

Test Measurement Error and Inference from Value-Added Models

Cory Koedel
Rebecca Leatherman
Eric Parsons*

January 2012

It is widely known that standardized tests are noisy measures of student learning, but value added models (VAMs) rarely take direct account of measurement error in student test scores. We examine the extent to which modifying VAMs to include information about test measurement error (TME) can improve inference. Our analysis is divided into two parts – one based on simulated data and the other based on administrative micro data from Missouri. In the simulations we control the data generating process, which ensures that we obtain accurate TME metrics with which to modify our value-added models. In the real-data portion of our analysis we use estimates of TME provided by a major test publisher. We find that inference from VAMs is improved by making simple TME adjustments to the models. This is a notable result because the improvement can be had at zero cost.

* University of Missouri, Department of Economics. The authors thank Mark Ehlert for valuable research assistance. The usual disclaimers apply.

I. Introduction

Value-added models (VAMs) are commonly used to evaluate educational interventions in research. Moreover, school districts and state education agencies across the country are increasingly turning to VAMs to measure school and teacher performance, sometimes with high stakes attached.¹ In the case of teachers, the rapid expansion of VAM-based assessments is driven by a large body of research showing that teacher effectiveness varies considerably (for a review of the literature see Hanushek and Rivkin, 2011).² This variability is poorly reflected by traditional teacher evaluations (The New Teacher Project, 2009), which has led to a number of recent studies suggesting that teacher assessments based on student test scores be incorporated into the evaluation process at some level (Boyd et al., 2011; Goldhaber and Hansen, 2010; Hanushek, 2009; Podgursky and Springer, 2007; Staiger and Rockoff, 2010). While concerns about using VAMs for high-stakes decisions in education remain (Briggs and Domingue, 2011; Corcoran, 2010; Hill, 2009; Rothstein, 2010), there is a growing consensus that VAMs can play a productive, if not yet entirely determined, evaluative role in education (Harris, 2011).

The issue of test-based measurement error (TME) has received little direct attention in the research literature on VAMs.³ While it is well understood that standardized tests are noisy measures of student learning, researchers have focused primarily on other concerns related to value-added

¹ Recent “Race to the Top” legislation encourages states to design teacher evaluation systems based on student achievement. Some of the winning proposals attached consequences to VAM-based assessments including tenure denial and tenure revocation. Other federal programs, like the Teacher Incentive Fund, also encourage achievement-based teacher evaluations. In addition, locales from Washington DC to New York City to the state of Missouri are experimenting with VAM-based accountability in a variety of forms.

² Chetty et al. (2011) and Hanushek (2011) link variation in teaching effectiveness to future labor market outcomes for students. Both studies find that high-quality teachers have great economic value.

³ McCaffrey et al. (2003) point this out in a 2003 book that predates much of the VAM literature. But despite the rapid growth of research in this area in recent years, few studies have directly examined the implications of test measurement error for VAMs. A notable exception is Boyd et al. (2008), although their objective is somewhat different than the objective of our study.

modeling.⁴ A likely reason is that TME is typically assumed to be classical; however, even if TME is classical (which is a reasonable assumption, and one that we maintain here), it can still affect inference from VAMs. This is particularly true for VAMs that are designed to evaluate individual teachers because (1) teacher-level sample sizes are inherently small and (2) the share of the variance in student test-score gains attributable to TME is large. Even if TME does not introduce systematic bias into VAMs, its presence may still be problematic.

The contribution of the present study is to examine the extent to which inference from VAMs can be improved by incorporating TME information directly into the models. We focus our analysis on VAMs that are designed to estimate teacher effects.⁵ Beginning in a simulated data environment where we control the data generating process, we incorporate information about TME into VAMs under ideal estimation conditions. We describe the estimation conditions as “ideal” because our knowledge of the data generating process ensures that we obtain accurate TME metrics. That is, we do not need to rely on *estimates* of TME as is typically the case with real data. Our simulation analysis shows that incorporating TME information into VAMs has the potential to meaningfully improve model performance.

