (such as absences and disciplinary record) can account for why test-score measures do not perfectly predict a teacher's long-run impact. For example, Jackson (2012) finds that including non-cognitive VA in addition to test-score VA increases the percentage of explained variation of teachers' impacts on high school graduation. Mulhern and Opper (2023) account for a teacher's multidimensional impact on students by estimating test-score VA for a given teacher as a function of a teacher's estimated impact on both cognitive and non-cognitive outcomes. Chetty, Friedman, and Rockoff (2014b) find that accounting for a teacher's non-cognitive VA also increases the predictability of a teacher's impact on adult wages and teen pregnancy.

Conventional test-score VA models may therefore miss important information regarding teacher quality. Teachers have different test-score impacts on different types of students. In particular teachers differ in their test-score VA for higher versus lower achieving students. Eastmond et al. (2024), for example, finds that San Diego teachers have different VA for students above versus below the median lagged test-score for math and reading. Biasi, Fu, and Stromme, 2021 uses Wisconsin data to show that teachers have a comparative advantage in either lower- or higher-achieving students.

3.1 Documenting Differences in Teacher VA for High Versus Low Achieving Students

I now describe my methodology for documenting that teachers have different underlying, or true, test-score VA for higher- versus lower-achieving students. My objective is to estimate the correlation between a teacher's true VA for these two groups of students. I first discuss how I define higher- and lower-achieving students. Then I describe the log-likelihood function I combine with a maximum likelihood estimator to estimate the correlation between a teacher's underlying VA for higher-achieving students and a teacher's underlying VA for lower-achieving students. I discuss my assumptions and then present results below in Table 2.

I divide students within a class into top and bottom students based on their lagged test-score in a given subject. I first calculate the distribution of lagged math and reading scores for all students within a given school and grade. I then classify students into "top-" or "bottom-" performing students based on how their lagged achievement compares to this broader distribution. I use three different definitions of top vs. bottom students for each subject. First, I define a "Top/Bottom 50" split in which I classify a top student as having a lagged test score above the median, and a bottom student as having a lagged test score below the median. Second, I define a "Top/Bottom 30" split in which top students have a lagged achievement in the top 30th percentile, and bottom students have a lagged achievement in the bottom 30th percentile. Third, I define a "Top/bottom 25" split using the top and bottom 25th percentile to classify top and bottom students.

For each definition of top- and bottom-performing students, I estimate a teacher's VA for math and reading separately for top and bottom students according to equations 1 - 4. For example, I estimate a teacher's unadjusted math VA for top students using the Top/Bottom 50 split by estimating equations 1 - 4 restricting my sample to only students whose lagged math score is above the school-grade-level median lagged math score in a particular year. I estimate a standard error for this unadjusted VA measure calculated as the standard deviation of student residuals within a class and "group" (top versus bottom) divided by the square root of the number of students within that class and group. As an example, if there are 16 top students and the standard deviation of these 16 students' math residuals is 1, I calculate the standard error of a teacher's top unadjusted VA as 0.25.

I assume that the variance of each of these estimated VA measures is the variance of a teacher's true VA plus noise. I also assume that both the estimated VA and the noise in these estimated values are normally distributed. These assumptions imply that I can use a maximum-likelihood estimator (MLE) to the estimate the correlation between teachers' true VA for top students and teacher's true VA for bottom students by minimizing

$$v_{j,s,t} \cdot \Sigma_{j,s,t} \cdot v_{j,s,t}, \tag{5}$$

where $v_{j,s,t}$ is a vector of a teacher j's estimated VA in year t for subject s. This matrix $v_{j,s,t}$ is defined as

$$v = \begin{bmatrix} \hat{VA}^{top}_{j,s,t} \\ \hat{VA}^{bot}_{j,s,t} \end{bmatrix},$$

where $\hat{VA}^{p}_{j,s,t}$ represents the unadjusted VA for teacher j in subject s in year t for either top-performing students ($\hat{VA}^{top}_{j,s,t}$) or bottom-performing students ($\hat{VA}^{bot}_{j,s,t}$).

I define $\Sigma_{j,s,t}$ as the variance-covariance matrix between a teacher's true VA measures for top and bottom students, denoted as

$$\Sigma = \begin{bmatrix} \sigma^2_{VA_{top}} & \sigma_{VA_{top}VA_{bot}} \\ \sigma_{VA_{top}VA_{bot}} & \sigma^2_{VA_{bot}} \end{bmatrix}.$$

I present the results of these MLE estimates below in Table 2. Columns 1 and 2 represent the estimates from a baseline specification. In columns 3 and 4 I control for the average classroom achievement and variance in classroom achievement in order to account for differences in signal to noise ratio of student residuals due to class size. For both specifications the estimated correlation in true VA measures for top and bottom students is less than 1. This correlation becomes weaker as the definition of top versus bottom students becomes stricter. These results provide an empirical motivation that there may be something to be gained when allowing different weights on different students within a VA model. Or, put slightly differently, it is unlikely when I allow for different weights on different students in a weighted average VA model those weights will be imply all students should be weighted equally.

4 Main Analysis: Estimating Weighted-Average Value-Added

In this section I first introduce a toy example in order to further build intuition as to why I expect my weighted average VA measure to weight students differently. I then discuss my econometric strategy for estimating my weighted VA measure and present results. From these results I posit three possible explanations and seek to understand to what extent each explanation could be driving my results.

4.1 Toy Example Estimation and Results

I construct a toy example in which high school graduation serves as the long-run outcome of interest. Suppose that a student's latent, or underlying, propensity to graduate high school follows a normal distribution according to the model

$$PR(Graduated_{i,j,t}|\tilde{Y}_{i,s,t-1}) = \Phi(\delta_0 + \delta_1\tilde{Y}_{i,s,t-1} + \delta_2\tilde{Y}^2_{i,s,t-1} + \delta_3\tilde{Y}^3_{i,s,t-1} + \tilde{X}_i\beta + \omega_{i,s,t}), \quad (6)$$

where $\tilde{Y}$ represents the lagged test score of student i from year $t-1$ in subject s. The vector of covariates $\mathbf{X}$ contains the same covariates as included in Equation 1.

From this data generating process, there are some students who are more at risk of not graduating high school than other students. Suppose that 80% of students graduate high school, as is the case in my North Carolina sample. The students who are more at risk of not graduating high school are more likely to be lower-achieving students. In a value-added setting, if the policy maker's goal is to maximize high school graduation, the optimal weighting scheme might place a higher weight on students more at risk of not graduating than on students who are high achieving. Higher-achieving students are more likely to graduate high school irrespective of having a teacher who has a high impact on high school graduation, and may receive a lower weight according to this framework.

For test-score effects to be p

I derive a set of empirical weights according to this framework by directly estimating the relationship between lagged achievement and a student's marginal propensity to graduate high school. First, I estimate Equation 6 in order to obtain each student's propensity to graduate. I then determine the slope of the normal distribution at each student's estimated latent propensity to graduate. This slope gives an approximation of how "at risk" a particular student is of not graduating high school. A steeper slope indicates a slightly better teacher might make the difference between a student graduating high school or not. Student's with the highest marginal propensity to graduate receive the highest weight. I normalize the weights within each classroom such that the sum of weights within a classroom sum to 1.

I estimate marginal propensity to graduate, and therefore student weights, separately using lagged math and reading scores. I divide students into vintiles (20 bins) of lagged achievement in each subject and report the average etimated weight within each bin. I then re-scale the weights such that the weight on the middle bin (bin 10) is equal to 1. Weights higher than 1 therefore represent students who are more at-risk, and therefore should receive a higher weight, than the median student. Weights lower than 1 similarly represent students less at-risk and students who should receive a lower weight than the median student. I present these results below as Figure 1.

I make the following assumptions to ensure a teacher's test-score effects are proportional to how at-risk a student is of not graduating high school. First, I assume a teacher's value-added (VA) in each bin is independent of a teacher's VA in every other bin. Second, that a teacher's VA for a student's particular bin is the only factor that affects that particular student's probability of graduating high school.

These toy example weights illustrate that, in a scenario in which students' risk of not graduating are correlated with lagged achievement, a weighted VA may increase the explanatory power of a teacher's long-run impact compared to an unweighted, conventional VA measure. The weights for both math and reading in Figure 1 imply that students with the lowest lagged achievement, those at the highest risk of not graduating, receive the highest weight. These weights decrease in a non-linear fashion as lagged achievement increases. The weights are lowest for students with the highest lagged-achievement. The weights imply that highest-achieving students should be weighted about 10% as much as the median student, which suggests that even the best students are slightly at risk of graduating high school. I now turn to the data to empirically decide how different students should be weighted within a VA model.

4.2 Weighted Average VA: Econometrics

I now discuss how I construct my alternative, weighted measure of value-added for teacher j in year t for subject s, which I will denote as $VA^*_{j,s,t}$. I construct VA^* with two goals in mind. First, VA^* should be a weighted average of student residuals. This ensures VA^* is comparable to existing methodologies for estimating value-added. Specifically, the conventional VA measure (a simple average of student residuals) is a special case of my weighted value-added measure. Second, VA^* should maximize the predictive power of a teacher's high school graduation VA.

I first divide a classroom into 5 bins, or quintiles, based on the distribution of lagged test scores in subject s within a given school and grade. I then assign a weight to each bin, which I denote β_k. I normalize the weights for each bin such that the weight on the middle bin (bin 3) is equal to 1. This normalization results in an intuitive interpretation of the estimated bin weights. Bin weights larger than 1 indicate bins of students who should be weighted more heavily than the median group of students. Bin weights smaller than 1 indicate bins of students should be weighted less heavily than the median group of students. I also consider the number of students in each bin when calculating bin weights.[5] This ensures that students in bins with fewer/more students do not mechanically receive a higher/lower weight simply as a function of the number of students in a particular bin. This also ensures the sum of weights within a classroom averages to 1 across classrooms.

