NATIONAL
CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

Accounting for Student Disadvantage in Value-Added Models

Eric Parsons
Cory Koedel
Li Tan

WORKING PAPER 179 • September 2018

Accounting for Student Disadvantage in Value-Added Models

Eric Parsons
University of Missouri

Cory Koedel
University of Missouri/CALDER

Li Tan
University of Missouri

Contents

Acknowledgments

Koedel is in the department of economics and Truman School of Public Affairs, and Parsons and Tan are in the department of economics, at the University of Missouri, Columbia. The authors gratefully acknowledge financial support from the University of Missouri Research Board and CALDER and thank Roddy Theobald and conference participants at AEFP 2017 for useful comments. The views expressed here are those of the authors and should not be attributed to the authors' institutions or the funders. Any and all errors are attributable to the authors.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A150137 to American Institutes for Research.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

Accounting for Student Disadvantage in Value-Added Models
Eric Parsons, Cory Koedel, Li Tan
CALDER Working Paper No. 179
September 2018

Abstract

We study the relative performance of two policy relevant value-added models – a one-step fixed effect model and a two-step aggregated residuals model – using a simulated dataset well grounded in the value-added literature. A key feature of our data generating process is that student achievement depends on a continuous measure of economic disadvantage. This is a realistic condition that has implications for model performance because researchers typically have access to only a noisy, binary measure of disadvantage. We find that one- and two-step value-added models perform similarly across a wide range of student and teacher sorting conditions, with the two-step model modestly outperforming the one-step model in conditions that best match observed sorting in real data. A reason for the generally superior performance of the two-step model is that it better handles the use of an error-prone, dichotomous proxy for student disadvantage.

1. Introduction

Value-added models (VAMs) are a commonly-used tool in research and policy applications for measuring how teachers affect student achievement. Recent experimental and quasi-experimental evaluations show that teacher value-added is a forecast-unbiased measure of teacher quality on average, at least in selected locales (Bacher-Hicks, Kane, and Staiger, 2014; Chetty, Freidman, and Rockoff, 2014; Kane, McCaffrey, Miller, and Staiger, 2013; Cullen, Koedel, and Parsons, 2016), and a large literature documents the informational content of value-added more broadly.[1] Numerous states and school districts across the U.S. have implemented systems that incorporate student achievement growth, typically in the form of value-added or a similar metric, into teacher evaluations (Steinberg and Donaldson, 2016).

The contribution of the present study is to empirically examine the relative performance of two value-added models (VAMs): a "one-step VAM" and a "two-step VAM" (or "aggregated residuals VAM"). In addition to being common in research, both modeling structures have been used as policy tools by state and local education agencies in recent years. The policy and political considerations associated with the choice between the two models are covered in Ehlert, Koedel, Parsons, and Podgursky (2014, 2016). This study provides a complementary technical evaluation focused on how the models identify teacher effects and implications for estimation accuracy.

We evaluate one- and two-step VAMs using a flexible, simulated dataset where the data generating process (DGP) is well-grounded in research and reflects realistic sorting conditions. Two aspects of our DGP are particularly important. First, we generate student test scores by calibrating the simulated data to match empirical regularities established by a large body of previous value-added

[1] The literature is too large to list all of the studies here. Notable examples of research linking value-added to other measures of teacher quality include Harris and Sass (2014), Jacob and Lefgren (2008), and Kane, Taylor, Tyler, and Wooten (2011). See Koedel, Mihaly, and Rockoff (2015) for a literature review; recent studies by Chetty et al. (2017) and Rothstein (2017) continue the debate over the quality of value-added measures.

research. A well-known but oft-ignored aspect of education research is that the commonly-used poverty measure, free/reduced-price meal (FRM) status, is coarse and error prone (USDA, 2007; Bass, 2010; Harwell and LeBeau, 2010; Hoffman, 2012; Michelmore and Dynarski, 2017). We build this aspect of real-world evaluations into our simulations by generating student scores using a continuous measure of income parameterized based on Chetty et al. (2014), and then mimicking the proxy approach to controlling for poverty with the binary and noisy FRM indicator. To the best of our knowledge, we are the first to examine the effect of measurement error in controls for student disadvantage on value-added estimation.

The second, complementary aspect of our simulation design is that we create a baseline sorting scenario of students to schools that reflects real world sorting. The DGP governing student sorting is based on data from Census tracts that we map onto elementary-school catchment areas in eight urban and suburban school districts in the Kansas City, Missouri metropolitan area. Income distributions available at the Census-tract level allow us to construct a realistic student sorting condition by income. Our baseline sorting condition is an intuitive reference point for thinking about alternative sorting scenarios.

A key feature differentiating the two VAMs we evaluate is the source of identifying variation used to estimate the control-variable coefficients. In the one-step VAM, identification is achieved entirely by leveraging within-teacher variation, while the two-step VAM leverages both within- and between-teacher variance. A concern with the one-step VAM that has gone unaddressed in previous research is that its reliance on within-teacher variation, combined with the fact that the control variables are measured with error (in particular, FRM status), results in amplified attenuation bias in the control-variable coefficients (e.g., see Ashenfelter and Krueger, 1994; Griliches, 1979). This can lead to the model failing to fully control for student characteristics, thereby causing bias in estimates of teacher value-added. In contrast, attenuation bias in control-variable coefficients is less of a concern

with the two-step VAM because it does not rely solely on within-teacher variance for identification. However, in the presence of non-random sorting of students to teachers, the two-step VAM has the potential to "overcorrect" for student characteristics, generating a different type of bias.

We are not the first to compare these modeling structures in terms of technical performance, although the literature is thin given the research and policy importance of model selection. Chetty et al. (2014) estimate both types of models and find that the two-step approach has marginally lower average forecasting bias, but this result is not focal to their analysis and they provide no explanation. Guarino, Reckase, and Wooldridge (2015) and Guarino, Maxfield, Reckase, Thompson, and Wooldridge (2015) examine these two modeling structures, among others, in related simulation studies. They find that the one-step VAM generally outperforms the two-step VAM, but their data generating process does not allow student covariates to affect achievement beyond lagged test scores. This goes against substantial empirical evidence that student characteristics predict current achievement even conditional on lagged performance.[2] In another simulation study, Zamarro, Engberg, Saavedra, and Steele (2015) allow for student disadvantage to impact student test scores but only consider models that control for student disadvantage perfectly; i.e., they do not incorporate measurement error into their analysis. These previous simulation studies also only consider fairly extreme student-teacher sorting conditions – i.e., random student-teacher assignments or assignments with a strong correspondence between teacher quality and student performance.

We find that one- and two-step VAMs perform similarly over a wide range of estimation conditions. Under the most realistic conditions, estimates from the two-step VAM are more accurate. Only when sorting conditions become extreme does the one-step VAM outperform the two-step

[2] Examples of studies showing that student characteristics predict achievement conditional on lagged scores include Goldhaber, Walch, and Gabele (2013) and Johnson, Lipscomb, and Gill (2015). The conditional importance of student characteristics is likely driven in part by imperfect accounting for measurement error in lagged tests (Lockwood and McCaffrey, 2014) but is an empirical regularity nonetheless.

VAM. The reason the two-step model generally performs better is that it is less adversely affected by the use of an error-prone, dichotomous proxy for student disadvantage. A somewhat surprising result is that concern about overcorrection bias in the two-step VAM is of little practical importance in the most plausible student-teacher sorting scenarios. This is because while control-variable coefficients in the two-step VAM exhibit less attenuation bias, they are still attenuated to some degree, which by itself pushes the model toward favoring teachers of high-income students. The two-step VAM's overcorrection bias works in the opposite direction, and as such leads to an improvement in the accuracy of value-added estimates.

2. Models and Theoretical Rationale

This section provides theoretical background for the models. Portions of the text draw on and extend the conceptual framework in Ehlert et al. (2016).

2.1 One-Step (Fixed Effect) Value-Added Models

The one-step VAM is the most prevalent modeling structure in research studies that estimate teacher value-added (e.g., Aaronson, Barrow, and Sander, 2007; Goldhaber and Hansen, 2010; Hanushek, Kain, O'Brien, and Rivkin, 2005; Rothstein, 2010) and is currently used in some policy applications. The precise set of conditioning variables changes across applications of the model, but the general structure is as follows:

$$Y_{ijst} = \beta_0 + Y_{ijs(t-1)}\beta_1 + \mathbf{X_{ist}}\boldsymbol{\beta_2} + \overline{\mathbf{X}}_{\mathbf{st}}\boldsymbol{\beta_3} + \boldsymbol{\theta_s} + \varepsilon_{ijst} \quad (1)$$

In (1), Y_{ijst} is a test score for student *i* in subject *j* taught by teacher *s* in year *t*, $\mathbf{X_{ist}}$ is a vector of student characteristics for student *i*, $\overline{\mathbf{X}}_{\mathbf{st}}$ is a vector of teacher-average student characteristics in year-*t*, $\boldsymbol{\theta_s}$ is a vector of teacher fixed effects, and ε_{ijst} is the error term. Typical controls in the **X**-vector include student race, gender, FRM eligibility, English-language-learner status, special education status, mobility status, and grade-level.

The identifying condition for equation (1) to recover unbiased, causal estimates of teacher effects is that the control variables are sufficient to capture student-teacher sorting; i.e., teacher assignments to students are conditionally independent. Even if we assume all relevant factors that influence sorting are included in the models conceptually, *how* they are included and measured has implications for whether they under- or over-correct for student circumstances, which can affect the accuracy of value-added estimates. A related concern is about the ability of VAMs to account for prior student performance in the presence of test measurement error (e.g., see Lockwood and McCaffrey, 2014), a point that we return to in more detail below.

By virtue of the one-step estimation, the coefficient vectors $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ in equation (1) are identified using within-teacher variation only. Specifically, the coefficients on individual student characteristics ($\boldsymbol{\beta}_2$) are identified by comparing students who differ along observed dimensions with the same teacher, and the coefficients for the teacher-averaged characteristics ($\boldsymbol{\beta}_3$) are identified using variation in the classroom composition of students taught by the same teacher over time. As a concrete example of the latter, for an elementary teacher with classrooms where the share of FRM-eligible students is 0.80, 0.85, and 0.78 over a three year period, this is the variation used to identify the effect on test scores, rather than (what would typically be much larger) differences across teachers.

While conceptually appealing, relying exclusively on within-teacher variance for identification exacerbates attenuation bias in control-variable coefficients from measurement error (see Ashenfelter and Krueger, 1994; Griliches, 1979). To illustrate, we use a simple bivariate example analogous to controlling for classroom-average characteristics in equation (1). In the example, we have a single regressor, $\overline{Z}_{st}$, observed for each unit, *s*, over two time periods, *t*, where $\overline{Z}_{st} = \overline{Z}^{*}_{st} + \xi_{st}$. $\overline{Z}^{*}_{st}$ is the true value and ξ_{st} is classical measurement error with $\text{var}(\xi_{st}) = \sigma^2_{\xi}$ for t = 1,2 and $\text{cov}(\xi_{s1}, \xi_{s2}) = 0.$

Consider estimating ordinary least squares (OLS) and fixed-effects (FE) coefficients, denoted by δ, from the following regressions:

$$Y_{ist} = \overline{Z}_{st}\delta_{OLS} + u_{ist} \quad (2)$$

$$Y_{ist} = \overline{Z}_{st}\delta_{FE} + \psi_s + e_{ist} \quad (3)$$

The FE coefficient in the presence of classical measurement error in $\overline{Z}_{st}$ is attenuated more than the OLS coefficient, as can be seen by the following formulas taken from Ashenfelter and Krueger (1994):

$$p\lim\hat{\delta}_{OLS} = \delta_{OLS}\left(1 - \frac{\text{var}(\xi)}{\text{var}(\xi) + \text{var}(\overline{Z}^*)}\right) \quad (4)$$

$$p\lim\hat{\delta}_{FE} = \delta_{FE}\left(1 - \frac{\text{var}(\xi)}{\left[\left(\text{var}(\xi) + \text{var}(\overline{Z}^*)\right)(1-\rho)\right]}\right) \quad (5)$$