Next, we extend our analysis to examine real data from the state of Missouri and readily-available TME metrics. The TME metrics that we consider are called Conditional Standard Errors of Measurement (CSEMs), and can be obtained from test publishers for most large-scale standardized tests nationwide. They are constructed using item response theory and provide estimates of the

⁴ For example, studies that examine the potential for VAM estimates to be biased by non-random student-teacher sorting (Goldhaber and Chaplin 2011; Kane and Staiger, 2008; Kinsler, 2011; Koedel and Betts, 2011; Rothstein, 2010), and studies that explore the general instability of VAM estimates over time (Aaronson, Barrow and Sander, 2007; Kane and Staiger, 2002), are common. Test measurement error contributes to the imprecision and instability of VAM estimates, as we show below, but other factors have received the bulk of the attention in research thus far (e.g., year-to-year sampling variability).

⁵ The importance of TME adjustments to VAMs decreases as sample sizes increase within units. When the units of analysis are schools or school districts, within-unit sample sizes are typically large enough that TME adjustments are of little practical importance (we touch on this point briefly when we present our findings). But this is not the case for teachers.

TME variance in student test scores at all points in the underlying score distribution. Consistent with our simulation results, and the claim that CSEMs provide useful information about TME, we find that the Missouri VAMs perform better when we incorporate the CSEMs directly into the models. We benchmark the improvement in model performance by comparing it to the improvement that would come from increasing within-teacher sample sizes in the data. Modifying a typical VAM to include the CSEM information produces an improvement in model performance similar to what would be observed if the average within-teacher sample size were increased by 11-17 percent.

Our findings support the immediate adaptation of VAMs to incorporate readily-available CSEM data. The gains in model performance that come from incorporating the CSEMs are modest, but can be had at zero cost and are notable given the increased reliance on VAMs for high-stakes teacher assessments nationwide. Our analysis also suggests that even larger improvements in model performance may be possible if better TME metrics can be developed. We draw this conclusion by comparing our findings from the simulations to our findings from the real-data models. Our simulations provide an upper bound of sorts on how much we can hope to gain by incorporating TME information into VAMs, and the benefits in the real-data models do not reach the frontier.

II. Background

Value-Added Models

We focus our analysis on the most-commonly used models for teacher evaluation in current practice.⁶ From an informal review of the various VAMs in use in different locales across the nation, we conclude that the typical VAM takes the following general form:

$$Y_{it} = \beta_0 + Y_{it-1}\beta_1 + X_{it}\beta_2 + T_{it}\theta + \varepsilon_{it} \quad (1)$$

⁶ Previous research discusses the basic modeling assumptions that underlie VAMs. See Harris and Sass (2006) and Todd and Wolpin (2003) for methodological overviews.

In equation (1), Y_{it} is a test score for student i in year t , Y_{it-1} is the lagged score, X_{it} is a vector of student-level covariates, and T_{it} includes indicator variables for students' teacher assignments. Many variants of this general modeling structure can be found, including models that (1) control for multiple measures of lagged performance, (2) control for school- and/or classroom-level aggregated student characteristics, (3) estimate the teacher effects as random instead of fixed, and (4) are estimated in multiple levels. In addition, it is common in the research literature to see models that incorporate multiple levels of fixed effects (i.e., schools, students), although we are not aware of any such models in present policy application.⁷

In the analysis that follows we focus on the basic modeling structure presented in equation (1) and its gainscore analog. The gainscore version of (1) forces $\beta_1 = 1$ and moves the lagged-score term to the left-hand side of the equation. Although the gainscore VAM has some undesirable properties (Andrabi et al., 2011) and is less common in application than the general VAM, it offers a key benefit for our analysis: namely, all of the TME is in the dependent variable.

Test Measurement Error

It is well understood that test scores are noisy measures of student learning, and test publishers have long provided TME information along with student scores. However, in practice this information is rarely incorporated into VAMs.⁸ An important aspect of TME is that it is not uniform across the test-score distribution – scores at the center of the distribution are measured with less error than scores in the tails. The intuition is that standardized tests are well-designed to assess student learning for “targeted” students (who score near the center of the distribution), but

⁷ Another class of models is based on transformed student scores, where the typically-suggested transformation is into percentiles (Ballou, 2009; Betebenner, 2008; Neal and Barlevy, forthcoming). We do not examine these models here, but the general concerns about TME that we raise will apply to transformed student scores as well.

⁸ Recent exceptions include the models being estimated in Washington DC (Isenberg and Hock, 2010) and New York City (Value-Added Research Center, 2010). Of note, however, is that neither of the technical reports associated with these models cites any research that directly assesses the TME issue within the VAM context. Our findings indicate that both models make TME corrections that are inferior to alternatives that would be easy to implement.

not for students whose level of knowledge is not well-aligned with the content of the exam (who score in the tails of the distribution).