I therefore construct VA^* as

$$VA^*_{j,s,t} = \sum_i \Big(\frac{\beta_k \mathbf{1}\{i \in k\}\mathbf{1}\{i \in j\} r_{i,j,k,s,t}}{\sum_i \sum_k \beta_k \mathbf{1}\{i \in k\}\mathbf{1}\{i \in j\}} \Big), \tag{7}$$

where $r_{i,j,s,t}$ represents student i's residual in subject s in year t, and student i is in teacher j's class. Note that an alternative and equivalent formulation of VA^* is given by

$$VA^*_{j,s,t} = \frac{1}{\sum_i \sum_k \beta_k \mathbf{1}\{i \in k\}\mathbf{1}\{i \in j\}} \sum_i \beta_k \mathbf{1}\{i \in k\}\mathbf{1}\{i \in j\} r_{i,j,k,s,t}, \tag{8}$$

where $\sum_i \sum_k \beta_k \mathbf{1}\{i \in k\}\mathbf{1}\{i \in j\}$ represents the sum of the bin weight times the number of students in each bin, summed over all bins within class j.

A conventional VA model is a special case of my weighted average measure. A simple average of student residuals is equivalent to setting $\beta_k = 1$ for all bins $k \in K$. Under such a

[5] For classes with zero students in a particular bin, I set the sum of residuals and the number of students in these bins to 0 rather than dropping such a class from my analysis.

scenario my value-added estimator becomes

$$VA^*_{j,s,t} = \frac{1}{N_{j,s,t}} \sum_i r_{i,j,k,s,t}, \tag{9}$$

where $\sum_i \sum_k \beta_k \mathbf{1}\{i \in k\}\mathbf{1}\{i \in j\}$ becomes the number of students in class j, or $N_{j,s,t}$ and $\beta_k \mathbf{1}\{i \in k\}\mathbf{1}\{i \in j\}$ collapses to 1 for all i, leaving $\sum_i r_{i,j,s,t}$. I allow for such a set of weights to occur if weighting all students equally is empirically the most predictive of a teacher's high school graduation VA.

It is important to distinguish between the estimated weight on a particular bin and the implied weight on each student's residual within that particular bin. β_k represents the estimated weight on a particular bin k. $\frac{\beta_k}{\sum_i \sum_k \beta_k \mathbf{1}\{i \in k\}}$ is the weight on each student residual within particular bin k. As an intuitive example, consider a set of estimated weights using 3 bins. Suppose there are 20 students in the class, with 6 students in bin 1, 8 students in bin 2, and 6 students in bin 3. Also suppose the estimated bin-weights for each bin are $\beta_1 = 2$, $\beta_2 = 1$, and $\beta_3 = 0.5$ The denominator of $VA^* = \sum_i \sum_k \beta_k \mathbf{1}\{i \in k\}\mathbf{1}\{i \in j\}$ is equal to $2 \cdot 6 + 1 \cdot 8 + 0.5 \cdot 6 = 23$. The estimated weight on each student's residuals in bin k is $\frac{\beta_k}{23}$, or $\frac{2}{23}$ on each student's residual in bin 1, $\frac{1}{23}$ on each student's residual in bin 2, and $\frac{0.5}{23}$ on each student's residual in bin 3. These student-level weights sum to 1 within the classroom.

I choose the weights (β_k's) to maximize the predictive power of this new VA measure on a teacher's long-run impacts. The longest-run outcome I observe in my data is a student's high school graduation. Therefore I seek to maximize the predictive power of this new VA measure on a teacher's high school graduation VA. I estimate the weights to minimize the sum of squared errors in the predicted high school graduation VA of teacher j in year t using the residuals of teacher j's students. I obtain these residuals using step 1 of the conventional VA methodology as described in Equation 1.

Of course, using the same students to calculate both the weights for this new VA measure and a teacher's high school graduation VA would introduce mechanical bias in my estimates (as noted in Jackson, 2018). Therefore I combine Jackson's out-of-sample methodology with the methodology for calculating shrunken VA measure using multiple years for a given teacher described in Mulhern and Opper (2023). I estimate a teacher's pooled high school graduation VA by estimating Equation 1 through Equation 4 with high school graduation as the outcome variable. I then to calculate a teacher j's adjusted high school graduation VA using all years other than year t, which I denote as $\tilde{VA}^{grad}_{j,-t}$.

I then use non-linear least squares to minimize the squared prediction error of a teacher's out-of-sample high school graduation VA. I define my optimization problem as

$$\min_{(\beta_k)s} \left[\tilde{VA}^{grad}_{j,-t} - \beta_0 - VA^*_{j,s,t}\right]^2, \tag{10}$$

where $VA^*_{j,s,t}$ represents the weighted VA measure for subject s using the Equation 1 residuals for students taught by teacher j in year t. I choose an intercept β_0 and weights on each bin 1 through 5 with weight on bin 3 normalized to 1.

4.3 Weighted Average VA: Results

I present my initial estimated weights for math and reading below as Figure 2. My estimates reject that a conventional VA model is most predictive of a teacher's long-run impact using either math or reading test scores. In particular for both subjects the higher-achieving students (bins 4 and 5) receive a higher weight than the median student. For reading, lower-achieving students receive a lower weight. The estimated weight on the lowest-achieving students based on lagged reading scores is approximately two-thirds as large as the weight on the median students.

There are three potential explanations for the results in Figure 2. The first explanation is that the noisiness in student residuals differ for higher- versus lower-achieving students. In particular the literature has found that lower-achieving students often have more noisy test score residuals than higher-achieving students (e.g. Kane and Staiger, 2008; Koedel, Mihaly, and Rockoff, 2015.The lower weights on the lowest-achieving students for reading, for example, might be due to relatively high variance of residuals for these students relative to the variance of residuals in the middle bin.

The second explanation is that a teacher's VA for certain bins may be more predictive about a teacher's impact on other bins. This is a small sample efficiency explanation rather than a true effects interpretation. It could be that teachers do truly differ in their effects across different bins, but the bins that receive the highest weight receive the most information about a teacher's impacts on the largest number of students, both within and outside of a particular bin.

The third possible explanation is a "true effect" interpretation. A teacher's VA measure likely captures a combination of many aspects of teaching, such as pacing, classroom management, inspiring critical thinking and creativity, and clarity of instruction. Perhaps being able to increase test scores for particular bins weights particular aspects of teaching more heavily. The bins that capture the most generally useful aspects of teaching receive the highest weight.

4.4 Weighted VA: Result Drivers

I now seek to understand to what extent each of the three possible explanations might explain my estimated weights shown in Figure 2. I begin by asking to what extent relative differences in the noisiness of student residuals between bins might explain my estimated weights. A straightforward way to assess this is to estimate a set of weights entirely determined by the relative variance of student residuals within each bin. I estimate this variance-only set of weights for math and reading as follows. First, I estimate the variance of student residuals in a particular subject across all classrooms and across all years. I then normalize these variances such that the variance of student residuals in the middle bin is 1. The variance of student residuals in the other bins therefore becomes the relative variance of student residuals compared to the other bin. I then calculate the weight for each bin as the inverse of the relative variance of student residuals. For example, suppose the variance of student residuals for reading scores in bin 1 (the lowest bin) is twice the variance of student residuals for reading scores in bin 3. I define the variance-only reading weight for bin 1 as 0.5.

I report these variance-only weights as gray squares (alongside the estimated weights from Figure 2 as blue dots with red standard error bars) below as Figure 3. These variance-only weights confirm that, for both subjects, residuals for lower-achieving students are noisier than for higher-achieving students. This relative noisiness explains some but not all of my results. In particular, this explanation seems to explain the lower weight on lower-achieving students, but does not fully account for the high weights on higher-achieving students. This relative noise story does a relatively better job of explaining the weights for the middle bins (bins 2 and 4) for both subjects.

Perhaps these variance-only weights do not fully explain my results because while I estimate variances across all years and classrooms, Equation 10 estimates average weights using classroom-year observations where the number of students in any given classroom and year is by definition limited to between 10 and 35 students. If I had larger classrooms, perhaps the weights I would obtain by estimating Equation 10 would more closely mirror the variance-only weights I show above in Figure 3.

To test this, I artificially increase the number of students defined as being in the same classroom by estimating weights using what I denote as a "reverse out-of-sample" methodology. Instead of using a teacher's out of sample high school graduation VA, I use a teacher's estimated high school graduation VA using only students in year t. Instead of calculating VA^* using students in year t, I treat the data as if all students taught by teacher j not in year t were in one large class taught by teacher j in year t. I then estimate

$$\min_{(\alpha_k)s} \left[VA^{grad}_{j,t} - \alpha_0 - VA^*_{j,s,-t} \right]^2, \tag{11}$$

where $VA^{grad}_{j,t}$ teacher j's high school graduation VA calculated only for students in year t.

$VA^*_{j,s,-t}$ is defined as

$$VA^*_{j,s,t} = \frac{1}{\sum_i \sum_k \beta_k \mathbf{1}\{i \in k\}\mathbf{1}\{i \in j\}} \sum_i \beta_k \mathbf{1}\{i \in k\}\mathbf{1}\{i \in j\} r_{i,j,k,s,-t}, \tag{12}$$

or the weighted sum of residuals for all students who had teacher j in any year other than year t, which I denote $r_{i,j,k,s,-t}$.

I report estimated weights for this reverse out-of-sample approach below as Figure 4. I include the initial weights from Figure 2 as hollow gray squares for comparison. The bin weights for the first two bins are qualitatively similar for both math and reading. For math (Figure 4a), there is no statistical difference between the reverse out-of-sample weight and the initial weight for bin 4. For math and reading, the weight on the highest-achieving students (bin 5) is higher than the corresponding weight from Figure 2.