Above, δ_{OLS} and δ_{FE} represent population regression coefficients in the absence of measurement error. $\text{var}(\overline{Z}^*)$ is the variance of the true underlying values and $\text{var}(\xi)$ is the measurement error variance, with $\text{var}(\overline{Z}) = \text{var}(\overline{Z}^*) + \text{var}(\xi)$ (with classical measurement error, $\text{cov}(\overline{Z}^*, \xi) = 0$). In equation (5), ρ is the correlation between $\overline{Z}_{s1}$ and $\overline{Z}_{s2}$, which will be large and positive in many contexts, including ours (i.e., the average characteristics of students assigned to a teacher in consecutive years are likely highly correlated). This implies potentially significant attenuation bias in the FE estimator relative to OLS. In fact, under the right conditions, bias from measurement error could even lead to $\hat{\delta}_{FE}$ being estimated with the wrong sign. For example, if the error-variance share is 0.10 then ρ above 0.90 would produce a negative value for $\left(1 - \frac{\text{var}(\xi)}{\left[\left(\text{var}(\xi) + \text{var}(\overline{Z}^*)\right)(1-\rho)\right]}\right)$. This scenario is not implausible – for example, in Missouri, the year-to-year correlation in the school-level percentage of FRM-eligible students is 0.95, and it is reasonable to believe that similar conditions may be met in other settings. When we estimate one- and two-step VAMs at the school level using data

from Missouri, the coefficient on the FRM eligibility share is negative and significant using the two-step VAM but positive and significant using the one-step VAM (results suppressed for brevity).[3]

In summary, the sole reliance on within-teacher variance to identify $\boldsymbol{\beta_2}$ and $\boldsymbol{\beta_3}$ will cause amplified attenuation bias in equation (1). The implication is that the model will "undercorrect" for student circumstances and bias estimates of teacher value-added in favor of teachers who teach more advantaged students in more advantaged environments. The problem will be most severe when student characteristics, and in particular teacher-averaged student characteristics, are important predictors of student outcomes and measured with error. Student FRM eligibility – the focal control variable in our study – has been shown to be particularly coarse and error-prone (USDA, 2007; Bass, 2010; Harwell and LeBeau, 2010; Hoffman, 2012; Michelmore and Dynarski, 2017).

2.2 Two-Step (Aggregated Residuals) Value-Added Models

The two-step analog to the one-step VAM shown in equation (1) is:

$$Y_{ijst} = \gamma_0 + Y_{ijs(t-1)}\gamma_1 + \mathbf{X_{ist}}\boldsymbol{\gamma_2} + \bar{\mathbf{X}}_{\mathbf{st}}\boldsymbol{\gamma_3} + \eta_{ijst} \quad (6)$$

$$\eta_{ijst} = \boldsymbol{\tau_s} + \zeta_{ijst} \quad (7)$$

The variables in equation (6) are as defined above; $\boldsymbol{\tau_s}$ in equation (7) is the vector of teacher effects analogous to $\boldsymbol{\theta_s}$ in equation (1).

The key distinguishing feature of the two-step VAM is that it partials out differences in test-score performance between students with different characteristics, and in different schooling environments, *before* estimating the teacher effects. The two-step model has been used to estimate teacher effects in recent high-profile research studies (Chetty et al., 2014; Kane et al., 2013) and also in policy applications.

[3] The intuition conveyed here is also informative for thinking about measurement error in the FRM indicator for individual students, although the structure of measurement error is not the same owing to the binary classifications.

Via the two-step estimation, the control-variable coefficients are not subject to the same degree of attenuation bias as in the one-step VAM (although they are not immune entirely, per equation (4)). The tradeoff is that due to the sequential estimation, the two-step VAM attributes all differences between students along the measured dimensions in equation (6) to those characteristics. Most pressing for the present application is that differences in teacher quality may align with student characteristics, in which case they are purged from the residuals prior to estimating equation (7). The implication is that the two-step VAM will "overcorrect" for student characteristics, leading to biased estimates of teacher value-added.

2.3 Empirical Bayes Models

We also briefly discuss Empirical Bayes (EB) models, which have been evaluated in other recent simulation studies (Guarino, Reckase, and Wooldridge, 2015; Guarino et al., 2015; Zamarro et al., 2015), often in comparison to variants of the one and two-step VAMs. Empirically, EB estimates are a middle ground between estimates from the one- and two-step VAMs. To see why, note that EB estimates can be calculated via generalized least squares by partially demeaning the control variables, as shown in this EB analog to equation (1):

$$Y_{ijst} - \kappa\overline{Y}_{ijt} = \beta_0 + (Y_{ijs(t-1)} - \kappa\overline{Y}_{ij(t-1)})\beta_1 + (\mathbf{X_{ist}} - \kappa\overline{\mathbf{X}}_{\mathbf{it}})\boldsymbol{\beta_2} + (\overline{\mathbf{X}}_{\mathbf{st}} - \kappa\overline{\overline{\mathbf{X}}}_{\mathbf{t}})\boldsymbol{\beta_3} + (\upsilon_{ijst} - \kappa\overline{\upsilon}_{ijt}) \qquad (8)$$

where $\upsilon_{ijst} = \psi_s + \omega_{ijst}$ is a composite error consisting of a teacher random effect, ψ_s, and a residual error term. The means that are subtracted from each term in equation (8) are calculated at the teacher-level, and the weighting factor, κ, is a model parameter that falls between 0 and 1. κ is a function of the variance of the teacher random effects, the variance of the error term, and the number of observations per teacher (n_s), with $\lim_{n_s \to \infty} \kappa = 1$ (Wooldridge, 2000).

The above transformation is insightful for comparing the EB model to the one and two-step VAMs. Specifically, if $\kappa = 0$ then equation (8) is equivalent to the first-step of the two-step model

(equation (6)), while if $\kappa = 1$ then equation (8) is equivalent to the teacher fixed effects model in equation (1). Thus, the one- and two-step models bound the EB model. In results omitted for brevity, we confirm this intuition by showing that output for the EB model consistently falls in-between the output from one- and two-step VAMs, as expected. Given this intuitive result, we focus the analysis below on the one- and two-step VAMs, acknowledging that EB estimates are an in-between case.

3. Simulation Details

3.1 Overview

Table 1 provides detailed documentation of our simulation structure. We begin with a baseline condition where students sort to schools by family income as indicated by the overlap of Census tracts and school catchment areas (see below for details), teachers are randomly assigned to schools, and students are randomly assigned to teachers within schools. We set targets for key coefficient estimates from the one- and two-step VAMs in this baseline condition, where the targets are taken from the value-added literature. The target values are shown in the "Targets" column of Table 1, with key references listed in the "Sources" column. The "Final Baseline Values" columns show the final parameter values used in the DGP and estimates taken from regressions based on our simulated data. The calibration process is iterative and complex given the many dimensions that we target per Table 1. In essence, it can be summarized as follows: we reverse-engineer a DGP by producing a simulated dataset that, when put into the standard regression frameworks, produces estimates consistent with what researchers have found in analyses of real data along numerous dimensions.[4]

Before getting into the details of how we construct the DGP, it is instructive to briefly touch on the value of our simulation design. Although simulation-based studies have obvious limitations

[4] Table 1 is informative about the key aspects of our simulation design but does not provide the entire parameterization of the simulation for presentational reasons. Appendix G provides supplementary material that, when combined with the information in Table 1 and files available online from the authors, including our baseline simulation program, can be used to fully reconstruct the simulated data environment.

and no simulated data environment can hope to capture perfectly all aspects of a real data environment, the questions we explore can best be answered with a simulation study. The reason is that with any real dataset measurement error in the non-test-score control variables is unknown. Our control of the DGP ensures that measurement error in our data is properly understood, which is critical given our focus on modeling the achievement returns to family income. Our simulations also facilitate a straightforward expansion of results to consider hypothetical changes to key evaluation conditions via adjustments to the DGP.

Table 1. Parameter and Coefficient Targets and Values for the Data Generating Process (DGP).

	Equation	Description	Targets	Final Baseline Values		Sources
				One-Step	Two-Step	
λ_1	(8), DGP	Teacher effect decay (one-year lag)	0.2-0.5	0.35	0.35	Lockwood et al. (2007), Kane and Staiger (2008), Jacob et al. (2010), Chetty et al. (2014b)
σ_Q	(9), DGP	SD of time-invariant TQ distribution	0.15	0.15	0.15	Winters and Cowen (2013), Koedel et al. (2015)
σ_V	(9), DGP	SD of time-varying TQ distribution	0.125	0.125	0.125	Winters and Cowen (2013), Koedel et al. (2015)
β_1	(11), REG	VAM lagged exam score coefficient	0.7-0.8	0.77	0.78	Koedel et al. (2015)
β_2^F	(11), REG	VAM FRM indicator coefficient	(-0.1)-(-0.15)	-0.15	-0.14	Ehlert et al. (2016)
$\beta_2^{\bar{F}}$	(11), REG	VAM classroom agg FRM coefficient	0.1 (one-step) -0.2 (two-step)	-0.09	-0.30	Ehlert et al. (2016)
β_2^I	(12), REG	VAM true-income coefficient	0.04-0.07	0.07	0.07	Chetty et al. (2014a) & personal correspondence
$\beta_2^{\bar{I}}$	(12), REG	VAM true aggregate income coefficient	0.02	0.03	0.02	Chetty et al. (2014a) & personal correspondence
ρ^θ	-	Year-to-year teacher effect correlation	0.4-0.6	0.53	0.50	McCaffrey, Sass, Lockwood, & Mihaly (2009), Koedel et al. (2015)
F^{FP}	-	FRM Indicator False Positive Rate	10%-50%	20%	20%	USDA (2007), Bass (2010), Harwell and LeBeau (2010), Hoffman (2012)

Notes: In the *Equation* column, DGP indicates a parameter from the data generating process, and REG indicates a targeted coefficient from VAM regressions. The values in the *Targets* column are from the listed sources, while the *Final Baseline Values* columns report the simulation parameters or output values, listed separately for the one-step and two-step VAMs. $\beta_2^{\bar{F}}$ and $\beta_2^{\bar{I}}$ are taken from versions of equations (11) and (12), respectively, that are extended to include the classroom aggregate variables. The target values for β_2^I and $\beta_2^{\bar{I}}$ are taken from models where income is measured in \$10,000 increments. ρ^θ is calculated using teacher effects estimated on non-overlapping student data (a second set of 60 students – three cohorts of 20 students each – is generated for each teacher to facilitate the comparison). This table provides most of the information needed to fully parameterize our simulations; the rest is available in Appendix G and in the baseline simulation program available online from the authors.

3.2 Generating Student Test Scores

We specify the DGP for a student test score at time *t* as a function of fixed student ability, past and present teacher quality, test measurement error, and a single control variable – household income of the student – that is included at both the individual level and aggregated to the classroom level to allow for classroom environment effects. Chetty et al. (2014), who gained access to income data from the Internal Revenue Service (IRS), provide evidence on the relationship between family income and achievement growth within the value-added framework. Using the IRS data, they estimate that a $10,000 increase in parental income is conditionally associated with a 0.065 standard deviation increase in the grade-8 test score for individual students (averaged across math and reading; see their online appendix D).[5] Based on this estimate, our DGP allows student test scores to be affected linearly by income via a continuous underlying variable (we relax the linearity assumption in Appendix A; our results are substantively unaffected).

A test score for student *i* in year *t* in our simulated data is constructed as follows:

$$Y_{it} = \alpha_i + \theta_{it} + \theta_{i(t-1)}\lambda_1 + I_i\lambda_2 + \overline{I}_{it}\lambda_3 + (\zeta_{1it} + \zeta_{2it}) \quad (9)$$

Equation (9) is not a regression; it is a data generating process. In the equation, α_i is fixed student ability, θ_{it} is the quality of the teacher assigned to student *i* in time *t*, λ_1 is a decay parameter for the lagged teacher effect, I_i is a continuous measure of student *i*'s family income, which we treat as fixed for each student and draw from a distribution based on U.S. Census tracts, and $\overline{I}_{it}$ is the average income of students in student *i*'s classroom in time *t*. The last two terms in parentheses, ζ_{1it} and ζ_{2it}, are error terms that represent test measurement error and residual error in the model (i.e. deviations

[5] We are not aware of other studies that provide similar estimates, which is likely due to a lack of income data beyond the FRM-eligibility proxy. However, even using the imperfect proxy, researchers have long identified important differences in educational outcomes by income. For example, using data from multiple districts across the United States, McCall, Houser, Cronin, Kingsbury, and Houser (2006) show that high-poverty students, as measured by FRM eligibility, have lower test scores and lower test-score growth relative to their low-poverty peers.

from the error-free scores for all other reasons, as discussed in Boyd et al., 2013), respectively. ζ_{2it} is specified to have homoscedastic variance, while ζ_{1it} is heteroskedastic, reflecting the fact that standardized tests are typically designed to be more precise measures of achievement in the middle of the ability distribution (Koedel, Leatherman, and Parsons, 2013; Lockwood and McCaffrey, 2014; Stacey, Guarino, and Wooldridge, 2016). Specifically, ζ_{1it} is distributed as $N(0,\sigma_{1it})$, where σ_{1it} is a u-shaped function of student i's error-free exam score $(\tilde{Y}_{it} = \alpha_i + \theta_{it} + \theta_{i(t-1)}\lambda_1 + I_i\lambda_2 + \overline{I}_{it}\lambda_3)$. The functional form used to determine σ_{1it} is specified based on published conditional standard error of measurement (CSEM) data taken from the Missouri statewide exam.