Test publishers produce estimates of the TME associated with student scores called conditional standard errors of measurement (CSEMs), which are derived using item response theory.⁹ They are “conditional” in the sense that they are based on each student’s placement in the overall score distribution – each observed score is associated with its own CSEM. Figure 1 shows publisher-provided CSEM data from the Missouri Assessment Program (MAP) exam. The CSEMs are plotted against observed test scores for the 2009 cohort of Missouri fifth graders. The U-shape in the figure reflects the relative precision of test scores in the center of the distribution relative to the tails. The key question that we aim to answer is this: if we adjust VAMs to account for the fact that some student scores are measured with more error than others, can we improve inference?

Our analysis requires growth-based TME metrics because VAMs are models of test-score growth. However, CSEMs are estimated and reported by publishers for test-score levels. We follow Thompson (2008) to convert level-score CSEMs into gainscore CSEMs, which we use as the TME metrics for the bulk of our analysis.¹⁰ To illustrate the construction of the gainscore CSEMs we write student i ’s test score in year t as $S_{it} = A_{it} + \lambda_{it}$; where S_{it} is comprised of a true signal component, A_{it} , and a TME component, λ_{it} . The TME variance in the level score is the variance of λ_{it} . The student’s test-score gain, G_{it} , can be written as $G_{it} = (A_{it} - A_{it-1}) + (\lambda_{it} - \lambda_{it-1})$. Under the maintained assumption that the TME is classical $\text{cov}(\lambda_{it}, \lambda_{it-1}) = 0$, so the TME variance in the gainscore is $\{\text{var}(\lambda_{it} - \lambda_{it-1})\}$.

⁹ See Lord (1980, 1984) and Hambleton et al. (1984) for more information about item response theory (IRT) and its application toward the construction of CSEMs. Although IRT-based CSEMs are what test publishers generally provide, others approaches are available (e.g., see Thorndike, 1951; or for a more-recent example, see Boyd et al., 2008).

¹⁰ Later on, we also consider a slight modification to Thompson’s (2008) approach that produces a different TME metric for application in the lagged-score VAM.

$\lambda_{it}) + \text{var}(\lambda_{it-1})\}$. Noting that the publisher-reported CSEMs estimate the standard error of the level-score TME, it follows from above that the formula for the gainscore CSEM is:

$$CSEM(G_{it}) = \sqrt{CSEM(S_{it})^2 + CSEM(S_{it-1})^2} \quad (2)$$

III. Simulations

We begin our evaluation by constructing a simulated data environment where we control the data generating process. We focus on a simple gainscore model for the simulations. Again, the benefit of the gainscore framework is that all of the TME is in the dependent variable, which allows for a straightforward TME adjustment. Later, when we turn to the real data, we evaluate gainscore and lagged-score VAMs.

Data Construction

We construct each student's gainscore in the simulations as the summation of three components: (1) a student-specific component, α , (2) a TME component, λ , and (3) a teacher-effect component, τ :

$$G_i = \alpha_i + \lambda_i + \tau_i \quad (3)$$

Each component is drawn from an underlying distribution for each student. An important feature of the simulations is that the components contribute appropriately to the total variance in student gainscores (i.e., each component's variance share in the simulations is proportional to its real-world variance share).

We start with the teacher-effect components. Rockoff (2004) reports that teacher effects explain roughly 4 percent of the variation in student test-score levels; therefore, we estimate the variance share of the gainscores attributable to the teacher effects as $\text{var}(\tau) = 0.04 * \frac{\text{var}(S)}{\text{var}(G)}$, where $\text{var}(\tau)$ is the variance of the teacher effects, $\text{var}(S)$ is the variance of student test-score levels, and

$\text{var}(G)$ is the variance of student test-score gains.¹¹ We estimate the test-score variances (in levels and gains) using test data from the 2009 cohort of grade-5 students in Missouri (the gainscore variance is for the gain between grades 4 and 5). The teacher effects are drawn from a normal distribution and explain 9 percent of the total variance in gainscores in the simulated data.