I conclude from these alternative sets of weights that the pattern of weights I observe in Figure 2 cannot be fully explained by differences in the noisiness of residuals for lower-achieving versus higher-achieving students. In particular, the weights implied by such a difference in noisiness story for the highest-achieving students are larger than those I obtain in Figure 2. This explanation does account for the lower weights on lower-achieving students.[6]

4.4.1 Explanation 2: Are Certain Bins More Predictive of Other Bins?

I now turn to investigating to what extent the second explanation may be driving my results. Recall that explanation two is also a story about small-sample efficiency. A teacher's impact on a particular bin may be more informative about a teacher's impact on students in other bins. The high weight on the highest bins may simply reflect that a teacher's impact on the highest-achieving students is also more predictive about a teacher's impact on other students within the class in other bins. This explanation does not rule out that a teacher may have different impacts on higher versus lower-achieving students, which is explanation 3.

If the weights shown in Figure 2 are driven by such an explanation, then the outcome measure I am trying to best predict using student residuals should not change the observed pattern of the weights. Suppose this explanation is true. The highest-achieving students receive a higher weight because a teacher's impact on their high school graduation is more informative about a teacher's impact on high school graduation for other students in the same class. If I estimate weights to predict, say, a teacher's impact on students' test-scores in the subsequent grade, I would still expect that a teacher's impact on the highest achieving students to be more informative of a teacher's impact on other students (to the extent that

[6] I also rule out that smaller classes are driving my results. I report estimated weights for math and reading restricted to smaller (10-19 students) and larger (20-35 students) classes in Appendix A.

a teacher's high school graduation VA and future test-score VA are correlated). I would expect to observe a lower weight on lower-achieving students and a higher weight on the highest-achieving students for any outcome for which I can estimate a teacher's out of sample value-added.

I use a teacher's test-score VA for the next two years as my alternative outcomes. That is to say I estimate Equation 10 replacing a teacher's out of sample high school graduation VA with a teacher's out of sample VA on math and reading scores in both the next grade and the next next grade. For example, I estimate a third grade teacher's VA on her students' math and reading scores in fourth and fifth grade. I define a teacher's VA on her students' test scores in the next grade as a "1-year post" VA and scores in the next next grade as a "2-year post" VA. More formally, I define teacher j's pooled out-of-sample VA for test scores in subject s in future period $t+\tau$ as $\tilde{VA}^s_{j,t+\tau}$. I define $VA^*_{j,s,t}$ using student residuals for subject s in the current year (using Equation 7) and then estimate

$$\min_{(\Delta_k)s} \left[\tilde{VA}^s_{j,t+1} - \Delta_0 - VA^*_{j,s,t}\right]^2, \tag{13}$$

separately for each subject and each future time period.

I report results from estimating Equation 13 below in Figure 5. I include the initial weights from Figure 2 as hollow gray boxes. The results provide empirical support that this small-sample story may have a lot of explanatory power. I observe almost no differences in weights regardless of which outcome I estimate weights to try and predict.

An alternative way to assess such an small-sample explanation is by estimating weights to predict the impact of teacher j on test scores for the next cohort of students in year $t+1$. By definition the students used to calculate the outcome I am trying to predict on the right-hand side of the regression (a teacher's impact on test scores in the next year $t+1$) and the student residuals I am using as my left-hand side variables (from period t) are not the same set of students. Therefore I simply estimate a teacher's test-score VA separately for each year and estimate

$$\min_{(\gamma_k)s} \left[VA^s_{j,t+1} - \gamma_0 - VA^*_{j,s,t}\right]^2, \tag{14}$$

where $VA^s_{j,t+1}$ is teacher j's test-score VA in year $t+1$ for subject s.

I report results from estimating Equation 14 as Figure 6 below. I again notice the same pattern of weights. This further confirms the highest weights may reflect that a teacher's impacts on the highest-achieving students are simply more predictive of a teacher's impacts on other students in the same class. Recall that the relatively high noisiness of student residuals for the lowest-achieving students seems to explain the low weights on the lowest-achieving students.

4.4.2 Explanation 3: True Differences

While small-sample efficiency can explain both the large weights on the highest-achieving students and the small weights on the lowest-achieving students, recall that explanation 2 does not explicitly rule out a true-differences story for the lower- or higher-achieving students. In this section, I therefore assess the extent to which any and all of the weights I observe in Figure 2 can also be explained by true differences in teacher impacts.

I begin by asking to what extent the lower weights on the lower-achieving students might be due to true differences. Suppose instead of estimating weights to predict a teacher's high school graduation VA for all students, I estimated weights to predict a teacher's high school graduation VA specifically for the lowest-achieving students. I would expect to see higher weights for the lower-achieving students, compared to the Figure 2 weights, when I estimate weights to predict a teacher's high school graduation VA for low-achieving students ("low-achieving HS grad VA").

I therefore define each teacher's out of sample high school graduation VA for students in the bottom b percentile of the achievement distribution as $\bar{VA}^{grad}_{j,b,t}$. I define a student to be in the lowest-achieving group based on a student's combined lagged math and reading standardized test scores, relative to all students within the same school, year, and grade. I normalize this total test score to be a standard normal for each school, grade, and year. I use four alternative definitions of lowest-achieving students, (i) bottom 50% (ii) bottom 30%, (iii) bottom 25%, and (iv) bottom 20%. For each definition, I estimate

$$\min_{(\lambda_k)s} \left[\tilde{VA}^{grad}_{j,b,-t} - \lambda_0 - VA^*_{j,s,t} \right]^2, \tag{15}$$

where $VA^*_{j,s,t}$ is defined as described above in Equation 7, and $\tilde{VA}_{j,b,-t}$ is teacher j's low-achieving HS grad VA.

I report estimated bin-weights predicting these low-achieving high school graduation VA measures below as Figure 7. I also include the Figure 2 estimates (using all students within a classroom) as hollow gray boxes. Surprisingly, the weights on the lower-achieving and higher-achieving students do not differ when I change which students I use to calculate a teacher's high school graduation VA. The implication of this result is that even if I am predicting a teacher's impact on the high school graduation of the lowest-achieving students, it is a teacher's impact on the *highest*-achieving students that receives the highest weight. This is evidence that it is not just small-sample efficiency but true differences in teacher quality that drive my results.

The finding that the highest-achieving students are most predictive of outcomes for the lowest-achieving students warrants additional investigation. If this is a real result, I would also expect the highest-achieving students to receive the highest weight when I predict a teacher's impacts on test scores for any one particular group of students. To test this, I estimate weights that best predict a teacher's out-of-sample test-score VA for students in one particular bin. I repeat this for each of the five bins for both math and reading.

I calculate five bin-specific VA measures for each subject, one for each bin. For example, I calculate a teacher's bin 1 reading VA as a teacher's reading VA only for students who are in the bottom 20th percentile of lagged achievement in reading. I then estimate

$$\min_{(\Lambda_k)s} \left[\tilde{VA}^{s}_{j,b,-t} - \Lambda_0 - VA^{*}_{j,s,t} \right]^2, \tag{16}$$

where $\tilde{VA}^{s}_{j,b,-t}$ is the out-of-sample bin-specific VA for teacher j in year t limited to students in bin b for subject s. $VA^{*}_{j,s,t}$ is the weighted average of student residuals in year t for subject s as defined in Equation 7.

I report estimation results for Equation 16 for each subject below as Figure 8. I include the initial Figure 2 weights as hollow gray squares. I observe the same pattern of weights for both subjects *regardless* of which bin-specific VA measure I am trying to predict.

I can again estimate an alternative set of weights using a teacher's impact on the next cohort of students (similar to my approach described in Section 4.4.1). Here, I estimate each teacher's bin-specific VA for math and reading scores in the next year $t+1$. I use the residuals of all students in year t to estimate the weights that best predict each of the five possible bin-specific VA measures in $t+1$. I report results below as Figure 9. I find similar results to those shown in Figure 8. I table a more detailed discussion of the implications of these results to Section 5.

4.5 Estimating Weights Jointly vs Separately

Until this point I have estimated the optimal weights for math and reading separately. This is consistent with the way a conventional VA model is constructed, and so the estimated weights can easily be compared to a model in which all weights are equal. Now I allow students' math and reading residuals to jointly predict a teacher's out-of-sample high school graduation VA. This joint-estimation is more consistent with the literature finding that a teacher's impact is multidimensional. In particular, it follows the idea from Mulhern and Opper (2023) that a teacher's math VA should incorporate information about a teacher's estimate reading VA, and vice versa.

The way I estimate weights for math and reading jointly follows directly from how I initially estimated weights in Equation 10. I define teacher j's weighted VA for subject s in year t as $VA^{*}_{j,s,t}$ based on Equation 7. A teacher's out-of-sample high school graduation VA constructed in the usual way and denoted $\tilde{VA}^{grad}_{j,-t}$. I then estimate

$$\min_{(\kappa_k)s} \left[\tilde{VA}^{grad}_{j,-t} - \kappa_0 - VA^*_{j,read,t} - VA^*_{j,math,t}\right]^2, \tag{17}$$

where the only difference from Equation 10 is that I include VA^* for both subjects.

I report these results below as Figure 10, and I include the Figure 2 weights as hollow gray squares. The weights for the middle-achieving students (bins 2 and 4), whether I estimate weights using both subjects or each subject separately does not change the weights. For math, adding in a teacher's impacts on reading scores increases the weight on the lowest-achieving student from below to above 1. The weight on the highest-achieving students also increases. For reading, the weight on the lowest-achieving students decreases when I include a teacher's impacts on math scores. The weight on the highest-achieving students also increases.

4.6 How Much More Predictive is a Weighted VA Compared to a Conventional Value-Added

I set out to ask whether a weighted average of student residuals might be a more predictive measure of a teacher's long-run outcomes. So far, I have only shown that the estimated weights on different students are not equal, implying that a weighted VA is more predictive than a conventional VA. In this section, I construct each teacher's weighted VA measure based on the estimated weights from Section 4 in order to answer the question of just how much more predictive is this weighted VA compared to a conventional VA?