Parameterized following Winters and Cowen (2013), θ_{st} consists of time-invariant (Q_s) and time-varying (V_{st}) teacher quality components. For teacher *s* we parameterize quality in time *t* as:

$$\theta_{st} = Q_s + V_{st}. \tag{10}$$

Equation (10) is also a data generating process. The time-invariant and time-varying components for each teacher are parameterized independently and distributed normal with mean zero and standard deviations $\sigma_Q = 0.15$ and $\sigma_V = 0.125$, respectively.

The other DGP parameters are adjusted to achieve the target year-to-year correlation of estimated teacher effects within teachers, ρ^{θ}, of 0.40-0.60 per Table 1. Although this year-to-year correlation target may initially seem high, it is appropriate given the models we estimate. Studies finding lower year-to-year correlations of value-added have typically employed specifications that include school and student fixed effects, which research demonstrates add noise but provide little benefit in terms of bias reduction (Koedel et al., 2015).

To arrive at parameter values for λ_2 and λ_3 in equation (9), we work backward from estimates in the literature. First, income is accounted for in standard models using the FRM indicator. Denoting the indicator as F, we apply the following data generating process to produce student FRM status:

$$F_i = \begin{cases} 1 \text{ if } I_i + v_i \leq f(s_i) \\ 0 \text{ otherwise} \end{cases} \tag{11}$$

where I_i indicates true income per above, v_i is measurement error in observed income, and $f(s_i)$ is the income threshold for FRM eligibility, which is a function of student i's family size, s_i.[6] v_i is specified as normally distributed with mean zero and a variance set so that the errors in the resulting FRM eligibility indicator align with the targeted error rate shown in Table 1. Students are flagged as FRM eligible if their observed income (true income plus error) falls below the FRM threshold value.

Estimates from variants of the following regression using real data on FRM status are widely available:

$$Y_{it} = Y_{i(t-1)}\beta_1 + F_{it}\beta_2^F + X_{it}\beta_3 + \varepsilon_{it} \tag{12}$$

After appropriately accounting for measurement error in F per the parameterization in Table 1, simultaneously parameterizing student ability, α_i, and test-measurement error such that the relationship between contemporaneous and lagged test scores for individual students in our simulated data matches commonly-available estimates (also per Table 1), and allowing for teacher effects, we parameterize λ_2 in equation (9) using an iterative process so that we obtain a value of β_2^F in our version of equation (12) that is consistent with the target value reported in Table 1.

[6] s_i is drawn from a school catchment area distribution constructed using Census tract data, similarly to family income.

Next, we estimate the following parallel version of the regression in equation (12) with the simulated data, where again the distribution of the continuous income variable is determined by US Census data:

$$Y_{it} = Y_{i(t-1)}\beta_1 + I_{it}\beta_2^I + X_{it}\beta_3 + \varepsilon_{it}. \quad (13)$$

Equation (13) replaces the noisy, coarse proxy for income in Equation (12) with the true income value. In addition to obtaining a proper estimate of β_2^F in equation (12), our parameterization of λ_2 in equation (9) should also produce an estimate of β_2^I from equation (13) in line with what is reported in Chetty et al. (2014), which it does. This suggests that the parameterizations of the DGP in equations (9), (10), and (11) are properly capturing the complex relationships between true income, measured FRM status, and test scores (including measurement error).

The aggregated income parameter in equation (9), λ_3, is backward-induced using a similar strategy (the iterative process of obtaining values for λ_2 and λ_3 occurs simultaneously). In the end, Table 1 shows that our simulated dataset fits empirical regularities established by the value-added literature quite well.

We use the above-described DGP to generate data for three cohorts of 12,000 students each. Six hundred teachers are also simulated, producing a student/teacher ratio of 20:1 per cohort. For simplicity we keep all teachers in the data in all three years so we observe three classrooms per teacher. Multiple student cohorts are required because without multiple cohorts, the schooling environment controls would not change within teachers and thus could not be included in the one-step VAM.

3.3 Student and Teacher Sorting Scenarios

The other important aspect of the simulation design is student and teacher sorting. While it is useful to assess model performance across a wide range of hypothetical sorting conditions, which is

straightforward, it is more difficult to construct a useful real-world scenario. However, doing so is necessary to be informative about the relevant tradeoffs across modeling structures.

As noted previously, our approach to constructing a realistic sorting scenario of students to schools is to overlay Census tract maps with elementary-school catchment areas. We manually collected elementary school catchment area maps for eight school districts, consisting of 95 urban and suburban elementary schools, and overlaid them onto a Census tract map of Jackson County, Missouri. This allows us to identify the share of each tract that falls within each school catchment area. Census data (from the 2008-2012 American Community Survey) provide income distributions for each tract, from which the income variables for the DGP described by equation (9) can be drawn for individual students. This allows for realistic income distributions at each of the 95 schools. We then resample five schools at random to create a 100-school sample. Our process ignores private school and other non-resident-based enrollment, but corrections can be made to account for non-resident enrollment patterns by comparing the resulting FRM eligibility percentages for schools from the simulation to those published by the state education agency. Results from an analysis that accounts for non-resident student enrollment, shown in Appendix B, are very similar to what we report in the main text.

We draw student ability from the same distribution for all students so that differences in average achievement across schools are driven by income differences. The 600 simulated teachers are assigned to the 100 simulated schools, with 6 teachers assigned to each school. We consider scenarios where teachers are randomly assigned to schools and scenarios where there is a positive correlation between teacher quality and school-average household income. The latter scenarios are particularly important for comparing the efficacy of the one- and two-step VAMs because the two-step VAM will misattribute teacher quality differences that align with student characteristics to the student characteristics.

A final issue is the sorting of students to teachers within schools. Research indicates that systematic within-school sorting is limited (e.g., Clotfelter, Ladd, and Vigdor, 2006; Isenberg et al., 2016). Moreover, mechanically, differences in student-teacher sorting by income driven by cross- and within-school sorting are analytically indistinguishable in our simulations and VAMs.[7] Therefore, for presentational convenience we initially focus on teacher sorting at the school level and randomly assign students to teachers within schools. We subsequently consider scenarios that allow for within-school sorting, including sorting along dimensions other than income.

3.4 Assessing VAM Performance

We evaluate the accuracy of estimated value-added using two summary measures: (a) the correlation with true values and (b) the mean squared error (MSE). In addition, we also consider whether teachers serving particular types of students (e.g., disproportionately low-income) are systematically affected by the use of alternative modeling structures. This gets at a key policy question regarding model choice: which types of teachers are more affected, and in which direction, when inaccuracies occur?

4. Results

4.1 Primary Results

Table 2 presents results from our baseline simulations. Teachers are randomly assigned to schools and students are assigned to schools by income based on the Census-tract mapping. Student assignments dictate the degree of within- versus between-school (and thus teacher) variance in income. Each simulation is conducted 250 times, and we report the mean values over the 250 replications.[8] We apply post-estimation shrinkage to all estimates following the procedure in Koedel et al. (2015).

[7] This would not be the case if our VAMs included school fixed effects. However, it is the norm in the modern value-added literature to omit school fixed effects (Koedel et al., 2015).

[8] The replication value of 250 was chosen empirically based on the convergence of model results.

Table 2. Accuracy of the One-Step and Two-Step Value-Added Estimates Compared to True Teacher Quality Values. Baseline Simulation Conditions. 250 Replications.

	True Income		FRM Proxy	
	1-Step	2-Step	1-Step	2-Step
Correlation with True Teacher Quality (Rho)	0.7000	0.7066	0.6609	0.6795
Mean Squared Error (MSE)	0.0176	0.0169	0.0208	0.0186

Notes: See Table 1 for the baseline simulation values.

Row (1) of the table shows correlations between the estimated teacher effects and true values (Q_s), and row (2) shows the MSE. In the first set of columns, we use the continuous income variable directly in the regression without measurement error. The models using true income give a baseline comparison, but they cannot be feasibly estimated in most applications due to the lack of availability of household income data. The second set of columns shows results from the more policy-relevant case where FRM eligibility is used as a noisy proxy, with the baseline error rate from Table 1.

Unsurprisingly, estimated teacher effects from models that use the continuous, perfectly-measured income variable are more accurate than estimates from models that use the FRM indicator. The models perform similarly when the direct income variable is used, but the two-step VAM performs better with the FRM proxy. The superior performance of the two-step VAM when we use the FRM proxy is in line with expectations for two reasons. First, the one-step model suffers from amplified attenuation bias. Second, the random assignment of teachers to schools obviates the key limitation of the two-step model – specifically, it will not overcorrect in the first step because there is no systematic relationship between teacher quality and student characteristics.

Table 3 compares our baseline results (with the 20% FRM error rate per Table 1) to simulations where the FRM misclassification rate is parameterized to 0%, 10%, 30%, and 40%, respectively. For ease of presentation, Table 3 and all subsequent tables follow the same general structure of Table 2 in terms of reporting rho (the correlation between estimated and true teacher quality) and the MSE. Not surprisingly, model performance weakens as the FRM error rate increases.

The reduction in performance is larger for the one-step VAM, which is consistent with the preceding discussion, although the effect in both models is modest. For example, moving from a 20% to 40% misclassification rate reduces the correlation between estimated teacher quality and true values by 0.0200 and 0.0115 for the one- and two-step VAMs, respectively.

Lowering the FRM misclassification rate also provides insight into model performance that is increasingly relevant given the search for better measures of student disadvantage (Michelmore and Dynarski, 2017). Even with no misclassification, forcing the poverty measure to be binary when the poverty effect is not binary weakens model performance. This can be seen by comparing the results in the second set of columns in Table 3 to the results in Table 2 where we include the accurate, continuous measure of student income in the model. Hence, even as better measures of student disadvantage are developed, our results suggest that the one-step model will still suffer from amplified attenuation bias if the typical binary-measurement approach is retained for income status.[9]

Table 3. Accuracy of the One-Step and Two-Step Value-Added Estimates Compared to True Teacher Quality Values. Various FRM Misclassification Rates. 250 Replications.

	Baseline (20% Error)		0% Error		10% Error		30% Error		40% Error	
	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step
Rho	0.6609	0.6795	0.6753	0.6807	0.6712	0.6804	0.6500	0.6757	0.6409	0.6680
MSE	0.0208	0.0186	0.0192	0.0186	0.0197	0.0186	0.0220	0.0189	0.0231	0.0196

Notes: All other parameters aside from the FRM error rate are set to the baseline values reported in Table 1. As in Table 2, Rho indicates the correlation between teachers' value-added estimates and true quality, and MSE is the mean squared error.

Next we investigate the sensitivity of our findings to changes in student-teacher sorting. Table 4 shows results that allow for a positive correlation between teacher quality and student income – i.e., positive student-teacher sorting – from models that use the noisy FRM indicator with the baseline error rate as reported in Table 1. We consider scenarios where the correlation between teacher quality and true student income is 0.1, 0.2, 0.3, and 0.6. To obtain these correlations in our simulation

[9] Use of a categorical variable, as in Michelmore and Dynarski (2017), would reduce but not eliminate this issue.

framework, pairs of teachers are randomly selected and their classrooms swapped, with the switch maintained if it moves the correlation in the desired direction and reversed otherwise. This process continues until the specified target correlation is met.