Next we use data from the same cohort of grade-5 Missouri students to determine the TME variance share. We begin by estimating students' gainscore CSEMs as in equation (2), then we take the weighted average of the squared CSEMs across the entire distribution of scores from the state. This provides an estimate of the TME variance in student gainscores, which we divide by $\text{var}(G)$ to obtain the TME variance share. Our calculations indicate that TME explains 50 percent of the total gainscore variance. In constructing students' gainscores we draw from the actual distribution of Missouri gainscore CSEMs, and then rescale the draws by a constant to ensure that the overall TME-variance share is 50 percent in the simulated data.¹²

Finally, the remaining components to the gainscores – the student components – are drawn from a normal distribution weighted so that the student components sweep up the residual gainscore variance (i.e., the variance not explained by teachers or TME). Conceptually then, the variability in the student components contains everything that one might observe in a student score other than TME and teacher effects. This includes student ability, luck, non-schooling inputs, etc. The role for the student-specific components is to characterize the contribution of non-TME and non-teaching factors that influence student scores. The variance share of the student component is approximately 41 percent.

¹¹ This result is also supported by the larger literature on teacher effects (Hanushek and Rivkin, 2011).

¹² The 50-percent figure may initially seem high, but note that the variance in student scores attributable to differences in student ability is much smaller in test-score gains than test-score levels. Conversely, the variance in student scores attributable to TME increases moving from levels to gains (by virtue of each gain being the product of two noisily-estimated level scores). Using a test that is noisier in levels than the MAP test in Missouri and an alternative approach to estimating the TME variance, Boyd et al. (2008) find that an even larger share of the variance in student gainscores is attributable to TME (≈ 84 percent).

After we create the three properly-weighted distributions that represent the student, TME and teacher-effect components, we construct a dataset consisting of 2000 students. The process for assembling the student scores is as follows. First, we take 2000 draws from the student-component distribution. We rank the student components from largest to smallest in absolute value and then assign draws from the TME distribution to the student components so that the largest TME variance metrics are assigned to the student components that are largest in absolute value.¹³ Noting that the TME metrics measure the TME *variance* associated with a given score, not the realized TME, we take a random draw from a normal distribution with the indicated variance to create the realized TME for each student. Finally, we randomly assign the students to teachers, and in doing so we attach the teacher-effect components and complete the observed scores. The random assignment of students to teachers is consistent with the objective of examining TME under ideal estimation conditions, although the student-teacher assignment process is not a crucial feature of our analysis.

We examine the role of TME under several different class-size scenarios in the simulations. Noting that in practice some VAMs evaluate teachers based on as few as five students (e.g., see Isenberg and Hock, 2010), we consider class sizes of 5, 10, 20 and 40.¹⁴ We redraw the student scores, re-attach new TME metrics, and re-assign students to teachers 1,000 times for each class-size scenario. We report our results by averaging our findings across the 1,000 simulations.

Analysis and Results

We estimate the following gainscore model using the simulated data:

$$G_i = \delta + T_i\tau + \varepsilon_i \tag{4}$$

¹³ In reality, the largest CSEMs are associated with the largest and smallest level scores as shown in Figure 1. Assigning the TME metrics in the simulations based on the gainscore magnitudes, therefore, is imperfect; but it is of little consequence in practice because the key issue for the simulations is how TME is distributed across teachers. More specifically, the key condition for the TME adjustments to improve inference from the models is that high- and low-TME students are dispersed across classrooms. This point is illustrated in Appendix A.

¹⁴ We run the simulations using 2000 students in all class-size scenarios. So, for example, when class sizes are set to 5 students we include 400 teachers in the analysis; when class sizes are set to 10 we include 200 teachers; and so on.

In (4), G_i is student i 's gainscore and T_i is a vector of teacher indicator variables. The simulations are designed so that student covariates are irrelevant (i.e., the student components of the test scores do not depend on any covariates, and even if they did, students are assigned to teachers at random). We also abstract conceptually from schools for simplicity, although putting additional structure on the simulations to incorporate schools would not affect our findings qualitatively.¹⁵

Because we know the data generating process we can compare the teacher-effect estimates from the models, $\hat{\tau}$, to the true teacher effects, τ . We estimate the models with the different class sizes as described above and assign equal-sized classes to all teachers. Because all of the TME is in the dependent variable in equation (4), the TME adjustment involves weighting the model by the inverse of the TME metric (i.e., the standard deviation of the TME variance). The intuition behind the weighting is simple: student scores measured with more precision will be assigned higher weights relative to noisier student scores. So, for example, if a teacher has 15 students with relatively precise scores and 5 students with relatively noisy scores, the 15 students will contribute more toward the teacher's estimated effect.