I use the same definition of a teacher's out-of-sample high school graduation VA ($\tilde{VA}^{grad}_{j,-t}$) as I did in Section 4. For each subject, I construct a teacher's conventional VA in year t for subject s as the average of student residuals ($VA_{j,s,t}$). I calculate that same teacher's weighted VA ($VA^*_{j,s,t}$) using the weights that best predict a teacher's VA on 2-year post test scores (Figure 5). I purposefully do not use the weights that best predict a teacher's VA on high school graduation (Figure 2) since a teacher's high school graduation VA will also be the variable I am using to test the relative predictive power of the weighted versus conventional VA measures.[7]

I define the increase in predictive power of a weighted VA measure (compared to a conventional VA measure) using the percentage of explained variation in teachers' high school graduation VA. Intuitively, I perform the following steps. First, I regress a teacher's out-of-sample high school graduation VA on a teacher's conventional VA for subject s. I calculate

[7] As discussed in Section 4, the estimated weights are almost equal for all outcome variables. Therefore the results I obtain in this section are also robust to using the estimated weights on different teacher outcomes.

the R-squared from this regression. I then regress a teacher's out-of-sample high school graduation VA on a teacher's *weighted* VA measure, and again calculate the R-squared. I then calculate the percentage increase in the R-squared when I use the weighted VA compared to the R-squared when I use the conventional VA as my main independent variable. I limit my sample to teachers with both a non-missing conventional and weighted VA for a particular subject.

I begin by showing these results without any other controls in Table 3. I report results for math in columns 1 and 2, and results for reading in columns 3 and 4. I cluster standard errors at the teacher level. The R-squared is higher when I use a weighted VA to predict high school graduation VA for both math ($\sim$ 20%) and reading ($\sim$ 8%) compared to a conventional VA in each subject. Both a weighted and conventional VA, however, explain less than 1% of the variation in teachers' high school graduation VA.

This is consistent with the literature which finds that (i) test-score impacts fade out over time and (ii) a teacher's impacts non-cognitive outcomes are more informative than a teacher's impacts on test scores. I address the second point by adding a teacher's pooled out-of-sample VA for both suspensions and my behavioral index (described in Section 2) as controls in the regressions I describe above. If my results are consistent with the literature, I would expect to see smaller increases in the percent of variation explained using a weighted versus conventional VA. I would also expect to see larger R-squared values overall when I include a teacher's non-cognitive impacts compared to those I observed in Table 3.

I report results controlling for a teacher's impact on non-cognitive outcome below in Table 4. Columns 1 through 4 show results for math, while columns 5 through 8 show results for reading. The first two columns for each subject show results without non-cognitive VA controls. My sample is limited to teachers with non-missing test-score VA and non-missing non-cognitive VA measures. I therefore have a smaller number of observations compared to Table 3. Including the baseline specification ensures comparability between specifications with and without controls.

My results are consistent with the evidence that a teacher's impacts on non-cognitive outcomes matter more for predicting a teacher's high school graduation VA than a teacher's test-score impacts. For both subjects, the R-squared roughly doubles when I include non-cognitive VA measures, regardless of which test-score VA I use. The percentage increase in explained variation due to a weighted test-score VA is also much lower when I include these non-cognitive VA measures. For math, the percentage increase in explained variation decreases from $\sim$ 27% to $\sim$ 7.5% when I include non-cognitive VA measures. For reading, the percentage increase in explained variation decreases from $\sim$ 9% to $\sim$ 4%.

Finally, I conduct the same analysis as shown in Table 4 except for I use both a teacher's conventional math and reading test-score VA to jointly predict a teacher's high school graduation VA. For the weighted VA, I use the weights I obtain in Figure 10. I report results below as Table 5. I observe similar results to Table 4 in which (i) the R-squared is much higher when I include a teacher's non-cognitive VA measures and (ii) the increase in percentage of explained variation when I use weighted VA compared to conventional VA decreases (from $\sim 13\%$ to $\sim 6\%$) when I include non-cognitive VA measures as controls.

4.7 Wrap-Up

I have shown in this section that a conventional VA measure is not the most predictive of a teacher's long-run impacts, specifically a teacher's high school graduation VA. A more predictive measure is a weighted VA in which the highest-achievement students (according to baseline achievement) should receive a higher weight than the median student. Lower-achieving students should receive a lower weight, more so for reading test scores than for math test scores. The increase in predictive power (as measured using the percentage increase in R-squared values) is modest for math and reading. The increase in predictive power is smaller when regressions include a teacher's impact on non-cognitive outcomes.

I find that the weights I estimate in Figure 2 are due to both small-sample efficiency and true differences in teacher quality. Lower-achieving students have more noise in their test-score residuals, and this seems to explain why lower-achieving students receive a lower weight. While higher-achieving students have relatively less noise in their residuals, this explanation implies a weight on the highest-achieving students which is lower than the weights I observe. The highest-achieving students instead seem to receive a higher weight at least in part because a teacher's impact on the highest-achieving students is most predictive of students in other bins. This is highlighted by the result that the highest-achieving students receive the highest weight even when predicting a teacher's long-run impacts on specifically the lowest-achieving students. In fact, the highest-achieving students receive the highest weight no matter which specific part of the class I am using to measure a teacher's impact. This evidence is most consistent with the explanation that teachers also differ in their impacts on students with different baseline achievement.

5 Discussion

A weighted average of student residuals improves the accuracy of test-score VA measures in predicting a teacher's long-term impacts relative to an unweighted average. This is a modest improvement (with estimates around $\sim 20\%$ for math and $\sim 6\%$ for reading) when predicting a teacher's impact on high school graduation. Test-score VA, even a weighted VA, explains very little of the variation in a teacher's long-run impact. A teacher's non-cognitive impacts are still more important than a teacher's test-score even if a test-score VA is constructed (i) accounting for a teacher's impacts in both math and reading and (ii) with weights on different students estimated with the explicit goal of best predicting a teacher's long-run impact.

Across all outcomes, the highest-achieving students (based on baseline achievement) receive the highest weight. For reading, the lowest-achieving students receive the lowest weight. On average, I estimate the highest-achieving students should be weighted about 1.5 times as much as the median student. For reading, the lowest-achieving students should be weighted about 0.5 times as much as the median student. These results speaks to two aspects of value-added models in general. First, VA models estimate teacher effects using a small sample (students in a single classroom). My results imply that weighting students differently improves the small-sample efficiency of VA models in predicting teachers' long-run impacts. The highest-achieving students simply have more predictive power compared to lower-achieving students in predicting how a teacher affects outcomes across all students within a class.

Second, these results are suggestive as to why policymakers might care about teachers who are estimated to be good at raising test scores. One possible answer is that raising test scores is important for its own sake. Students with higher test scores are just more likely to have good long-run outcomes. The other possible answer is that being good at raising test scores is an indicator for the type of teacher who is good at promoting long-run outcomes in students. This latter explanation is suggested by my empirical result that the highest-achieving students should always receive the highest weight, even when estimating weights to predict long-run outcomes for the lowest-achieving students.

This finding suggests that teachers might differ in their teaching styles. I hypothesize that being able to increase test scores for the highest-achieving students requires different skills than being able to increase test scores for other students in a classroom. Raising test scores for the highest-achieving students, for example, might require teaching students more advanced critical thinking skills than being able to raise test scores for the lowest-achieving students. For those students, it might be more important to be able to improve basic literacy and numeracy skills.

This means that the ability to increase test scores for any student is not important because the test-scores themselves are predictive of a student's long-run outcomes. Instead, being able to increase test scores (for the highest-achieving students in particular) is important because it reveals whether a particular teacher has latent characteristics which make her good at promoting skills in her students that translate into better long-run outcomes.

References

Aucejo, Esteban et al. (2022). "Teacher effectiveness and classroom composition: Understanding match effects in the classroom". In: *The Economic Journal* 132.648, pp. 3047–3064.

Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger (2014). *Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles.* NBER Working Paper 20657. National Bureau of Economic Research.

Backes, Benjamin et al. (2023). *How to Measure a Teacher: The Influence of Test and Nontest Value-Added on Long-Run Student Outcomes.* Working Paper. CALDER Center.

Biasi, Barbara, Chao Fu, and John Stromme (2021). *Equilibrium in the market for public school teachers: District wage strategies and teacher comparative advantage.* Tech. rep. National Bureau of Economic Research.

Blazar, David and Matthew A. Kraft (2017). "Teacher and Teaching Effects on Students' Attitudes and Behaviors". In: *American Economic Review* 107.5, pp. 146–150. DOI: `10.1257/aer.p20171049`.

Chamberlain, Gary E. (2013). "Predictive Effects of Teachers and Schools on Test Scores, College Attendance, and Earnings". In: *Proceedings of the National Academy of Sciences* 110.43, pp. 17176–17182. DOI: `10.1073/pnas.1315746110`.

Chetty, Raj, John Friedman, and Jonah Rockoff (2014a). "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates". In: *American economic review* 104.9, pp. 2593–2632.

— (2014b). "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood". In: *American economic review* 104.9, pp. 2633–2679.

Chetty, Raj et al. (2011). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR". In: *The Quarterly Journal of Economics* 126.4, pp. 1593–1660. DOI: `10.1093/qje/qjr041`.

Condie, Scott, Lars Lefgren, and David Sims (2014). "Teacher heterogeneity, value-added and education policy". In: *Economics of Education Review* 40, pp. 76–92.

Delgado, William (2020). "Heterogeneous teacher effects, comparative advantage, and match quality: Evidence from Chicago public schools". In: *Manuscript, Boston Univ.*

Deming, David (2009). "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start". In: *American Economic Journal: Applied Economics* 1.3, pp. 111–134. DOI: `10.1257/app.1.3.111`.