We can express the correlations between teacher quality and student income in terms of the estimated gaps they create in teacher quality across students who differ by (noisy) FRM status. This is useful for benchmarking because such gaps have been reported many times in the literature using real data. For example, Isenberg et al. (2013) show that on average in 29 school districts, the gap in teacher value-added estimated from a one-step VAM between FRM and non-FRM students is 0.02-0.03 student standard deviations. In a follow-up study, Isenberg et al. (2016) find a smaller gap of roughly 0.005 student standard deviations.[10] Sass et. al (2012) also use a one-step VAM and find gaps in teacher quality by student FRM eligibility of 0.01 to 0.03 student standard deviations in Florida and North Carolina.[11] Goldhaber, Quince, and Theobald (2018), again using a one-step VAM, find gaps in teacher quality between FRM and non-FRM students in North Carolina and Washington on the order of 0.02-0.03 student standard deviations.

Although our analysis suggests that gaps estimated using a one-step VAM will be biased, they are still useful for calibration. In our simulations, correlations between teacher quality and continuous student income of 0.10, 0.20, 0.30, and 0.60, as reported in Table 4, correspond to gaps in teacher quality by (noisy) FRM status of 0.035, 0.038, 0.040, and 0.049, respectively, when teacher quality is estimated using a one-step VAM to match the above studies. Thus, benchmarking the gaps estimated in our simulations against available research (as discussed in the previous paragraph) suggests that the

[10] An explanation for the differing results is that Isenberg et al. (2016) include classroom characteristics in their models while Isenberg et al. (2013) do not. A reason given by the authors is that additional years of data were available for the later study, providing more within-teacher variation to leverage in their one-step VAM (Isenberg et al., 2016).
[11] These authors estimate many models. The range of values reported in the text is for estimates from models comparing students by individual FRM status with partial persistence, student covariates, and un-shrunken value-added. Many of the other models considered by these authors imply even smaller, and sometimes negative, gaps by student FRM status.

scenarios where the correlation between teacher quality and income is in the range of 0.00-0.10 most closely reflect real-world conditions.

Table 4. Accuracy of the One-Step and Two-Step Value-Added Estimates Compared to True Teacher Quality Values. Various Teacher Sorting Scenarios. 250 Replications.

	Baseline (0 Corr)		0.1 Corr		0.2 Corr		0.3 Corr		0.6 Corr	
	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step
Rho	0.6609	0.6795	0.6683	0.6825	0.6780	0.6835	0.6851	0.6825	0.7054	0.6609
MSE	0.0208	0.0186	0.0215	0.0185	0.0201	0.0183	0.0198	0.0183	0.0189	0.0189

Notes: The correlation values represent the correlation between true teacher quality and student income, both aggregated at the school level. All other parameters are set to the baseline values reported in Table 1. As in Table 2, Rho indicates the correlation between teachers' value-added estimates and true quality, and MSE is the mean squared error.

With this context we turn to the results. First, somewhat surprisingly, note that there is a modest *improvement* in the accuracy of value-added estimates from both models when we introduce limited sorting (up to a 0.20 income-quality correlation). Moreover, there is a consistent improvement as the correlation rises through 0.60 for the one-step VAM. The improvement in both models, and particularly the two-step model, when sorting increases may initially seem counterintuitive. However, there is a straightforward explanation: the bias introduced into the models in the positive student-teacher sorting scenarios is positively correlated with true teacher quality and offsets other biases in the models. Put differently, if the true state of the world is that higher quality teachers are sorted to higher income schools, then models that are biased in favor of teachers in high income schools, such as the one-step model, are essentially adjusting the teacher effect estimates toward the truth, albeit in an unintentional and ad hoc manner.

To unpack this explanation further, start with the one-step VAM and the initial condition where the correlation is 0.00 (random teacher assignments). Because the coefficients on individual and aggregate FRM status from the achievement regression suffer from attenuation bias, teachers who by happenstance receive more high-income students appear to be more effective. This is due to the incomplete accounting for student-income effects in the model. However, this bias is uncorrelated

with true teacher quality because teachers are randomly distributed. Teachers who by happenstance get a good income draw are rewarded by the bias, and those who get a bad income draw are harmed, but these gains and losses are unrelated to true quality values.

Next consider increasing the correlation between teacher quality and true student income via the sorting process, as shown in the later columns of Table 4. Attenuation bias in the FRM controls remains an issue – that is, because the attenuated coefficients do not fully capture the value of income, the bias continues to favor teachers of high income students. But now, with positive student-teacher sorting, teachers who have more high-income students are truly better on average. Thus, the sorting bias directionally aligns with the truth. As positive sorting increases, the bias increasingly goes in the same direction as the truth because the best teachers are more and more likely to have the highest income students. On net, the result is that the accuracy of model predictions improves with positive student-teacher sorting. Of course this only works up to a point, but over the range of sorting conditions we consider in Table 4 the net effect of increased sorting bias is improved model performance, at least for the one-step VAM.

As a verification of this mechanism, in Appendix C we show analogous results under conditions with negative sorting – i.e., where higher-quality teachers are on average assigned to lower income students. In this scenario the best teachers now have students who based on their income should perform worse, but the model does not fully account for this, so the bias in their value-added estimates is negative. The converse is true for low-value-added teachers. As predicted, Appendix C shows that the performance of the one-step VAM deteriorates rapidly as negative sorting becomes more severe. The rationale is the opposite of what we see in Table 4. The two-step VAM exhibits a similar but more muted pattern in Appendix C, which is consistent with the above-described differences in how the models work.

Turning to the results for the two-step VAM in Table 4, attenuation bias is still an issue, but much less so because the model leverages between-teacher variance to help identify the parameters in equation (6). Therefore, the "correlated bias" described in the previous paragraph improves accuracy by less as the correlation increases because the initial effect of attenuation bias is smaller. At the same time, overcorrection bias as described in Section 2.2 is increasingly an issue as the correlation between teacher quality and student income rises. This is because differences in teacher quality by income load onto the first-stage parameter estimates in equation (6). At low levels of sorting, increases in positive student-teacher sorting modestly improve accuracy on net, like with the one-step VAM. However, when the correlation reaches the 0.30 level, overcorrection bias becomes more important and the net effect changes direction. This is a manifestation of the model tradeoffs discussed in Section 2.

In summary, the results in Table 4 indicate that over the range of realistic sorting scenarios, the two-step VAM marginally outperforms the one-step VAM in terms of accuracy, suggesting that the attenuation bias issue dominates the overcorrection bias issue. With high levels of sorting, the one-step VAM exhibits superior performance because overcorrection bias in the two-step VAM becomes more problematic.

These results may initially seem at odds with findings from Guarino, Reckase, and Wooldridge (2015) and Guarino et al. (2015), who argue that the one-step VAM produces the most accurate estimates. While there are design differences between these studies and ours that make direct comparisons difficult, we note that they only find that the one-step VAM outperforms the two-step VAM with non-random student-teacher sorting. Moreover, although neither of these papers provides clear metrics documenting the degree of student-teacher sorting in the simulations, the descriptions of the student grouping and teacher assignment procedures found in both papers suggest substantial sorting. This makes their findings most comparable to our high correlation case (0.6), in which our

results are directionally similar. However, we show that the two-step VAM outperforms the one-step VAM in more moderate sorting scenarios that research suggests are more realistic.

It is also noteworthy that Guarino, Reckase, and Wooldridge (2015) and Guarino et al. (2015) use a DGP where student test scores depend only on time-invariant student ability, teacher quality, and a random error term. Their DGP does not incorporate economic disadvantage at the individual or classroom levels. By construction, their setup prevents attenuation bias owing to noisy control variables from affecting model performance, while the pattern of results we show is influenced significantly by the attenuation bias issue.

Next we allow for within-school sorting of students to teachers. The implications of within-school sorting will be similar to the implications of cross-school sorting because the same underlying factors are relevant. A conceptual difference, however, is that it is more plausible that sorting occurs along dimensions other than income within schools. Thus we also consider within-school sorting along the dimensions of fixed ability and lagged test scores (where the latter is inclusive of error).

The results are reported in Table 5. We start with selected scenarios from Table 4 where the cross-school sorting conditions are different – specifically, we use the cases where school-level teacher quality and student income are correlated at the levels of 0.00, 0.20, and 0.60. Results from the main settings without within-school sorting, which match what we show in Table 4, are reported in the first set of columns of Table 5 for ease of comparison. On top of the baseline cross-school sorting conditions indicated by the rows, each column in the table is for a different within-school sorting condition. For example, in vertical-panel-2/horizontal-panel-2, we show results where sorting across schools generates a school-level correlation between teacher quality and student income of 0.20 and, on top of that, within-school sorting of students to teachers generates a 0.10 correlation between student income and teacher quality (measured at the teacher level within schools). We show results for within-school correlations of 0.10 between teacher quality and student income, ability, and lagged

test scores; we also show results from a stronger within-school sorting condition based on lagged test scores – a 0.20 correlation – given the attention that sorting on test scores has received in research (we are not aware of compelling empirical support for this focus in elementary schools, although within-school student sorting is generally a more significant concern in later grades).

The results in Table 5 are broadly consistent with the patterns documented in Table 4. The general themes that the models perform similarly across estimation conditions and, as sorting becomes more severe, the one-step VAM performs relatively better, are apparent.

Table 5. Accuracy of the One-Step and Two-Step Value-Added Estimates Compared to True Teacher Quality Values. Various Within-School Teacher Sorting Scenarios. 250 Replications.

	Within-School Sorting Conditions		Baseline (0 Corr)		0.1 Corr (Income)		0.1 Corr (Ability)		0.1 Corr (Lagged Test)		0.2 Corr (Lagged Test)	
			1-Step	2-Step	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step
Cross-School Sorting Conditions (Table 4)	Baseline (0 Corr)	Rho	0.6609	0.6795	0.6655	0.6822	0.6643	0.6822	0.6620	0.6792	0.6631	0.6791
		MSE	0.0208	0.0186	0.0207	0.0186	0.0207	0.0185	0.0207	0.0185	0.0207	0.0185
	0.2 Corr	Rho	0.6780	0.6835	0.6818	0.6847	0.6807	0.6858	0.6784	0.6820	0.6795	0.6814
		MSE	0.0201	0.0183	0.0201	0.0183	0.0201	0.0183	0.0201	0.0183	0.0201	0.0183
	0.6 Corr	Rho	0.7054	0.6609	0.7084	0.6595	0.7075	0.6628	0.7053	0.6575	0.7063	0.6555
		MSE	0.0189	0.0189	0.0189	0.0191	0.0189	0.0190	0.0189	0.0191	0.0189	0.0191

Notes: The correlation values in each row represent the correlation between true teacher quality and student income, both aggregated at the school level, as reported in Table 4. The correlation values in each column represent additional within-school sorting along the stated dimension (student household income, student fixed ability, and lagged test scores) built on top of the relevant cross-school sorting scenario. All other parameters are set to the baseline values reported in Table 1. As in Table 2, Rho indicates the correlation between teachers' value-added estimates and true quality, and MSE is the mean squared error.

4.2 Extensions

In this section we explore extensions that modify the DGP and estimation conditions and procedures. First, we consider the sensitivity of our findings to modifying the timespan of the data. In research it is not uncommon for studies to use more than three years of data to estimate teacher value-added (e.g., Chetty, Friedman, and Rockoff, 2014; Sass et al., 2012). At the other end of the spectrum, in policy applications estimates of teacher value-added are often based on just one year.

Table 6 shows results from extensions of the simulation that cover seven years of data for each teacher and just one year. Reducing the data to a single year mechanically prevents separate identification of the classroom aggregate coefficients and teacher fixed effects in the one-step model, thus the classroom aggregates must be omitted.[12] We also exclude classroom aggregates from two-step VAM to allow for a straightforward comparison. A caveat to the analysis over the 7-year span is that the "fixed" component of teacher quality is likely to drift some over time (Chetty, Friedman, and Rockoff 2014), particularly for new teachers, which is not built into our DGP or models. Nonetheless, the results in Table 6 permit general insight into comparative model performance as the number of years available for estimation changes.

Data availability impacts model performance in the expected ways. Specifically, performance for both the one- and two-step VAMs improves as we move from one to three to seven years of data. The relative performance of the models is largely unchanged across the scenarios, with the two-step model continuing to outperform the one-step model in the more realistic low- to moderate-sorting cases and the one-step model outperforming the two-step model in the high-sorting cases.

[12] Of course, the same mechanical identification problem applies to the two-step model, but the two-step model makes no attempt at separate identification via the partialing out of all covariates in the first stage.

Table 6. Accuracy of the One-Step and Two-Step Value-Added Estimates Compared to True Teacher Quality Values. Various Number of Years of Student Outcome Data used in Model Estimation. 250 Replications.