Table 1 shows the simulation results. In the first horizontal panel we report correlations between the actual teacher effects and the estimates from the weighted and unweighted models. In the second and third panels we document how the inaccuracies in the estimated teacher effects correspond to cross-quintile movements in teacher rankings; the second panel shows the share of teachers for whom the estimated effect assigns them to the wrong quintile relative to the true rankings, and the third panel shows the share of "major" errors, which we define as occurring

¹⁵ Most of the variance in student test scores in Missouri occurs within, and not between, schools (also see Kane and Staiger, 2002). Per Figure 1, this ensures that there is considerable dispersion of TME within schools. Only if the within-school variability in TME were small would the division of students into schools become an important issue for our analysis.

whenever a teacher's placement in the distribution is off by more than one quintile (e.g., a 15th percentile teacher in the true distribution whose estimate suggests she belongs in the 43rd percentile).

The first thing to notice in the table is that for all class-size scenarios the TME weighting produces notable improvements in inference from the VAM. The correlations between the estimated teacher effects and the true effects are consistently larger in the TME-weighted models, and in most cases the differences translate to real consequences for teachers' quintile rankings. This is particularly apparent in the the third panel of the table. For example, in the 20-student class size scenario the TME weighting is associated with a three percentage point reduction in the number of teachers for whom a major classification error is made (from 22 to 19 percent).

The benefits of the TME-weighting are similar over the range of class sizes that we consider. In omitted results, we take the class sizes up to 100 students per teacher and confirm that when within-teacher sample sizes are large, the error correction has essentially no effect. As a brief aside, this is why our analysis applies to teacher evaluations more so than school or district evaluations – in most cases schools and districts can be evaluated with large enough sample sizes so that the influence of TME is mostly obviated.

To summarize our simulation findings, they show that TME adjustments to VAMs have the potential to meaningfully improve inference from the models. But the gains in Table 1 come under ideal estimation conditions, where a notable feature of the analysis is that we have access to accurate TME metrics. Next we turn to our real-data analysis.

IV. Missouri Data

The most accessible TME metrics available in practice are CSEMs provided by test publishers. CSEMs are constructed based on item response theory and provide estimates of TME in student scores throughout the score distribution. The extent to which inference from VAMs can be

improved by incorporating the CSEM information is an empirical one, and to the best of our knowledge it has gone previously unaddressed in prior research.

For this portion of our analysis we use statewide administrative panel data from Missouri. The data allow us to link students and teachers at the classroom level and include the typical demographic information about students. We model math scores for grade-5 students in 2008-2009 and 2009-2010. We limit our analysis to students for whom gainscores can be constructed, and to teachers who do not change schools and teach more than 10 students in each year. Basic summary information about the data is provided in Table 2.¹⁶

V. Real-Data Analysis and Results

Analysis

We use the Missouri data to estimate lagged-score VAMs, as described in equation (1), and analogous gainscore VAMs. Our VAMs include controls for student race, gender, free/reduced-price lunch status, language status, mobility status, and whether the student has an individualized education plan (i.e., special education status). We estimate single-year VAMs for reasons that will become clear shortly, which rules out the inclusion of school- and classroom-level aggregates in the models because their effects are not separately identified from the teacher effects. In results omitted for brevity we also estimate two-step models that allow us to bring in these school and classroom-level controls; our findings from the two-step models are very similar to what we report below.

For the gainscore VAM the TME adjustment is the same as in the simulations – we weight the regression by the inverse of the gainscore CSEM. But the proper TME adjustment in the lagged-score VAM is more complicated because some of the TME is on the right-hand side of the equation. Fuller (1987) develops a statistical framework that could be used, in theory, to properly

¹⁶ As with other similar statewide datasets we cannot link all students in Missouri to their teachers. But for the students we can link to teachers, approximately 84 percent of all students, we have a high degree of confidence that the linkages are correct. Details on the process that we use to create the student-teacher links are available upon request.

model the TME in the lagged-score model; however, a concern is that the suggested approach would produce estimates that are highly sensitive to outliers (Fuller, 1987), and may perform poorly with student achievement data for this reason.¹⁷ As an alternative, in the few instances where TME is explicitly acknowledged and modeled in lagged-score VAMs, researchers have made the assumption that the TME variance in the lagged score is constant across the test-score distribution. In this case a straightforward errors-in-variables correction based on Fuller (1987) can be applied (Isenberg and Hock, 2010; Value-Added Research Center, 2010). However, Figure 1 shows that the TME variance is clearly not constant across the distribution, and it is not immediately obvious how the violation of this assumption will affect model performance in these cases.