Eastmond, Tanner S. et al. (2024). "From Value Added to Welfare Added: A Social Planner Approach to Education Policy and Statistics". Unpublished manuscript.

Fox, Lindsay (2016). “Playing to teachers’ strengths: Using multiple measures of teacher effectiveness to improve teacher assignments”. In: *Education Finance and Policy* 11.1, pp. 70–96.

Gershenson, Seth et al. (2022). “The long-run impacts of same-race teachers”. In: *American Economic Journal: Economic Policy* 14.4, pp. 300–342.

Gilraine, Michael and Nolan G. Pope (2021). *Making Teaching Last: Long-Run Value-Added*. NBER Working Paper 29555. National Bureau of Economic Research.

Graham, Bryan S et al. (2023). “Teacher-to-classroom assignment and student achievement”. In: *Journal of Business & Economic Statistics* 41.4, pp. 1328–1340.

Hanushek, Eric A and Steven G Rivkin (2010). “Generalizations about using value-added measures of teacher quality”. In: *American economic review* 100.2, pp. 267–271.

Jackson, C Kirabo (2018). “What do test scores miss? The importance of teacher effects on non–test score outcomes”. In: *Journal of Political Economy* 126.5, pp. 2072–2107.

— (2012). *Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina*. NBER Working Paper 18624. National Bureau of Economic Research.

Kane, Thomas J. and Douglas O. Staiger (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*. NBER Working Paper 14607. National Bureau of Economic Research.

Koedel, Cory, Kata Mihaly, and Rockoff (2015). “Value-added modeling: A review”. In: *Economics of Education Review* 47, pp. 180–195. ISSN: 0272-7757. DOI: `https://doi.org/10.1016/j.econedurev.2015.01.006`. URL: `https://www.sciencedirect.com/science/article/pii/S0272775715000072`.

Konstantopoulos, Spyros (2009). “Effects of teachers on minority and disadvantaged students’ achievement in the early grades”. In: *The Elementary School Journal* 110.1, pp. 92–113.

Lavy, Victor and Rigissa Megalokonomou (2024). *Alternative Measures of Teachers’ Value-Added and Impact on Short- and Long-Term Outcomes: Evidence From Random Assignment*. NBER Working Paper 32671. National Bureau of Economic Research.

Lavy, Victor, M Daniele Paserman, and Analia Schlosser (2012). “Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom”. In: *The Economic Journal* 122.559, pp. 208–237.

Mulhern, Christine and Isaac Opper (2023). “Measuring and summarizing the multiple dimensions of teacher effectiveness”. In.

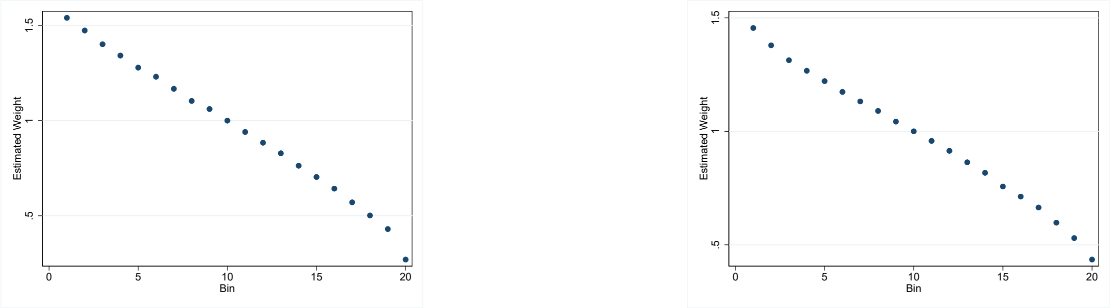
Nielsen, Eric (2019). “Test questions, economic outcomes, and inequality”. In.

Petek, Nathan and Nolan G. Pope (2023). “The Multidimensional Impact of Teachers on Students”. In: *Journal of Political Economy* 131.4, pp. 1057–1107. DOI: 10.1086/722227.

Rivkin, Steven, Eric Hanushek, and Kain (2005). “Teachers, Schools, and Academic Achievement”. In: *Econometrica* 73.2, pp. 417–458. DOI: 10.1111/j.1468-0262.2005.00584.x.

6 Tables and Figures

Figure 1: Toy Model Estimated Weights By Vintiles of Lagged Achievement


(a) Math Weights

(b) Reading Weights

Notes: Each dot represents the average weight on students with a lagged achievement within a particular vintile of lagged achievement. Figure 1a shows weights based on lagged math achievement, and Figure 1b shows weights based on lagged reading achievement. Weights are calculated as the estimated slope in a student's predicted latent, or underlying, propensity to graduate according to the model given by Equation 6. Weights are normalized within each classroom such that the sum of student-level weights sum to one within each classroom. I restrict my analysis to fourth and fifth grades in classrooms with between 10 and 35 students between 1998 and 2011.

Table 1: Summary Statistics of Student Data

Variable	Mean	SD	Min	Max
Female	0.510	(0.500)		
Black	0.235	(0.424)		
Hispanic	0.053	(0.225)		
White	0.483	(0.500)		
Asian	0.0159	(0.125)		
Economically Disadvantaged	0.838	(0.368)		
Student With Disabilities	0.157	(0.364)		
Academically Gifted	0.0282	(0.166)		
English Language Learner	0.125	(0.111)		
Ever Suspended	0.36	(0.48)		
Graduated High School	0.805	(0.396)		
-Ln(1+Absences)	-0.0973	(0.437)	-4.95	0
-Days Suspended	-0.0880	(1.314)	-447	0
Classroom Size	22.869	(3.73)	10	35
Student-Year Observations		2,587,625		
Students		1,633,504		

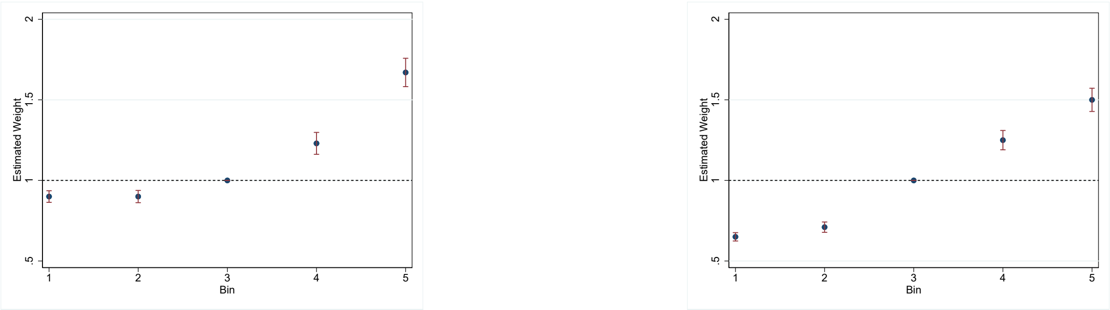
Notes: Table reports summary statistics for student demographic, non-cognitive outcomes, and high school graduation. I restrict my sample to students in classes with between 10 and 35 students. I do not report min and max values for indicator variables that take values of either 0 or 1. I calculate absences as the inverse of the natural log of 1 plus the number of days a student is absent (consistent with Mulhern and Opper, 2023). I multiple days suspended by negative 1 such that fewer days suspended is a better student outcome. My analysis sample is limited to 1998-2011 (due to the lack of lagged test score information in 1997) for students in fourth and fifth grade. Lagged test-scores are required for inclusion in my analyses, and third-graders do not take a standardized test in second grade.

Table 2: Estimated Latent Correlation Among Top/Bottom Value-Added

		Baseline		Classroom Moments	
Split	Subject	Correlation	Std. Error	Correlation	Std. Error
		(1)	(2)	(3)	(4)
Top/Bottom 50%	Math	0.836	0.00227	0.811	0.0024
	Reading	0.682	0.00431	0.645	0.00436
Top/Bottom 30%	Math	0.590	0.00463	0.590	0.00470
	Reading	0.439	0.00672	0.313	0.00743
Top/Bottom 25%	Math	0.509	0.00531	0.531	0.00524
	Reading	0.362	0.00722	0.212	0.00795

Notes: Top/Bottom XX% represent the split used to determine value-added (VA) for top and bottom students. For example, Top/Bottom 50% Math indicates I define top math students as students with a lagged math score above the median lagged math score within a particular school, grade, and year. I define bottom 50 math students as students with a lagged math score below or equal to this same median. Correlation (columns 1 and 3) and Std. Error (columns 2 and 4) represent the estimated correlation coefficient and standard error from a maximum likelihood estimator as described in Section 3.1. VA is defined as the average of unadjusted residuals defined in Equation 1 restricted to either top or bottom students. Columns 1 and 2 do not include any additional controls. In columns 3 and 4 I include the class-level mean and variance of lagged achievement as additional controls. Restricted to fourth and fifth grade classes of between 10 and 35 students from 1998 to 2011.

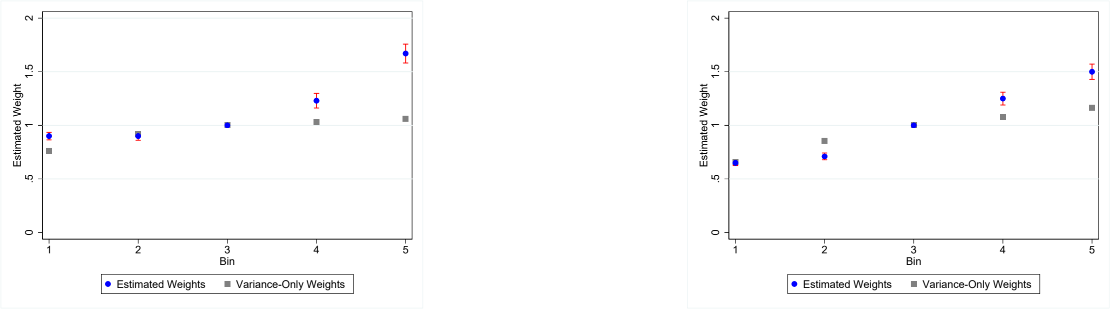
Figure 2: Empirical Bin-Weight Estimates for Math and Reading


(a) Math

(b) Reading

Notes: Each dot represents the estimated weight on each bin from estimating Equation 10. Bars indicate the 95% confidence interval around each estimated weight. The weight on bin 3 is defined to be 1. Figure 2a shows weights for student math residuals and Figure 2b shows weights for student reading residuals. Student residuals are calculated based on Equation 1. Limited to fourth and fifth-grade classes with between 10 and 35 students.