	Number of Years of Student Outcome Data used in the Model		Baseline (3 years)		7 Years		1 Year (No Classroom Aggs)	
			1-Step	2-Step	1-Step	2-Step	1-Step	2-Step
Cross-School Sorting Conditions (Table 4)	Baseline (0 Corr)	Rho	0.6609	0.6795	0.6841	0.7053	0.5876	0.5969
		MSE	0.0208	0.0186	0.0204	0.0180	0.0252	0.0215
	0.2 Corr	Rho	0.6780	0.6835	0.6990	0.7075	0.6070	0.6123
		MSE	0.0201	0.0183	0.0198	0.0178	0.0246	0.0209
	0.6 Corr	Rho	0.7054	0.6609	0.7272	0.6866	0.6442	0.6340
		MSE	0.0189	0.0189	0.0185	0.0184	0.0232	0.0197

Notes: The correlation values in each horizontal panel represent the correlation between true teacher quality and student income, both aggregated at the school level, as reported in Table 4. Each pair of columns represents a model estimated using the given number of years of outcome data. All other parameters are set to the baseline values reported in Table 1. As in Table 2, Rho indicates the correlation between teachers' value-added estimates and true quality, and MSE is the mean squared error.

Next, we turn to the issue of test measurement error. The conceptual issues raised thus far share similarities with issues raised in the literature on test measurement error. Specifically, like measurement error in student income, measurement error in lagged test scores can also adversely affect estimates of teacher value-added by reducing the efficacy of the lagged-achievement control. Unlike measurement error in income, information about test measurement error is often available from test publishers, at least for the portion attributable to the test itself. In our framework this portion of the error is denoted by ζ_{1it}. Most research studies do not make adjustments to address test measurement error, but adjustments are often made in policy applications of VAMs (e.g., Isenberg and Walsh, 2014). In Table 7 we replicate selected results from above after implementing a feasible method of moments (FMOM) correction for test measurement error developed by Lockwood and McCaffrey (2014).[13]

[13] We use the *eivtools* R-package developed by J.R. Lockwood to implement the procedure. The "feasible" descriptor refers to the fact that test measurement error is not known in practice and must be estimated; thus feasible corrections are only possible based on estimates of test measurement error. Lockwood and McCaffrey (2014) develop a procedure

Table 7. Accuracy of the One-Step and Two-Step Value-Added Estimates Compared to True Teacher Quality Values. Feasible Method of Moments Correction. 250 Replications.

			Baseline (No TME Correction)		FMOM	
			1-Step	2-Step	1-Step	2-Step
Cross-School Sorting Conditions (Table 4)	Baseline (0 Corr)	Rho	0.6609	0.6795	0.6675	0.6679
		MSE	0.0208	0.0186	0.0186	0.0186
	0.2 Corr	Rho	0.6780	0.6835	0.6680	0.6659
		MSE	0.0201	0.0183	0.0185	0.0186
	0.6 Corr	Rho	0.7054	0.6609	0.6478	0.6269
		MSE	0.0189	0.0189	0.0190	0.0199

Notes: The correlation values in each horizontal panel represent the correlation between true teacher quality and student income, both aggregated at the school level, as reported in Table 4. Values in the FMOM columns are from models that apply the Lockwood and McCaffrey (2014) feasible method of moments correction to account for test measurement error. All other parameters are set to the baseline values reported in Table 1. As in Table 2, Rho indicates the correlation between teachers' value-added estimates and true quality, and MSE is the mean squared error.

The results from applying the FMOM correction are mixed. First we explain the findings for the one-step VAM, which are comparable to findings in Lockwood and McCaffrey (2014). In the no-sorting and low-sorting scenarios (again, the ones best supported by research), the FMOM correction improves accuracy as measured by the MSE, while the correlation slightly improves in the no-sorting scenario and slightly declines in the low-sorting scenario. The improvement in MSE matches results from Lockwood and McCaffrey (2014). And while these authors do not report correlations between estimates and true values, the lower correlation we report in the low-sorting scenario is predicted by their work. Specifically, the correlations are a function of both the MSE and the variance of value-added. Lockwood and McCaffrey (2014) show that the FMOM procedure reduces the variance of value-added estimates, which all else equal puts downward pressure on the correlation. This explains how the correlation and MSE can decline simultaneously.

The FMOM correction weakens the performance of the one-step VAM in the high-sorting

for estimating test measurement error and show that the feasible approach performs very similarly to the approach based on known test measurement error in simulations.

scenario as measured by both the correlation with true values and the MSE. This result is part of a broad pattern in Table 7 for the one-step VAM in which as positive student-teacher sorting increases, the FMOM procedure becomes less helpful. Similarly to the results presented in Table 4, the reason is that the procedure is removing bias in teacher value-added by dis-attenuating the lagged achievement control; but as student-teacher sorting becomes more positive, the bias that is removed is increasingly aligned with true values and working to offset other biases in the model.

Like with Table 4, Appendix C presents parallel FMOM results with *negative* student-teacher sorting to empirically support the correlated-bias explanation. The models in Appendix C introduce the same level of sorting, but now the bias generated by the sorting is negatively correlated with teachers' true values due the reversal of the sorting process. Thus, in Appendix C, improvements to the model that reduce the influence of sorting bias should lead to more accurate estimates of teacher effects, which is precisely what we find in all scenarios when we apply the FMOM correction. We conclude that the correction is working properly to reduce bias by better capturing the lagged-achievement effect, but depending on the direction of bias and degree of sorting, this can lead to more or less accurate estimates of value-added.

We also briefly touch on estimates from the two-step VAM, which is not studied by Lockwood and McCaffrey (2014). There is no evidence of a benefit from the FMOM procedure for the two-step VAM and if anything, it modestly reduces estimation accuracy. Two factors contribute to this result: (1) again, the loss of correlated bias worsens model performance, like with the one-step VAM, and (2) less benefit accrues from the FMOM procedure in the two-step VAM because the consequences of test measurement error are less severe to begin with. Notably, the two-step VAM is better positioned to leverage information about lagged aggregate test performance to reduce the effect of test measurement error even in the absence of the FMOM correction, as discussed by Lockwood and McCaffrey (2014). We briefly explore this explanation in Appendix E by showing that the FMOM

correction has a more positive influence on model performance when the model does not include classroom aggregates, including lagged aggregate achievement, as conditioning variables.

Finally, in Table 8 we vary the degree of student sorting to schools by true income. This investigation is distinct from the investigation of sorting in Tables 4 and 5 (and extended in Tables 6 and 7). Whereas in Tables 4 and 5 we allow for positive student-teacher sorting, here we maintain the baseline condition that teacher quality and student income are uncorrelated, but we sort students to schools by income to different degrees. In the real world, these types of changes would reflect differences in residential segregation by income. This dimension of model sensitivity is important because more sorting along the income dimension results in less within-teacher variance to be leveraged for identification. Following on the theoretical discussion from Section 2, we hypothesize that the gap in performance between the one- and two-step VAMs should widen, increasingly favoring the two-step VAM, as the within-school (and thus within-teacher) variance share of student income declines.

The column headers in Table 8 report the within- school variance share of the income variable; i.e., 73.4 percent of the variance in income occurs within schools in the baseline sorting condition. We consider three alternative scenarios. The two extreme cases are random assignment and perfect sorting. We also consider an intermediate case where the within-school variance is greatly reduced but remains non-negligible, at roughly 25 percent.[14]

[14] Converting student income to the noisy FRM proxy per above, we can compare the within and between school variance shares in student FRM status in the simulation to observed values in real data. We obtained data from an anonymous set of school districts in a different Midwestern metropolitan area to perform the comparison and find that our baseline simulation scenario is similar (especially when we correct for non-traditional enrollment as in Appendix B) but exhibits somewhat more within-school variance in student FRM status than in the real data.

Table 8. Accuracy of the One-Step and Two-Step Value-Added Estimates Compared to True Teacher Quality Values. Various Student Sorting Scenarios that Modify the Within-School Variance Shares of Student Income. 250 Replications.

	Baseline (within var=73.4%)		Random (within var=99.8%)		Intermediate (within var =25.3%)		Perfect (within var =0.1%)	
	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step
Rho	0.6609	0.6795	0.6839	0.6837	0.4912	0.6323	0.3281	0.5869
MSE	0.0208	0.0186	0.0180	0.0180	0.0583	0.0250	0.1734	0.0332

Notes: The within variance share is the within-school variance share of individual student income. Teachers are assigned to schools at random in all columns. All other parameters are set to the baseline values from Table 1. As in Table 2, Rho indicates the correlation between teachers' value-added estimates and true quality, and MSE is the mean squared error.

With random sorting of students to schools by income, Table 8 shows that estimation accuracy improves in both VAMs, and they perform nearly identically. This is as expected because random sorting provides adequate within-teacher variation to be leveraged by the one-step VAM. Also as expected, the two-step VAM exhibits relative performance gains as we shrink the within-school variance of income. When students are perfectly sorted to schools by income, the performance of both models declines because there is less variation in income, within and between teachers, to be exploited. Although this is not a plausible real-world scenario, it is instructive about the mechanisms driving our findings. The large relative degradation in performance of the one-step VAM is because students will either all be FRM-eligible or ineligible at most schools, based on the true income measure. In fact, within-teacher identifying variation that is not attributable to measurement error is coming entirely from a handful of schools where the average income in the school falls near the threshold income level for FRM eligibility; i.e., the small number of schools that have both FRM eligible and ineligible students.[15] On the whole, the strict sorting of students by true income increases the share of the within-teacher variance in measured FRM status attributable to measurement error.

An issue related to the results in Table 8 is student and teacher mobility. Our simulations do not permit the underlying distributions of income from the catchment areas to change over time, nor

[15] If FRM eligibility were solely a function of household income the variation would always be limited to a single school that falls right at the threshold value. The fact that the FRM eligibility threshold varies by family size allows for some additional within-school variability.

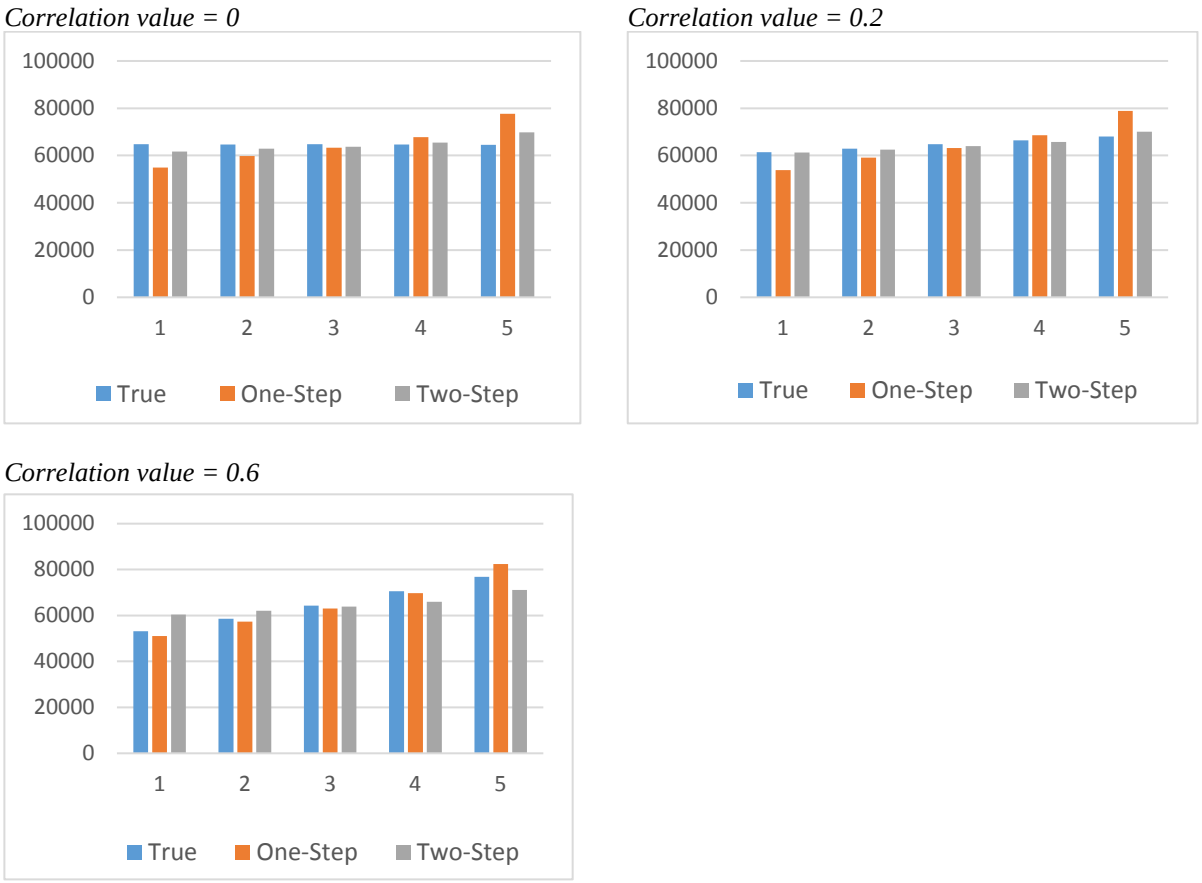
do they allow for teachers to change schools. Both teacher mobility and changing neighborhood demographics – e.g., neighborhood gentrification – would create additional within-teacher variance in student income over time. The clear inference from Table 8 is that the relative performance of the one-step VAM would improve.