In the absence of a clearly preferred modeling approach for the lagged-score VAM we consider four different TME adjustments. First, we ignore the TME variance in the current-year score and model the lagged-score TME as being constant across the distribution. This approach is conceptually unappealing given what we know about TME (Figure 1), but it is of interest because it is the only approach of which we are aware that is presently in use (Isenberg and Hock, 2010; Value-Added Research Center, 2010). Second, we maintain the incorrect assumption of constant error variance in the lagged score, and model it as such, but also weight the entire model by the CSEM of the current-year test score (in levels). Third, we ignore the fact that the TME in the lagged score enters on the right-hand-side of the equation entirely, and simply weight the lagged-score VAM by the gainscore CSEM – that is, we treat the TME in the lagged-score VAM the same way that we

¹⁷ In addition to Fuller (1987), other useful references include Brown (1982), Carroll and Gallo (1982) and Zamar (1989). Our read of the literature is that the costs associated with using a complex errors-in-variables model would outweigh the benefits in the VAM context. In addition to the primary concern that outlying data can result in poor performance (Fuller, 1987; Brown, 1982), some suggested approaches produce estimates that are asymptotically biased (Zamar, 1989). Furthermore, much of the work in this area has been performed using Monte Carlo methods, and in actual application it may not be straightforward to determine whether the errors-in-variables correction is successful. An alternative strategy involves using instrumental variables (IV) to address the lagged-score TME, with test scores in other years and/or subjects serving as instruments. But this requires more data and will therefore result in smaller student samples. The data requirements would be particularly problematic with instruments based on second- and third-lagged test scores because entire grade levels of students would no longer have sufficient score histories to be included in the analysis.

treat it in the gainscore VAM. Fourth, we apply modified lagged-score CSEM weights of the following form:

$$CSEM(G_{it}) = \sqrt{CSEM(S_{it})^2 + \hat{\beta}_1^2 * CSEM(S_{it-1})^2} \quad (5)$$

Equation (5) is analogous to equation (2) except that the lagged-score TME variance is downweighted by $\hat{\beta}_1^2$. $\hat{\beta}_1$ is the coefficient on the lagged-score term from the general VAM, estimated by unweighted OLS. While none of the TME adjustments that we consider for the lagged-score VAM are technically correct, the case can be made that no preferable alternative is available. In the absence of a robust strategy for properly modeling TME in the lagged-score VAM, we ask whether available imperfect solutions improve model performance.

An important limitation of this portion of our analysis is that, unlike in the simulations, we do not know the true values of the teacher effects. This makes it difficult to quantify the benefits of the TME adjustments because we cannot compare the estimated teacher effects to the true parameter values. Instead, we evaluate the benefits of the TME adjustments in terms of their influence on the year-to-year stability of the estimated teacher effects. Although teacher performance in the real world is not perfectly stable over time, there is clearly a stable component.¹⁸ We hypothesize that TME-adjusted VAMs will produce estimates of teacher effects that are more stable over time relative to unadjusted models. The mechanism is that the TME adjustments will lower the contribution of noisily estimated student scores in favor of more-precisely estimated scores, which will clarify the stable components of the estimates.¹⁹

¹⁸ This finding has been replicated by numerous studies. Examples include Aaronson et al. (2007), Boyd et al. (2011) and Goldhaber and Hansen (2010).

¹⁹ Note that the “stable” components of the teacher effects could include bias. For example, if a teacher is systematically assigned students who are poised for high growth in consecutive years, the weighting will clarify the consistency of this signal in addition to the consistency of the teacher’s effectiveness. For more on the sorting-bias issue in VAMs see Chetty et al. (2011), Goldhaber and Chaplin (2011), Kane and Staiger (2008), Koedel and Betts (2011) and Rothstein

To facilitate our stability analysis we estimate teacher effects separately for the 2008-2009 and 2009-2010 school years. We compare the year-to-year stability in teachers' estimated effects from the VAMs with different TME adjustments. As noted in the previous section, we restrict our analysis to teachers who teach at least 10 students and are employed at the same school for both years.²⁰

Results

Tables 3 and 4 show our results for the gainscore and lagged-score models, respectively. First, Table 3 shows year-to-year correlations in the teacher-effect estimates, and the stability in teachers' quintile assignments over time, for weighted and unweighted gainscore models. Consistent with the hypothesis that the weighting improves inference, the estimated teacher effects are more stable over time in the weighted models. Specifically, the year-to-year correlation in the estimates is higher, and there are marginally fewer teachers who change quintile rankings (although sometimes we need to take the shares out to three decimal places to see the differences).