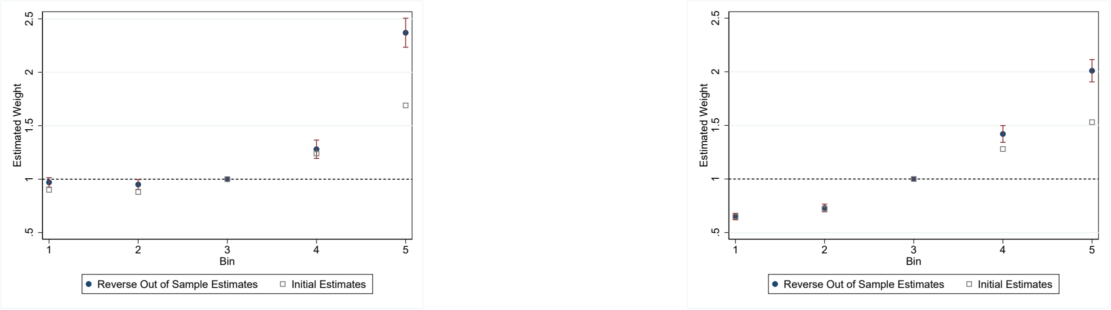
Figure 3: Estimated and Variance-Based Bin-Weights by Subject


(a) Math

(b) Reading

Notes: I report variance-only weights as gray squares in the above graph. The blue dots with error bars represent the initial estimated weights and their 95% confidence intervals from Figure 2. Figure 3a shows results for math, and Figure 3b shows shows results for reading. Limited to fourth and fifth-grade classes with between 10 and 35 students.

Figure 4: Reverse Out of Sample Bin-Weight Estimates by Subject

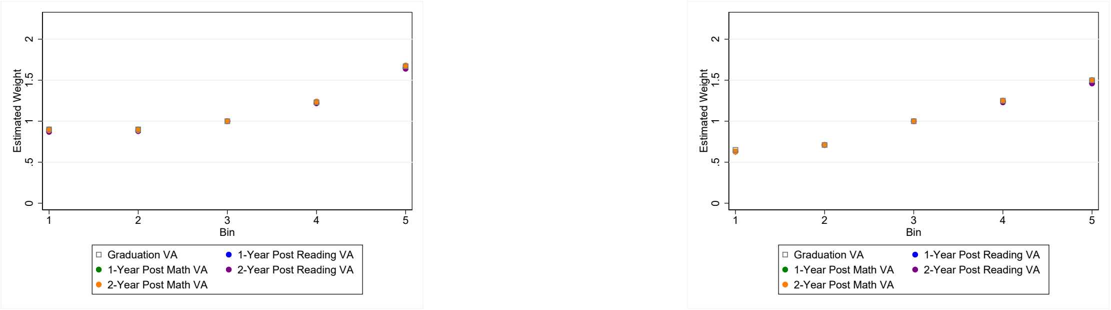


(a) Math

(b) Reading

Notes: The blue dots with error bars represent estimated out-of-sample weights and their 95% confidence intervals. I report the initial weights from Figure 2 as hollow gray squares. Figure 4a shows results for math, and Figure 4b shows shows results for reading. Limited to fourth and fifth-grade classes with between 10 and 35 students.

Figure 5: Estimated Bin-Weights Predicting Future Test-Score VA

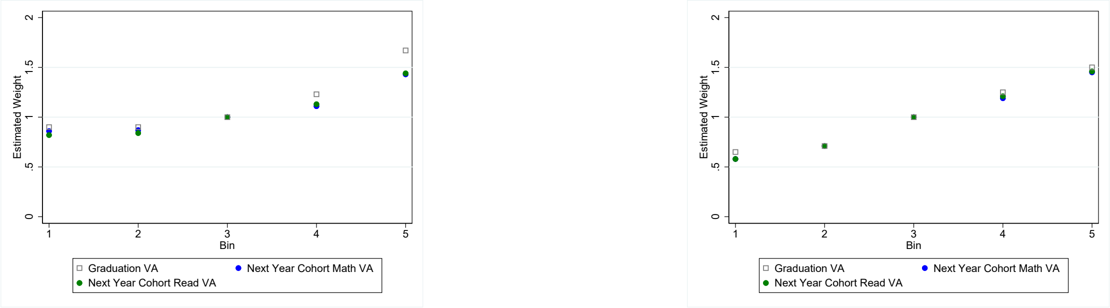


(a) Math

(b) Reading

Notes: Each dot represents an estimated weight. For each outcome, the weight on bin 3 is normalized to 1. I report initial weights from Figure 2 as hollow gray squares. Figure 3a shows results for math, and Figure 3b shows shows results for reading. Limited to fourth and fifth-grade classes with between 10 and 35 students.

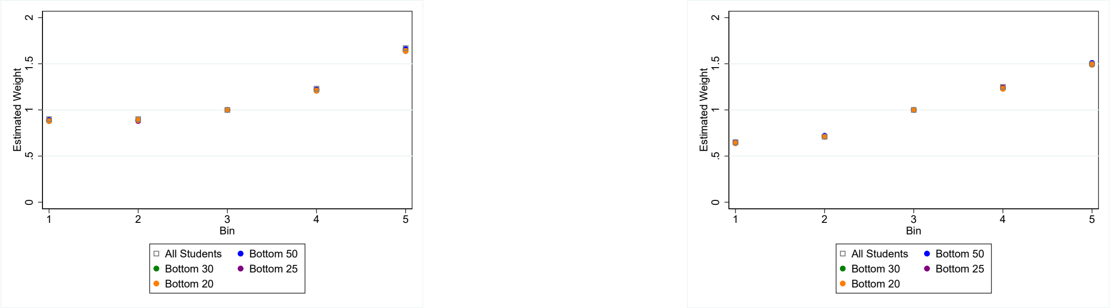
Figure 6: Estimated Bin-Weights Predicting Subsequent Cohort VA


(a) Math

(b) Reading

Notes: Each dot represents an estimated weight. For each outcome, the weight on bin 3 is normalized to 1. I report initial weights from Figure 2 as hollow gray squares. Next Year Cohort indicates a teacher's estimated VA using math or reading scores for students in the next year. Figure 6a shows results for math, and Figure 6b shows shows results for reading. Limited to fourth and fifth-grade classes with between 10 and 35 students.

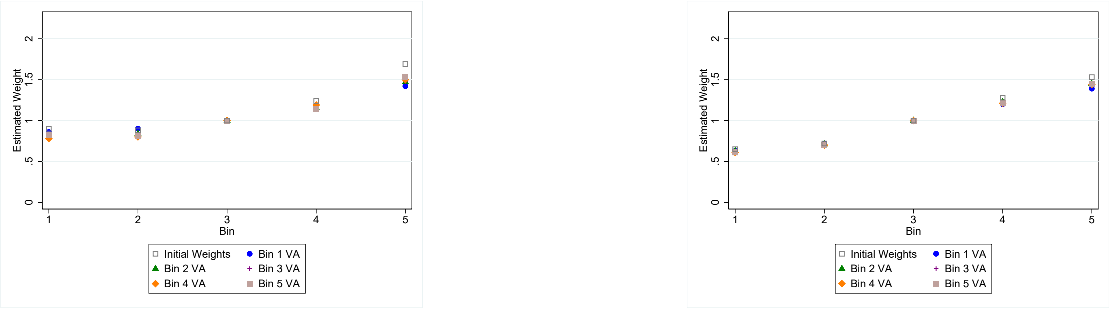
Figure 7: Estimated Bin-Weights Predicting Low-Achieving HS Grad VA


(a) Math

(b) Reading

Notes: Each dot represents an estimated weight. For each outcome, the weight on bin 3 is normalized to 1. I report initial weights from Figure 2 as hollow gray squares. The legend indicates which students I used to calculate a teacher's out-of-sample high school graduation VA. For example, "Bottom 50" indicates I calculated a teacher's out-of-sample high school graduation VA only for students in the bottom 50th percentile of combined baseline math and reading achievement. This baseline measure is a normalized sum of lagged math and reading standardized test scores at the school, grade, and year level. Figure 7a shows results for math, and Figure 7b shows shows results for reading. Limited to fourth and fifth-grade classes with between 10 and 35 students.