4.3 Systematic Bias

Thus far we have shown that the two-step VAM outperforms the one-step VAM in terms of overall accuracy in the most plausible sorting and data-quality conditions, while under extreme conditions the one-step VAM can perform better. That said, the differences in overall accuracy are typically small and if this were the only consideration in selecting a model, a reasonable conclusion is that both models perform similarly. However, average performance metrics like those presented in Tables 2-8 can mask heterogeneity across models in how errors are distributed. For example, in the context of use in accountability systems it is important to understand whether teachers serving students from different backgrounds are positively or negatively impacted by model choice. Even with small differences between models on average, meaningful differences for different types of teachers could exist.

We explore this issue in Figure 1, where we show the average household income of students (in dollars) taught by teachers who differ by quality quintile. Teachers are placed into quality quintiles using three different metrics: the truth (i.e., Q_s as parameterized in the known DGP) and estimates from the one- and two-step VAMs. Lower-numbered quintiles indicate a lower quality ranking. Each chart in Figure 1 presents results from a different student-teacher sorting condition – we show results for correlations between teacher quality and student income of 0.00, 0.20, and 0.60.

Figure 1. Average Classroom Income by Teacher Quality Quintile as Measured by True Teacher Quality and Estimates Produced by the One-Step and Two-Step VAMs. 250 Replications.



Notes: The correlation values represent the correlation between true teacher quality and average student income (i.e., the degree of parameterized student-teacher sorting by income). The vertical axis is measured in dollars of income and the horizontal axis divides teachers by quintile ranking. All parameters other than the correlations are set to the baseline values reported in Table 1.

The first chart in Figure 1 is from the random assignment condition where there is no correlation between teacher quality and student income. Correspondingly, the first bar in each quintile-group (blue), which shows average student income based on teachers' true quintile rankings, reveals no differences in average income across quintiles. The second and third bars reveal that estimates from both VAMs favor teachers who happen to be assigned higher-income students, a relationship that is stronger for the one-step VAM.

These results are as expected for the one-step VAM because amplified attenuation bias leads the model to under-correct for the control variables, falsely attributing a good or bad income draw to the teacher. The practical implication is that even without systematic sorting, teachers who are lucky and get more higher-income students are rewarded. The fact that the two-step VAM also favors teachers of higher-income students is perhaps surprising given that it is purposefully structured to create "proportional" teacher rankings. What drives the result in the two-step VAM is again attenuation bias from measurement error. Even though this bias is reduced in the two-step VAM, it is not entirely mitigated.

The other charts show results where the correlation between teacher quality and student income increases to 0.20 and 0.60. Recall from Table 4 that in terms of overall accuracy, the two-step VAM is marginally more accurate under the first condition and the one-step VAM is more accurate under the second. For the 0.20-correlation case, the two-step VAM, despite overall accuracy levels that are very similar to the one-step VAM in this condition (per Table 4), produces a teacher quality distribution by average income that is much closer to the true distribution because attenuation bias and overcorrection bias are working in opposite directions and largely cancel each other out. When we further strengthen student-teacher sorting in the last chart, the one-step VAM continues to favor teachers of high-income students relative to the truth, but the one-step VAM is more accurate in its classifications per the preceding analysis. In contrast, the two-step model's overcorrection bias now overpowers the influence of attenuation bias, and on net the two-step model favors teachers serving low-income students relative to the rankings based on true quality.

The results in Figure 1 may initially seem to contradict recent research showing that value-added estimates from one- and two-step VAMs are forecast unbiased on average (Bacher-Hicks et al.,

2014; Chetty et al., 2014; Kane et al., 2013).[16] However, there need not be a contradiction for two reasons. First, the income gaps in the top and bottom quintiles, while clearly visible, are not overwhelming, especially when they are mapped to student achievement. Second, what bias is present is concentrated in the tails of the distribution and limited in the middle. Thus, it may not show up strongly in summary measures of bias across the entire teacher distribution. Intuitively, Figure 1 shows that when there is bias from attenuated control-variable coefficients, the effect is more pronounced for teachers of classrooms where student characteristics differ most from the average classroom.

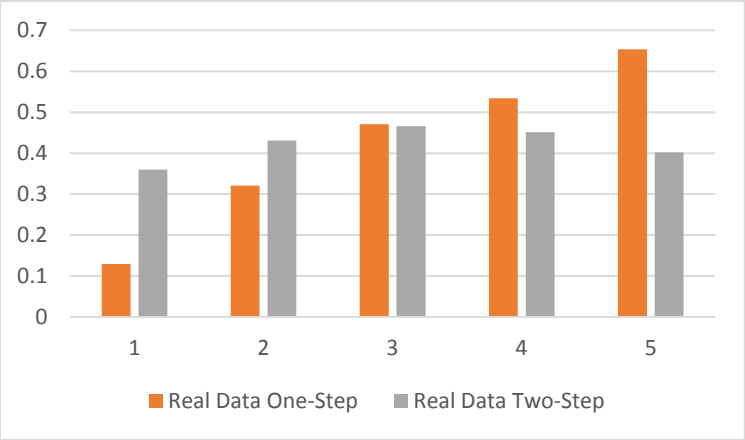
To help contextualize these results we develop a procedure within our simulation framework analogous to the one used by Chetty et al. (2014) to estimate average forecasting bias in value-added estimates. We relegate the details of the procedure to Appendix D, but conceptually the idea is to estimate how much of the variance in student test scores that cannot be explained by observable measures (e.g., student FRM status) can be explained by unobservable measures that we know given our simulation design (e.g., student family income). In Chetty et al., (2014), the analogy to the family income information we have in our simulations is information from IRS tax data.

Consistent with Chetty et al. (2014), we estimate small and statistically insignificant forecasting bias of value-added across the teacher distribution using both VAMs, on average. The bias is nominally larger in the one-step VAM, but in both VAMs the point estimate for the bias is below five percent. These results illustrate that a modest degree of bias in value-added estimates, concentrated among a faction of the sample, can go undetected in tests that focus on average bias across the entire distribution. This is hardly a newsworthy result in the abstract. However, while a small amount of bias may be ignorable if it is distributed evenly, if the bias is concentrated among certain teachers, as in Figure 1, it could be quite important, particularly in policy applications.

[16] The Kane et al. (2013) study provides evidence that is less directly related to our application because it tests for bias from sorting within schools only.

It is also worth noting that the model comparisons in Figure 1 generally match findings obtained from real data. This is illustrated in Figure 2, where we use data from an anonymous set of school districts in a different Midwestern metropolitan area to perform a similar analysis. True student income and true teacher quality are not observed in the real data, so we report results in terms of the shares of students who are ineligible for FRM and estimates of teacher value-added from the one- and two-step VAMs.

Figure 2. Average Share of Students Not Eligible for FRM by Teacher Quality Quintile, as Measured by Estimates from the One-Step and Two-Step VAMs Using Real Data.



Notes: The vertical axis measures the FRM-*ineligible* share to align the figure directionally with its simulation-based analogs in Figure 1. The horizontal axis divides teachers by quintile ranking. This figure is produced using administrative data from an anonymous set of school districts in a different Midwestern metropolitan area.

In summary, the one- and two-step VAMs perform similarly in terms of overall accuracy, and the bias present in estimates from each model is small on average across all teachers. However, a more-nuanced analysis reveals pockets of systematic bias that have different implications for which types of teachers are identified as the most and least effective in the different models. While the research literature has thus far devoted little direct attention to this issue, it is likely a key concern for policymakers and other stakeholders in consequential evaluation systems. It is beyond the scope of this article to delve into which types of errors are most acceptable, or even desirable, for public policy. We refer interested readers to Ehlert et al. (2016, 2014) for such a discussion.

5. Discussion and Conclusion

We use rich simulated data grounded in the value-added literature, and with real-world sorting conditions, to evaluate the performance of one- and two-step VAMs. We capture model performance by correlations of estimated teacher effects with true values and the MSE. These metrics speak to the accuracy of teacher classifications that would occur based on value-added rankings (see Appendix F).

An issue that has been largely overlooked in the value-added literature to date is the presence of control variable coefficients that are attenuated by measurement error. Estimated teacher effects from the one-step VAM are particularly prone to the influence of bias from this type of attenuation because the model isolates within-teacher variation to identify the control-variable coefficients (Ashenfelter and Krueger, 1994; Griliches, 1979). The direct effect of income in our DGP, and our use of a noisy proxy for income in the models, allows us to study the practical importance of this issue within the larger context of model selection, which depends on several analytic tradeoffs. While theoretically the tradeoffs associated with the models are straightforward to understand, the theory is ambiguous about which modeling approach is most accurate empirically.

We find that the two-step VAM produces estimates of teacher quality that are more accurate than estimates from the one-step VAM, albeit modestly, under the most likely conditions in teacher evaluation and research settings. When there is substantial student-teacher sorting the one-step VAM is more accurate, but available research suggests that such extreme sorting is unlikely, at least at the elementary level (Sass et al., 2012; Isenberg et al., 2013; Isenberg et al., 2016; Goldhaber, Quince and Theobald, 2018). In higher grades (i.e., in middle school and even more so in high school) student-teacher sorting may be more of an issue due to increased student tracking, the implications of which merit attention in future research.[17]

[17] The issue of how best to estimate value-added for high school teachers remains unresolved in the literature (Anderson and Harris, 2013; Jackson, 2014; Parsons et al., 2015).

We also find that at low to moderate levels of sorting, both models are modestly biased in favor of teachers in high-income schools. Although the bias is not large enough to show up in traditional tests of average bias across the full distribution, we show that such tests can miss pockets of systematic bias that disproportionately affect small groups of teachers. The tilt of the models in favor of teachers in high-income schools is the result of attenuation bias caused by the use of FRM-eligibility as a noisy proxy for continuous student disadvantage. In short, attenuation bias leads the models to under-account for the effect of family income, and correspondingly, they misattribute income-driven differences in achievement to teachers. There is a conceptually parallel issue with respect to test measurement error (Lockwood and McCaffrey, 2014).

The primary argument made by opponents of the two-step VAM is that it overcorrects for student disadvantage and may hide gaps in teacher quality across students who differ by socioeconomic status. A finding that may surprise some – although *ex post* it should not be surprising given the basic statistics of attenuation bias – is that even estimates from the two-step VAM are biased in favor of teachers of high-income students when sorting conditions are modest. Our results suggest that the intuitive concern about overcorrection bias in the two-step VAM only applies when there is a very high degree of student-teacher sorting, in which case overcorrection bias dominates attenuation bias caused by the noisy FRM proxy variable.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25*(1), 95-135.

Anderson, A. & Harris, D. (2013). *Bias of public sector worker performance monitoring: A structural model and empirical evidence from secondary school teachers.* Unpublished manuscript, Tulane University, New Orleans, LA.

Ashenfelter, O. & Krueger, A. (1994). Estimates of the economic return to schooling from a new sample of twins. *American Economic Review, 84*(5), 1157-1173.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). Validating teacher effect estimates using changes in teacher assignments in Los Angeles. NBER Working Paper No. 20657.

Bass, D. N. (2010). Fraud in the lunchroom? *Education Next, 10*(1), 67-71.

Boyd, D., Lankford, H., Loeb, S., and Wyckoff, J. (2013). Measuring Test Measurement Error: A General Approach. *Journal of Educational and Behavioral Statistics* 38(6), 629-663.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593-2632.

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

Clotfelter, C.T., H.F. Ladd, & J.L. Vigdor. (2006) Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*(4), 778-820.