Table 4 reports our findings for the lagged-score models. When we apply the simple gainscore-based weights (in columns 4 and 5 of the table) our findings for lagged score models are very similar to what we report for the gainscore model. The CSEM weighting results in modest improvements in the year-to-year correlations in the teacher effects, and in the stability of teachers' quintile assignments over time. This is a notable result – despite the fact that the weighting procedure is technically incorrect in the lagged-score models, the adjustment produces results very similar to what we find in the gainscore model where the weighting procedure *is* technically correct. In contrast, the models do not perform well when we make the errors-in-variables correction under

(2010). The evidence from these studies seems to be converging toward a consensus that while teacher-effect estimates from VAMs are biased, the bias is small.

²⁰ In unreported results we also estimate the models using the subsample of teachers who have more than 3 years of teaching experience because new teachers are likely to have less stable year-to-year effects. Our results from this supplementary analysis are similar to what we report in the text.

the assumption that the TME variance in the lagged score is constant.²¹ The year-to-year stability in the teacher effects is actually *lower* in the model that makes this correction without any corresponding adjustment for current-score TME (column 2), and the year-to-year stability is only marginally higher than in the unweighted model when we add the TME weights based on the current-year CSEM (column 3). Hence, our results show that it is costly to assume that the TME variance in the lagged score is constant.

Tables 3 and 4 show that incorporating the CSEM information can improve inference from VAMs, but it is not obvious how to interpret our findings. How much better do the TME-adjusted models in Tables 3 and 4 really perform? To quantify the benefits of the CSEM adjustments we benchmark them against the benefits associated with changing a well-understood feature of the data: within-teacher sample sizes. Specifically, we equate the gains in model performance that come from the CSEM adjustments with the gains that would come from increasing the number of students per teacher in the data.

We perform the benchmarking procedure using the year-to-year correlations in the VAM estimates. Note that the year-to-year correlations will get smaller as within-teacher sample sizes decline independent of the TME issue. With this in mind we begin with the correlations from the CSEM-adjusted VAMs. We randomly drop a single student from each teacher's classroom in each year, re-estimate the adjusted models, and re-calculate the correlations. Then we drop a second student from each classroom at random and repeat our analysis again, and so on. For each CSEM-adjusted VAM, there comes a point where the year-to-year stability in the teacher-effect estimates that we estimate using the restricted sample aligns with the year-to-year stability in the estimates

²¹ For the models where we assume that the TME variance in the lagged score is constant, and model it as such (columns 2 and 3), we use the unconditional standard error of measurement (SEM) of the testing instrument to estimate the TME variance. This approach is taken in the models used in New York City and Washington DC (Isenberg and Hock, 2010; Value-Added Research Center, 2010).

from the unweighted model based on the full sample. When this happens, we identify the within-teacher restricted sample size and call it R . We say that the CSEM weights offer benefits in model performance equivalent to increasing within-teacher sample sizes in the unweighted models by $\{(F-R)/R\}$ percent, where F is the average number of students per teacher in the full sample ($F \approx 21$, see Table 2). Put differently, if we were to begin with R students per teacher, we could obtain similar improvements in model performance by either (1) weighting the model using the CSEMs or (2) increasing the average within-teacher sample size from R to F without weighting.

Our results from the benchmarking exercise are reported in Table 5. In our preferred models where we use the gainscore-based CSEM weights (columns 1, 4 and 5), the benefits from the TME adjustments are equivalent to increasing average within-teacher sample sizes by 2-3 students. This finding is consistent in both the lagged-score and gainscore VAMs. Noting that $R \approx 18-19$, the benefits associated with the weighting correspond to increasing the average within-teacher sample size by approximately 11-17 percent.