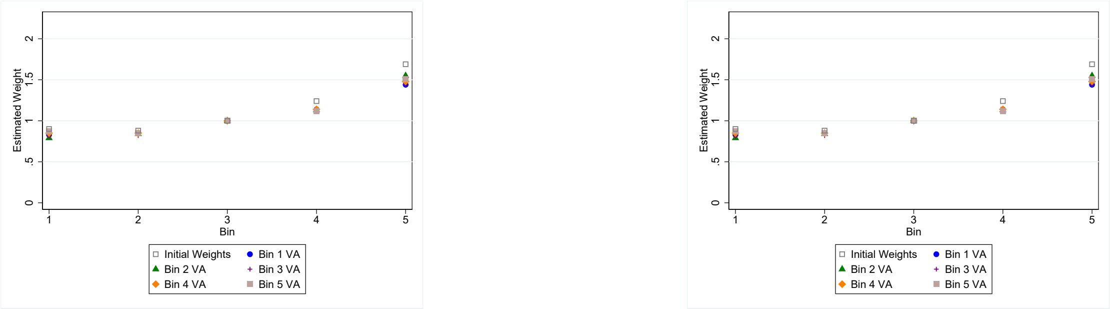
Figure 8: Estimated Bin-Weights Bin-Specific VA


(a) Math

(b) Reading

Notes: Each dot represents an estimated weight. For each outcome, the weight on bin 3 is normalized to 1. I report initial weights from Figure 2 as hollow gray squares. The legend indicates which students I used to calculate a teacher's out-of-sample test-score VA in each subject. For example, "Bin 1" in Figure 8a indicates I calculated a teacher's out-of-sample math VA using only the lowest-achieving students, or the students with a baseline standardized math test score in the bottom 20th percentile. This baseline measure is relative to standardized test scores in each subject at the school, grade, and year level. Figure 8a shows results for math, and Figure 8b shows shows results for reading. Limited to fourth and fifth-grade classes with between 10 and 35 students.

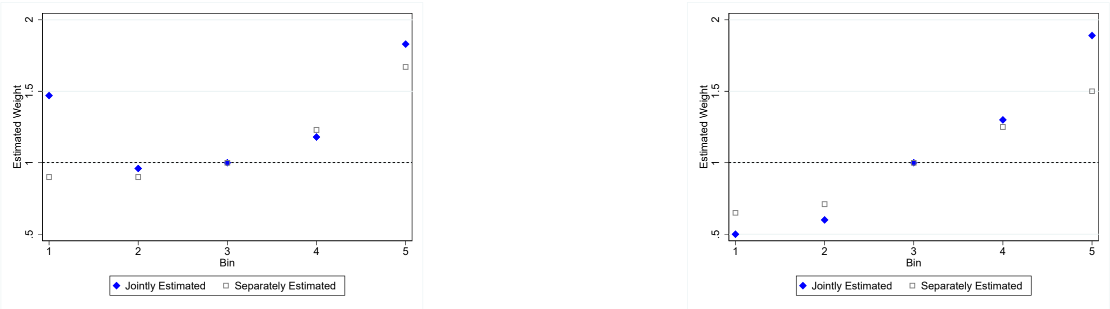
Figure 9: Estimated Bin-Weights Next Cohort Bin-Specific VA


(a) Math

(b) Reading

Notes: Each dot represents an estimated weight. For each outcome, the weight on bin 3 is normalized to 1. I report initial weights from Figure 2 as hollow gray squares. The legend indicates which students I used to calculate a teacher's test-score VA in the next year in each subject. For example, "Bin 1" in Figure 9a indicates I calculated a teacher's math VA in the next year using only the lowest-achieving students, or the students with a baseline standardized math test score in the bottom 20th percentile. This baseline measure is relative to standardized test scores in each subject at the school, grade, and year level. Figure 9a shows results for math, and Figure 9b shows shows results for reading. Limited to fourth and fifth-grade classes with between 10 and 35 students.

Figure 10: Jointly Estimated Bin-Weights for Math and Reading



(a) Math

(b) Reading

Notes: Each dot represents an estimated weight. For each subject, the weight on bin 3 is normalized to 1. I report initial weights from Figure 2 as hollow gray squares. The weights (shown in blue) are estimated to maximize the predictive power of predicting a teacher's high school graduation VA using both math and reading as shown in Equation 17 discussed in Section 4.5. Figure 10a shows results for math, and Figure 10b shows shows results for reading. Limited to fourth and fifth-grade classes with between 10 and 35 students.

Table 3: Predictive Power of Weighted vs Unweighted VA: Baseline

	Math		**Reading**	
	(1)	(2)	(3)	(4)
Conventional (Unweighted) VA	0.000566***		0.000967***	
	(0.000114)		(0.000127)	
Weighted VA		0.000581***		0.000984***
		(0.000106)		(0.000123)
Observations	84,704	84,704	84,678	94,678
R^2	0.0010	0.0012	0.0017	0.0019
% Increase in Explained Variation	–	19.56	–	8.15

Notes: Table 3 reports results from regressing a teacher's out-of-sample high school graduation value-added (VA) on a teacher's test-score VA in a particular subject. There are no other controls in the regression. I report clustered-standard errors at the teacher level in parentheses. Conventional (Unweighted) VA indicates test-score VA is calculated as the unweighted average of student residuals in a the subject indicated by the column heading. Weighted VA indicates test-score VA is calculated using the estimated weights predicting a teacher's impact on test-scores for students 2 years in the future. These weights are shown in Figure 5. The percentage (%) increase in explained variation for each subject is calculated as the percentage change in the R-squared value of the regression when I use the weighted VA as the regressor compared to when I use the conventional VA for the same subject as the regressor. Limited to fourth and fifth grade classes to between 10 and 35 students. $p < 0.001^{***}, p < 0.05^{**}, p < 0.1^{*}$.

Table 4: Predictive Power of Weighted vs Unweighted VA: Non-Cognitive

	Math				**Reading**			
	Baseline		**Non-Cognitive**		**Baseline**		**Non-Cognitive**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Conventional (Unweighted) VA	0.000508***		0.000439***		0.00984***		0.000898***	
	(0.000134)		(0.000133)		(0.000150)		(0.000147)	
Weighted VA		0.000538***		0.00463***		0.00100***		0.000921***
		(0.000125)		(0.000123)		(0.00146)		(0.00143)
Suspension VA			0.0112***	0.0112**			0.0112**	0.0112**
			(0.00421)	(0.00421)			(0.00420)	(0.00420)
Behavioral Index VA			0.00400*	0.00388*			0.00371*	0.00367**
			(0.00208)	(0.00208)			(0.00207)	(0.00207)
Observations	67,595	67,595	67,595	67,595	67,571	67,571	67,571	67,571
R^2	0.0008	0.0010	0.0023	0.0025	0.0016	0.0018	0.0031	0.0033
% Increase in Explained Variation	–	26.73	–	7.52	–	8.96	–	4.25

Notes: Table 4 reports results from regressing a teacher's out-of-sample high school graduation value-added (VA) on a teacher's test-score VA in a particular subject. I report clustered standard errors at the teacher level in parentheses. Conventional (Unweighted) VA indicates test-score VA is calculated as the unweighted average of student residuals in a the subject indicated by the column heading. Weighted VA indicates test-score VA is calculated using the estimated weights predicting a teacher's impact on test-scores for students 2 years in the future. These weights are shown in Figure 5. Suspension VA is defined as a teacher's out-of-sample VA for reducing number of days suspended. Behavioral Index VA is defined as a teacher's out-of-sample VA for increasing a student's behavioral index score (described in Appendix B). Columns 1,2, 4, and 5 represent analogous results from Table 3 restricted to teacher-years with non-missing non-cognitive VA measures. Columns 3,4,7, and 8 show results include the non-cognitive VA measures as controls. The percentage (%) increase in explained variation for each subject and whether or not I include non-cognitive VA measures as controls is calculated as the percentage change in the R-squared value of the regression when I use the weighted VA as the main regressor compared to when I use the conventional VA for the same subject as the main regressor. Limited to fourth and fifth grade classes to between 10 and 35 students. $p < 0.001^{***}, p < 0.05^{**}, p < 0.1^{*}$.

Table 5: Predictive Power of Weighted vs Unweighted VA: Joint Estimation

	Baseline		**Non-Cognitive**	
	(1)	(2)	(3)	(4)
Conventional Math VA	0.000143		0.000100	
	(0.000147)		(0.000146)	
Conventional Reading VA	0.000879***		0.000826***	
	(0.000158)		(0.000157)	
Weighted Math VA		0.000175		0.000135
		(0.000128)		(0.000128)
Weighted Reading VA		0.000827***		0.000779***
		(0.000138)		(0.000123)
Suspension VA			0.0111***	0.0111**
			(0.00420)	(0.00420)
Behavioral Index VA			0.00366*	0.00359*
			(0.00207)	(0.00208)
Observations	67,549	67,549	67,549	67,549
R^2	0.0017	0.0019	0.0031	0.0033
% Increase in Explained Variation	–	13.27	–	6.36

Notes: Table 5 reports results from regressing a teacher's out-of-sample high school graduation value-added (VA) on a teacher's test-score VA for both math and reading. I report clustered standard errors at the teacher level in parentheses. Conventional (Unweighted) VA indicates test-score VA is calculated as the unweighted average of student residuals for each subject. Weighted VA indicates test-score VA is calculated using the estimated weights predicting a teacher's impact on graduation using both math and reading test-score residuals. These weights are shown in Figure 10. Suspension VA is defined as a teacher's out-of-sample VA for reducing number of days suspended. Behavioral Index VA is defined as a teacher's out-of-sample VA for increasing a student's behavioral index score (described in Appendix B). Columns 1 and 2 report results without controlling for non-cognitive outcomes. In columns 3 and 4 I also control for a teacher's non-cognitive VA measures. The percentage (%) increase in explained variation for each subject and whether or not I include non-cognitive VA measures as controls is calculated as the percentage change in the R-squared value of the regression when I use the weighted VA as the main regressor compared to when I use the conventional VA for the same subject as the main regressor. Limited to fourth and fifth grade classes to between 10 and 35 students. $p < 0.001^{***}, p < 0.05^{**}, p < 0.1^{*}$.