Cullen, J.B., C. Koedel, & E. Parsons. (2015). The compositional effect of rigorous teacher evaluation on workforce quality. NBER Working Paper No. 22805.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2014). Choosing the right growth measure: Methods should compare similar schools and teachers. *Education Next, 14*(2), 66-71.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2016). Selecting growth measures for use in school evaluation systems: Should proportionality matter? *Educational Policy,* 30(3), 465-500.

Food Research and Action Center (2015). *FRAC Facts: Community Eligibility Provision.* Washington DC. Retrieved from http://frac.org/pdf/community_eligibility_amazing_new_option_schools.pdf

Forsberg, Dan (2015). Changes in Free/Reduced-Priced Lunch as a Measure of Student Poverty. The Governor's Office of Student Achievement. Atlanta, GA. Retrieved from https://gosa.georgia.gov/changes-freereduced-priced-lunch-measure-student-poverty

Goldhaber, D. & Hansen, M. (2010). Using performance on the job to inform tenure decisions. *American Economic Review* 100(2), 250-255.

Goldhaber, D., Walch, J., & Gabele, B. (2013). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy, 1*(1), 28-39.

Goldhaber, D., Quince, V., & Theobald, R. (2018). Has it always been this way? Tracing the evolution of teacher quality gaps in U.S. public schools. *American Educational Research Journal* 55(1), 171-201.

Griliches, Z. (1979). Sibling models and data in economics: Beginnings of a survey. *Journal of Political Economy, 87*(5), S37-S64.

Guarino, C., Reckase, M., & Wooldridge, J. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy, 10*(1), 117-156.

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015). An evaluation of empirical Bayes's estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics, 40*(2), 190-222.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). The market for teacher quality. NBER Working Paper No. 11154.

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

Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as an SES measure in education research. *Educational Researcher, 39*(2), 120-131.

Hoffman, L. (2012). *Free and Reduced-Price Lunch Eligibility Data in EDFacts: A White Paper on Current Status and Potential Changes*. Fairfax, VA: U.S. Department of Education.

Institute for Education Sciences (2015). *Request for applications. Statistical and researcher methodology in education. CFDA number: 84.305D*. Washington, DC.

Isenberg, E., Max, J., Gleason, P., Potamites, L., Santillano, R., Hock, H., & Hansen. M. (2013). Access to effective teaching for disadvantaged students (NCEE 2014-4001). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Isenberg, E., Max, J., Gleason, P., Johnson, M., Deutsch, J., Hansen. M., & Angelo, L. (2016). Do Low-Income Students Have Equal Access to Effective Teachers? Evidence from 26 Districts (NCEE 2017-4007). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Isenberg, E. & Walsh, E. (2014). Measuring school and teacher value added in DC, 2012-2013 School Year: Final Report. Mathematica Policy Research (01.17.2014). Retrieved from http://www.mathematica-mpr.com/~/media/publications/PDFs/education/value-added_DC.pdf

Jackson, C. K. (2014). Teacher quality at the high-school level: The importance of accounting for tracks. *The Journal of Labor Economics, 32*(4), 645-684.

Jacob, B.A., Lefgren, L., & Sims, D.P. (2010). The Persistence of Teacher-Induced Learning Gains. *Journal of Human Resources*, 45(4), 915-943.

Johnson, M., Lipscomb, S. & Gill, B. (2015). Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables. *Journal of Research on Educational Effectiveness*, 8(1), 60-83.

http://www.mathematica-mpr.com/~/media/publications/PDFs/education/value-added_pittsburgh.pdf

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating Measures of Effective Teaching Using Random Assignment. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, Thomas J. & Staiger, D.O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper No. 14607.

Kane, T.J., Taylor, E.S., Tyler, J.H., & Wooten, A.L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources* 46(3), 587-613.

Koedel, C., Leatherman, R., & Parsons, E. (2012). Test measurement error and inference from value-added models. *The B.E. Journal of Economic Analysis & Policy, 12*(1). (Topics)

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

Lockwood, J. R. & McCaffrey, D. F. (2014). Correcting for test score measurement in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics, 39*(1), 22-52.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics, 32*(2), 125-150.

McCaffrey, D.F., Sass, T.R., Lockwood, J.R., and Mihaly, K. (2009). The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy* 4(4), 572-606.

McCall, M.S., Hauser, C., Cronin, J., Kingsbury, G.G., & Houser, R. (2006). Achievement Gaps: An Examination of Differences in Student Achievement and Growth. Northwest Evaluation Association: Policy Report.

Michelmore, K., & Dynarski, S. (2017). The gap within the gap: Using longitudinal data to understand income differences in educational outcomes. *AERA Open* 3(1), 1-18.

Parsons, E., Koedel, C., Podgursky, M., Ehlert, M., & Xiang, P. B. (2015). Incorporating end-of-course exam timing into educational performance evaluations. *Journal of Research on Educational Effectiveness, 8*(1), 130-147.

Raudenbush, S. & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307-335.

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

Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic Review* 107(6), 1656-84.

Sass, T., Hannaway, J., Xu, Z., Figlio, D., & Feng, L. (2012). Comparison of the value Added of teachers in high-poverty schools and teachers in lower poverty schools. *Journal of Urban Economics 72*, 104-122.

Stacy, B., Guarino, C, & Wooldridge, W. (2016). Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve? Working Paper.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy 11*(3), 340-359.

U. S. Department of Agriculture, Food and Nutrition Services, Office of Research and Analysis (2007). *NSLP/SBP Access, Participation, Eligibility, and Certification Study – Erroneous Payments in the NSLP and SBP: Vol. I. Study Findings*. Alexandria, VA: U.S. Department of Agriculture.

Winters, M.A. & Cowen, J.M. (2013). Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies. *Journal of Policy Analysis and Management*, 32(3), 634-654.

Wooldridge, J. (2000). *Introductory Econometrics*. Mason, OH: South-Western Thomson Learning.

Zamarro, G., Engberg, J., Saavedra, J. E., & Steele, J. (2015). Disentangling disadvantage: Can we distinguish good teaching from classroom composition? *Journal of Research on Educational Effectiveness, 8*(1), 84-11.

Appendix A
Nonlinear Income Effects in the Data Generating Process

Our primary DGP allows family income to affect test scores linearly and is parameterized using income effects estimated from a regression in which income enters linearly (per Chetty et al., 2014). However, the true relationship between income and test scores may be nonlinear. A potential source of nonlinearity discussed by Ehlert et al. (2016) and Raudenbush and Willms (1995) is instructional practices that are correlated with schooling context (we avoid a lengthy discussion of this possibility here and instead refer interested readers to these papers). In this appendix we briefly consider the possibility of non-linear classroom environment effects by modifying the DGP as follows:

$$Y_{it} = \alpha_i + \theta_{it} + \theta_{i(t-1)}\lambda_1 + I_i\lambda_2 + \bar{\mathbf{I}}_{\mathbf{it}}\tilde{\boldsymbol{\lambda}}_{\mathbf{3}} + \zeta_{it} \quad \text{(A.1)}$$

Equation (A.1) is analogous to equation (9), but we replace the linear classroom average income control with a set of controls that allow for nonlinear test-score returns. Specifically, we divide classrooms into three groups: (1) the bottom quintile, (2) the middle three quintiles, and (3) the top quintile, by average family income. We then include a vector of indicators for these three groups, multiplied by parameter vector $\tilde{\boldsymbol{\lambda}}_{\mathbf{3}}$, in the DGP in place of the linear measure. It is important to note that the non-linear income effects are still calibrated to the Chetty et al. (2014) results – i.e., the non-linear DGP still produces coefficients in the linear regression model estimated by Chetty et al. that match what these authors report. The other aspects of the DGP are held fixed for this exercise.[18]

Table A.1 shows results that correspond to what we show in Table 4, but with nonlinear income effects in the DGP. For brevity we only report results for selected student-teacher sorting scenarios – correlations between teacher quality and student income of 0.00, 0.20, and 0.60. The results in Table A.1 are substantively similar to what we report in Table 4. Thus, our findings are not qualitatively sensitive to allowing for non-linearity in how income affects student test scores.

Appendix Table A.1. Accuracy of the One-Step and Two-Step Value-Added Estimates Compared to True Teacher Quality Values. Non-Linear Classroom Environment Data Generating Process Specification. Various Teacher Sorting Scenarios. 250 Replications.

	Baseline (0 Corr)		0.2 Corr		0.6 Corr	
	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step
Rho	0.6631	0.6833	0.6808	0.6876	0.7094	0.6655
MSE	0.0205	0.0182	0.0198	0.0179	0.0186	0.0185

Notes: The correlation values represent the overall correlation between true teacher quality and student income, both aggregated at the school level. All other parameters aside from the teacher quality correlation and the specification of the classroom environment effects in the DGP are set to the baseline values reported in Table 1. As in Table 2, Rho indicates the correlation between teachers' value-added estimates and true quality, and MSE is the mean squared error.

[18] In equation A.1, the parameterized effects of being in a bottom- and top-quintile classroom, relative to the middle quintiles, are -0.08 and 0.10, respectively. We also consider other non-linear parameterizations (still subject to the constraint that they preserve the Chetty et al. (2014) result from the linear regression) and obtain similar results.

Appendix B
Accounting for Non-Residential School Enrollment

Our primary simulation design uses Census tract income data that includes all families with school-aged children to define the distributions from which family income is drawn for each school. However, in reality, some students – often higher income students living in urban areas – attend private or other schools where enrollment is not based on residence. Thus, the school-level income distributions are skewed relative to what is observed in public schools in the KCMO metropolitan area. With this in mind, we also perform simulations that use school-level income distributions adjusted for exit to non-residential schools. To make the adjustment, we simulate 100,000 students for each school, divide the simulated students into two groups based on FRM status, and then randomly select the appropriate number of students from each bin to produce a total of 360 students per school (120 students per school cohort) where the FRM percentage matches the actual school-level FRM rates reported by the Missouri Department of Elementary and Secondary Education (DESE). For example, if a school's reported FRM rate is 65% then 234 students (65%) are randomly drawn from the school's FRM bin, while 126 students (35%) are drawn from the school's non-FRM bin. Results using the adjusted income distributions that are analogous to what we show in Table 4 are presented in Table B.1.

The results in Table B.1 are very similar to what we show in the main text. A reason for the similarity is that the within versus between-school variance shares of student income are similar in either case (in our baseline condition in the main paper, roughly 73% of the variance is within schools; with these adjusted distributions, the within-school variance is 64%). Given the high degree of correspondence between the results using the full and adjusted income distributions, we do not emphasize this issue strongly.

Appendix Table B.1. Accuracy of the One-Step and Two-Step Value-Added Estimates Compared to True Teacher Quality Values. School Income Distributions Adjusted to Match DESE Reported School-Level FRM Rates. Various Teacher and Student Sorting Scenarios. 250 Replications.

	Baseline (0 Corr)		0.2 Corr		0.6 Corr	
	1-Step	2-Step	1-Step	2-Step	1-Step	2-Step
Rho	0.6526	0.6688	0.6647	0.6738	0.6829	0.6617
MSE	0.0209	0.0198	0.0206	0.0195	0.0197	0.0197

Notes: The correlation values represent the overall correlation between true teacher quality and student income, both aggregated at the school level. All other parameters aside from the teacher quality correlation and the school-level student income distributions are set to the baseline values reported in Table 1. As in Table 2, Rho indicates the correlation between teachers' value-added estimates and true quality, and MSE is the mean squared error.

Appendix C
Negative Student-Teacher Sorting

Table 4 shows that for both VAMs, as positive student-teacher sorting increases between 0.00 and 0.20, accuracy of the value-added estimates improves. Moreover, this pattern is present for the one-step VAM through the most strict sorting scenario we consider (a 0.60 correlation between student income and teacher quality). In the text we explain that the source of improvement is correlated bias; the bias introduced into the models in the positive student-teacher sorting scenarios is positively correlated with true teacher quality and offsetting other biases in the models. The "overcorrection bias" in the two-step VAM eventually offsets the correlated sorting bias at high levels of sorting. This same phenomenon also explains why the FMOM correction does not improve model performance in the high-sorting cases for either model; i.e., it reduces correlated bias.

To test this explanation empirically we report results from simulations with *negative* student-teacher sorting – i.e., where teacher quality is negatively correlated with student income. With negative sorting, the bias introduced is in the opposite direction of the truth. Thus, bias reduction should improve the accuracy of teachers' value-added estimates. There is no evidence in the literature to support negative sorting as realistic, but these models are instructive about the mechanism driving our findings in the positive-sorting condition. If the correlated-bias explanation is correct, negative sorting should worsen model performance overall. Moreover, it should increase the accuracy gains from employing Lockwood and McCaffrey's FMOM correction.