A Appendix A: Additional Tables and Figures

I include summary stats for my estimated value-added (VA) measures by subject and portion of the class below in Table A1. Each row indicates both which students are included in estimating a teacher's VA and whether the VA uses math or reading test scores. The mean of all VA measures is 0, and instead I report the standard deviation (Std. Dev) of each VA measure. Column 1 indicates the students used in the calculation of each VA. Whole-Class, for example, indicates the regression residuals of all students are included in calculating a teacher's VA, while Top 50% indicates only residuals for students above the median of lagged achievement in a particular subject (relative to a student's school, year, and grade) are included. Column 3 indicates the raw, or unadjusted VA measures calculated based on Equation 1 and Equation 4. To calculate an adjusted, or shrunk VA measure, I follow the procedure for estimating VA using all years according to Mulhern and Opper (2023). I assume there is no drift in teacher VA. I calculate the Empirical Bayes estimate for each teacher and subject, and report the Std. Dev of these measures in Column 4. This is equivalent to assuming that a teacher's impact on math or reading scores is independent of the same teacher's impact on the other subject's test scores. I denote this as 1D shrinkage. I relax this assumption in Column 5, which I denote 2D shrinkage. Finally, I report the number of teachers for which I have non-missing data for Column 5 in Column 6.

I also rule out that this explanation is fully driving my results by estimating Equation 10 separately for smaller (10-19 students) and larger (20-35 students) classes.[8] I find no significant qualitative nor quantitative difference in my estimated weights for reading and math scores. I include those results below as Figure A1 (reading) and Figure A2 (math).

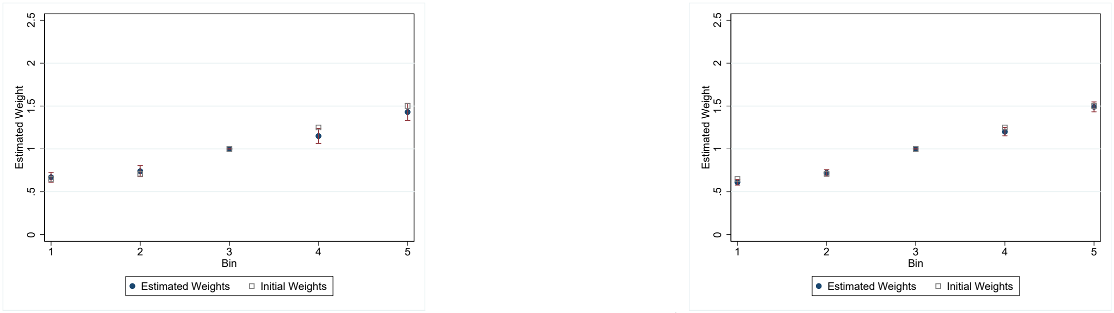
[8] Recall that I restrict all analyses to classes with between 10 and 35 students.

Table A1: Summary of Value-Added Measures by Dimensionality (Grades 3-5) Pooled Years

Group	Subject	Unadjusted Std. Dev	1D Std. Dev	2D Std. Dev	2D Teachers
(1)	(2)	(3)	(4)	(5)	(6)
Whole-Class (Typical VA)	Math	0.209	0.108	0.149	45,221
	Reading	0.155	0.0557	0.0711	45,221
Top 50%	Math	0.216	0.0851	0.132	35,763
	Reading	0.157	0.0282	0.0371	35,763
Bottom 50%	Math	0.230	0.102	0.146	41,209
	Reading	0.191	0.0541	0.0622	41,211
Top 30%	Math	0.224	0.0834	0.122	30,245
	Reading	0.169	0.0254	0.0298	30,246
Bottom 30%	Math	0.246	0.0969	0.140	38,988
	Reading	0.222	0.0529	0.0590	39,005
Top 25%	Math	0.230	0.0839	0.120	29,156
	Reading	0.177	0.0252	0.0296	29,156
Bottom 25%	Math	0.253	0.0948	0.135	38,026
	Reading	0.234	0.0529	0.0558	38,032

Notes: This table reports the student-weighted standard deviation of estimated teacher value-added (VA) measures for the students and subject indicated in columns 1 and 2. Top/Bottom XX% represent the split used to determine value-added (VA) for top and bottom students. For example, Top/Bottom 50% Math indicates I define top math students as students with a lagged math score above the median lagged math score within a particular school, grade, and year. I define Bottom 50% math students as students with a lagged math score below or equal to this same median. Column 3 represents the standard deviation (Std. Dev) of student residuals within a split and subject across teachers, or each teacher's unadjusted VA measure. For columns 4 and 5 I apply the methodology detailed in Mulhern and Opper (2023). In column 4 I assume a teacher's VA within a particular split is independent across math and reading. In column 5 I relax this assumption (which most aligns with Mulhern and Opper, 2023). I include the number of teachers included in my Column 5 estimates in Column 6. These estimates use all years available for a given teacher. I restrict to classrooms with between 10 and 35 students and to student test-scores in grades 4 and 5. Note that my effective sample is limited to 1998-2011 as there are no lagged test scores in the 1997 data, the earliest year I have available.

Figure A1: Reading Bin-Weight Estimates by Class Size

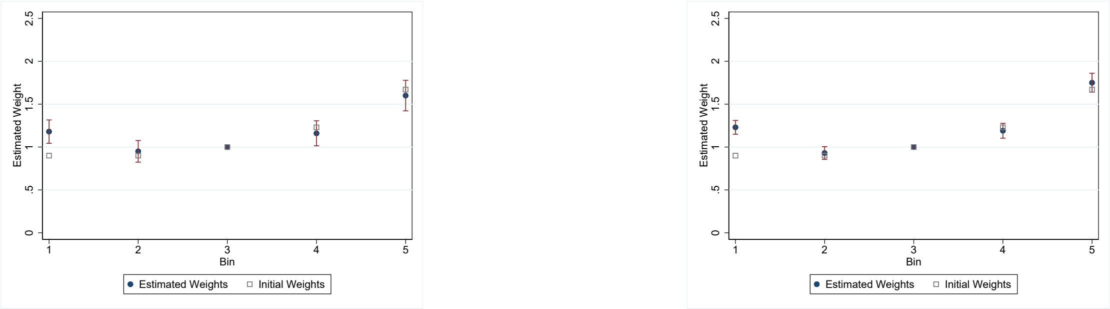


(a) Small (10-19 students) Classes

(b) Large (20-35 students) Classes

Notes: Each dot represents the estimated weight on each bin from estimating Equation 10. Bars indicate the 95% confidence interval around each estimated weight. The weight on bin 3 is defined to be 1. Figure 2a shows weights for student math residuals and Figure 2b shows weights for student reading residuals. Student residuals are calculated based on Equation 1. Limited to fourth and fifth-grade classes with between 20 and 35 students.

Figure A2: Math Bin-Weight Estimates by Class Size



(a) Small (10-19 students) Classes

(b) Large (20-35 students) Classes

Notes: Each dot represents the estimated weight on each bin from estimating Equation 10. Bars indicate the 95% confidence interval around each estimated weight. The weight on bin 3 is defined to be 1. Figure 2a shows weights for student math residuals and Figure 2b shows weights for student reading residuals. Student residuals are calculated based on Equation 1. Limited to fourth and fifth-grade classes with between 10 and 19 students.

B Appendix B: Creating Students' Behavioral Index

I discuss in this section my methodology for constructing a student's behavioral index which I use to calculate a teacher's behavioral index VA. I first conduct a principal component analysis (PCA) using the number of (i) short-term suspension days, (ii) long-term suspension days, (iii) number of incidents leading to short-term suspension, (iv) number of incidents leading to long-term suspension, (v) number of incidents leading to expulsion, (vi) number of incidents leading to detention, (vii) number of incidents leading to privileges being revoked, (viii) number of incidents leading to other consequences, (ix) number of incidents leading to a 1-year suspension, and (x) number of days in which a student is sent home from school. I report summary statistics for these measures below as Table B1.

Table B1: Summary Statistics of Student Data: Behavior

Variable	Mean	SD	Min	Max
Short-term Suspension Days	0.152	1.000	0	58
Long-term Suspension Days	0.003	0.351	0	90
Short-term Suspensions	0.229	0.721	0	15
Long-term Suspensions	0.000	0.021	0	1
Times Expelled	0.000	0.005	0	1
Times Detention	0.001	0.032	0	6
Times Privileges Revoked	0.001	0.048	0	9
Times Other Consequences	0.001	0.030	0	4
Times Received 1-year Suspension	0.000	0.004	0	1
Times Sent Home	0.008	0.145	0	13

Notes: Table reports summary statistics for student behavioral outcomes used to construct my behavioral index. These summary statistics are for students who appear in the suspension data. I restrict my sample to students in classes with between 10 and 35 students. Data is available from 2001 to 2011. I exclude 2001 due to a lack of available lagged-behavioral information for students.

I report the results from the PCA estimation using one principal component below in Table B2. I then generate a behavioral index as the predicted value from this PCA regression. I flip the sign of this index so that a more negative, i.e. lower, index represents worse behavior and an index closer to 0 represents better behavior or fewer disciplinary consequences.

When I merge this behavioral index onto the test-score data there are some students who do not appear in the suspension/disciplinary data. I interpret this as these students did not have any disciplinary incidents in a given year. I therefore assign these students as having a higher behavioral index than the best-behaving students with a non-missing behavioral index value. I then transform this behavioral index variable into a standard normal variable.

Table B2: Principal Component Analysis Loadings: Component 1

Variable	Component 1 Loading
Short-term Suspension Days	0.5317
Long-term Suspension Days	0.4345
Short-term Suspensions	0.5408
Long-term Suspensions	0.4686
Times Expelled	0.0459
Times Detention	0.0315
Times Privileges Revoked	0.0308
Times Other Consequences	0.0623
Times Received 1-year Suspension	0.0887
Times Sent Home	-0.0259
Eigenvalue	1.68
Variance Explained (%)	16.81

Notes: Table B2 shows results from a principal component analysis (PCA) on the variables included in the table. I include summary statistics for these outcomes above in Table B1. I use the predicted value from this PCA to generate an initial behavioral index for students. I multiply this inverse by -1 such that a behavioral index closer to 0 (smaller in absolute value) represent better-behaved students. Students not in the suspension data receive a behavioral index score higher than the best-behaved students in the suspension data.