Table C.1 shows results analogous to Tables 4 and 7 but where we impose negative correlations between teacher quality and student income. For brevity we report selected results from cases where the correlation between teacher quality and student income is set to -0.20 and -0.60. We also repeat the baseline "zero correlation" results for ease of comparison. Consistent with the positively-correlated-bias explanation for our main findings, when there is negative student-teacher sorting both models consistently perform worse as the degree of sorting increases. The performance of the one-step VAM deteriorates more rapidly. In addition, the FMOM correction improves model performance with negative sorting, particularly for the one-step VAM.

Appendix Table C.1. Accuracy of the One-Step and Two-Step Value-Added Estimates Compared to True Teacher Quality Values. Various Negative Teacher Sorting Scenarios. With and Without the Feasible Method of Moments Correction. 250 Replications.

		Baseline (0 Corr)		-0.2 Corr		-0.6 Corr	
		1-Step	2-Step	1-Step	2-Step	1-Step	2-Step
Baseline (No TME Correction)	Rho	0.6609	0.6795	0.6438	0.6713	0.5854	0.6184
	MSE	0.0208	0.0186	0.0212	0.0189	0.0228	0.0208
FMOM	Rho	0.6675	0.6679	0.6652	0.6650	0.6407	0.6250
	MSE	0.0186	0.0186	0.0186	0.0186	0.0193	0.0200

Notes: The correlation values represent the overall correlation between true teacher quality and student income, both aggregated at the school level. Values in the FMOM panel are taken from models that apply the Lockwood and McCaffrey (2014) feasible method of moments correction to account for test measurement error. All other parameters are set to the baseline values reported in Table 1. As in Table 2, Rho indicates the correlation between teachers' value-added estimates and true quality, and MSE is the mean squared error.

Appendix D
Estimating Average Forecast Bias

We estimate the average forecast bias in the value-added estimates used throughout our paper by replicating the procedure used by Chetty et al. (2014) in Table 3 of their study. The idea is to use information typically unobserved by the researcher to estimate the bias from VAMs that exclude such information. In their study the unobserved information comes from IRS tax records; our simulations facilitate a similarly-spirited test where the observed information is the noisy FRM indicator and the unobserved information is true family income, which we know because we control the DGP.

We begin by running the entire simulation three times with a fixed teacher quality vector, capturing the value-added estimates for each teacher from each simulation run, which we denote $\hat{\theta}_{sr}$ for teacher s in simulation run *r*.[19] We then estimate the following jack-knife regression:

$$\hat{\theta}_{s3} = \gamma_1 \hat{\theta}_{s1} + \gamma_2 \hat{\theta}_{s2} + \varepsilon_s. \tag{D.1}$$

The predicted values taken from the estimation of (D.1), $\tilde{\theta}_{s3}$, are saved and used later in the process. We do this for both the one- and two-step VAMs to produce structure-specific jack-knifed estimates.

Following Chetty et al. (2014), the next step is to produce residualized family income values that capture the portion of true income not explained by observed characteristics, which we do using the following regression:

$$I_{ist} = \alpha_0 + \mathbf{X}_{ist}\boldsymbol{\alpha}_1 + \boldsymbol{\tau}_s + \kappa_{ist} \ , \tag{D.2}$$

where I_{ist} is the family income of student *i* taught by teacher *s* in time *t*, $\mathbf{X}_{ist}$ is the vector of observable student characteristics included in the value-added models (in our application, student *i*'s FRM status, classroom FRM share, and prior year exam score), and $\boldsymbol{\tau}_s$ is a vector of teacher fixed effects.[20] Following the estimation of equation (D.2) the residualized income values are calculated as

$$\tilde{I}_{ist} = I_{ist} - \hat{\alpha}_0 - \mathbf{X}_{ist}\hat{\boldsymbol{\alpha}}_1 \ . \tag{D.3}$$

A parallel process is used to produce residualized values of average classroom income, $\tilde{\bar{I}}_{st}$, and student time-*t* test scores, $\tilde{Y}_{ist}$.

We then estimate the following prediction regression:

$$\tilde{Y}_{ist} = \delta_0 + \delta_1 \tilde{I}_{ist} + \delta_2 \tilde{\bar{I}}_{st} + \tau_s + \zeta_{ist} \tag{D.4}$$

in which the portion of student exam scores not explained by observable student characteristics is predicted based on the portion of family income characteristics that are not explained by observable

[19] We also allow a fraction of teachers to randomly change schools each run. This incorporates an appropriate level of persistent teacher sorting along the student income dimension.

[20] The simulation subscript $r = 3$ is suppressed for notational simplicity in all equations from (D.2) onward.

student characteristics. Put differently, we partial out the information present in the observable student characteristics included in the VAMs from both exam scores and family income and then estimate how much of the remaining variation in student exam scores can be explained by the additional information present in the true, unobserved family income values. The predicted values from equation (D.4), $\hat{Y}_{ist} = \hat{\delta}_0 + \hat{\delta}_1 \tilde{I}_{ist} + \hat{\delta}_2 \bar{\tilde{I}}_{st}$, are then used as the outcome variable in the following model:

$$\hat{Y}_{ist} = \beta_0 + \beta_1 \tilde{\theta}_{ist} + \upsilon_{ist} \ . \qquad \text{(D.5)}$$

This equation captures the extent to which variation in student test scores that is not explained by observable student characteristics, but is explained by unobservables (true family income in our case), is correlated with the estimates of teacher quality. As demonstrated in Chetty et al. (2014), the coefficient $\hat{\beta}_1$ in (D.5) is an estimate of the average forecast bias in the teacher effect estimates.

Appendix E
Test Measurement Error Correction when Classroom Aggregates are Omitted from the Models

Table 7 shows that the Feasible Method of Moments (FMOM) procedure improves accuracy in the one-step model (as measured by MSE) in the no- and moderate-sorting scenarios, while it provides no benefit and may slightly reduce accuracy in the two-step model. In the text we posit that the lack of improvement in the two-step model results from (a) the inclusion of classroom aggregates, which should reduce the bias associated with TME as discussed in Lockwood and McCaffrey (2014) and (b) the fact that the coefficients on the aggregate controls in the two-step model are not attenuated as strongly, as discussed in Section 2.

To empirically examine this explanation, in this appendix we report results from models that omit the classroom aggregates from the estimation of both the one and two-step models. Table E.1 reports two sets of results for each model – results estimated with no TME correction and results using Lockwood and McCaffrey's FMOM procedure. We also present our baseline results that include the classroom aggregates and do not apply the FMOM procedure for ease of comparison. The results show that when the classroom aggregates are not included in the models, the FMOM procedure improves the accuracy of both models, with the biggest impact seen in the one-step VAM.

Appendix Table E.1. Accuracy of the One-Step and Two-Step Value-Added Estimates Compared to True Teacher Quality Values. Various Teacher Sorting Scenarios. With and Without the Feasible Method of Moments Correction (Lockwood and McCaffrey, 2014). Classroom Aggregates Omitted. 250 Replications.

			Baseline (With Aggs, No TME Corr)		W/o Aggs, No TME Correction		W/o Aggs, FMOM	
			1-Step	2-Step	1-Step	2-Step	1-Step	2-Step
Cross-School Sorting Conditions (Table 4)	Baseline (0 Corr)	Rho	0.6609	0.6795	0.6484	0.6599	0.6654	0.6657
		MSE	0.0208	0.0186	0.0254	0.0207	0.0189	0.0188
	0.2 Corr	Rho	0.6780	0.6835	0.6690	0.6772	0.6578	0.6590
		MSE	0.0201	0.0183	0.0246	0.0199	0.0191	0.0190
	0.6 Corr	Rho	0.7054	0.6609	0.7072	0.7010	0.6206	0.6203
		MSE	0.0189	0.0189	0.0230	0.0185	0.0202	0.0202

Notes: The correlation values represent the overall correlation between true teacher quality and student income, both aggregated at the school level. Values in the FMOM column are taken from models that apply the Lockwood and McCaffrey (2014) feasible method of moments correction to account for test measurement error. Results in the baseline column replicate values from Table 4 and are taken from models that do not make a test measurement error correction and include classroom aggregates. Classroom aggregates are excluded from all other models. All other parameters are set to the baseline values reported in Table 1. As in Table 2, Rho indicates the correlation between teachers' value-added estimates and true quality, and MSE is the mean squared error.

Appendix F
Teacher Misclassifications in High- and Low-Income Schools

We measure the performance of the VAMs in the main text using two primary metrics – correlation with the true values and MSE. Educator accountability systems in practice typically employ some type of classification system to categorize teachers. The measures we use will be indicative of the degree of misclassification under general conditions. Nonetheless, to explore this contextual issue further we directly examine the performance of the models in terms of their ability to identify high- and low-performing teachers given fixed classification rules. We focus on classifications that identify the bottom and top 10 percent of teachers in the value-added distribution.

Appendix Table F.1 shows the numbers of teachers identified in the bottom and top deciles based on true teacher quality and by estimates from the one- and two-step VAMs. We further split the sample into high-income (top quintile) and low-income (bottom quintile) schools. With random sorting of teachers to schools in the first two rows, of the 60 total teachers in the top decile, the table shows that on average across simulations, 12 are teaching in top-quintile schools and 12 in bottom-quintile schools, as expected. As teachers are sorted to schools by quality in subsequent rows, imbalances emerge.

As predicted by Figure 1, estimates from both the one- and the two-step VAMs imply that there are more top decile teachers in high-income schools and more bottom decile teachers in low-income schools. This is true even in the random sorting case, with the difference most pronounced for the one-step VAM. These results reflect the fact that due to the income effect on student test scores, and our imperfect FRM proxy, a happenstance assignment to a high-income school leads to a higher value-added estimate. Moving down the table, as student sorting increases, true teacher quality is no longer evenly distributed across school types and the one-step VAM improves relative to the two-step VAM as measured by misclassification rates.

Appendix Table F.1. Teacher Misclassification Counts in High- and Low-Income Schools. Various Teacher Sorting Scenarios. 250 Replications.

			Number of Bottom Decile Teachers As Measured By:			Number of Top Decile Teachers As Measured By:		
			True TQ	1-Step	2-Step	True TQ	1-Step	2-Step
Cross-School Sorting Conditions (Table 4)	Baseline (0 Corr)	High-Income	12	6	11	12	24	18
		Low-Income	12	18	12	12	5	9
	0.2 Corr	High-Income	10	5	10	14	26	18
		Low-Income	15	19	12	10	5	9
	0.6 Corr	High-Income	5	3	9	21	29	19
		Low-Income	21	22	13	5	3	8

Notes: *High-income* refers to schools in the top quintile of the student household income distribution, while *low-income* refers to schools in the bottom quintile. Counts represent the average number of teachers who end up in each decile over 250 replications, rounded to the nearest integer value. All other parameters are set to the baseline values reported in Table 1.

Appendix G
Data Generating Process Supplement

This appendix provides information supplementary to Table 1 in order to give the full parameterization of the DGP, with the aim of easing future replications and extensions.

Appendix Table G.1 provides information on several additional parameters of the DGP not directly reported on in Table 1. We have also posted the baseline simulation program used to produce Table 2 online, along with supplementary data files containing the test measurement error heteroscedasticity function and the school income distribution data for our baseline student sorting scenario. Combined, Table 1, Appendix Table G.1, and the files available online provide all that is needed to replicate our baseline simulation.

Appendix Table G.1. Additional Parameters for the DGP.

	Equation	Description	Value
λ_2	(9), DGP	DGP parameter for continuous family income	0.13
λ_3	(9), DGP	DGP parameter for classroom average income	0.02
σ_{ζ_1}	(9), DGP	Mean standard deviation of test measurement error across all students (Heteroskedastic)	0.25
σ_{ζ_2}	(9), DGP	Standard deviation of residual model error, excluding TME (Homoskedastic)	0.25
σ_α	(9), DGP	Standard deviation of student ability	0.54
σ_v	(11), FRM	SD of measurement error in observed income	$30,000

Notes: The individual test measurement errors are heteroskedastic, so we report the mean value of the standard deviation of test measurement error across all students. σ_α is not a free parameter; its value is determined by the other parameters specified in the DGP.