We also briefly extend our analysis to consider “shrunkened” VAM estimates. Shrunkened estimates are constructed as the weighted average of a prior, typically the average teacher effect, and each teacher’s estimated effect based on available data. Teachers for whom more-precise information is available to estimate effectiveness have more weight applied to the data; teachers for whom less-precise information is available have more weight applied to the prior (for applications of shrinkage estimators in research see Chetty et al., 2011; Goldhaber and Hansen, 2010; Kane et al., 2008; Lockwood et al., 2002). A key benefit of shrinkage in the teacher context is that it addresses variability in the data in within-teacher sample sizes. Albeit indirectly, shrinkage estimators may also

partly account for TME in the sense that teachers who teach students with noisier scores will have more weight applied to the prior.²²

In Table 6 we evaluate shrunken VAM estimates from the gainscore and simple-weighted lagged-score models (columns 1, 4 and 5 in Table 5). We shrink the estimated effect for teacher j in

year t by multiplying it by the reliability ratio $r_{jt} = \frac{\hat{\sigma}_{\theta_t}^2}{\hat{\sigma}_{\theta_t}^2 + \hat{\sigma}_{\lambda_{jt}}^2}$, where $\hat{\sigma}_{\theta_t}^2$ is an estimate of the

variance of true teaching effectiveness (estimated as in Aaronson et al., 2007) and $\hat{\sigma}_{\lambda_{jt}}^2$ is an estimate

of the variance of the effect for teacher j , which we estimate by the square of teacher j 's standard

error. It should be no surprise that the baseline year-to-year stability in the shrunken estimates in

Table 6 is higher than in Table 5.²³ But even for the shrunken estimates we identify noticeable

benefits from directly augmenting the VAMs to incorporate CSEM information. We conclude that

the shrinkage procedure is a broad-brush approach to dealing with the inherent imprecision in

estimated teacher effects, and can be improved upon by applying direct measures of TME into

VAMs.

Finally, for comparative purposes we perform a similar benchmarking procedure to what we show in Tables 5 and 6 using the simulated data. In the 20-students-per-teacher simulation we find

that the improvement in model performance associated with the TME weighting is equivalent to

increasing within-teacher sample sizes by 33 to 43 percent.²⁴ While Tables 5 and 6 show that

²² This depends on how the shrinkage procedure is performed – shrinkage estimators are sometimes constructed in ways that will not account for TME at all. As an example, in some applications the only variability in the shrinkage factors across teachers comes from differences in within-teacher sample sizes. In such cases there is the implicit assumption that all student scores are equally-reliable measures of performance. In the analysis that follows we shrink the estimated teacher effects using a formula that allows for variation across teachers in TME to influence the shrinkage factors (because this is the only interesting case).

²³ Although the shrunken estimates are more stable, there is no consensus in the literature as to whether shrunken estimates are preferred. Shrunken estimates improve stability by introducing systematic bias (toward the prior) in the teacher effects.

²⁴ Per the preceding discussion, the performance standards by which we judge the models differ between the simulation and real-data analyses, which makes this comparison rough (i.e., in the simulations we measure model performance in

inference from VAMs can clearly be improved by the CSEM weighting, the discrepancy between the gains from the TME adjustments in the simulations and real-data models suggests further improvement may be possible.

VI. Conclusion

We examine the benefits associated with incorporating information about test measurement error into value-added models and highlight two key findings from our study. First, we use simulations to show that under ideal estimation conditions, modifying VAMs to include TME information leads to meaningful improvements in model performance. The benefits in model performance suggested by our simulations are larger than the benefits that we observe in our real-data analysis, which suggests that it may be feasible to produce TME metrics that are more accurate than the metrics currently provided by most test publishers. Even so, in our real-data models we confirm that inference from VAMs is improved when the models are modified to incorporate publisher-provided CSEMs, which are readily available for most large-scale standardized tests nationwide. The benefits associated with incorporating the CSEM information into VAMs are equivalent to what would be observed if we could increase within-teacher sample sizes by 11-17 percent. This is a notable result because CSEM data can be incorporated into VAMs at zero cost.

terms of how well the models replicate the true parameter values; in the real-data models we use the stability in the year-to-year estimates of teacher effects).

References

- Aaronson, Daniel, Lisa Barrow and William Sander. 2007. Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25(1), 95-135.
- Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja and Tristan Zajonc. 2011. Do Value-Added Estimates Add Value? Accounting for Learning Dynamics. *American Economic Journal: Applied Economics* 3(3), 29-54.
- Ballou, Dale. 2009. Test Scaling and Value-Added Measurement. *Education Finance and Policy* 4(4), 351-383.
- Betebenner, Damien. 2008. A Primer on Student Growth Percentiles. Policy Report. National Center for the Improvement of Educational Assessment.
- Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff .2011. Teacher Layoffs: An Empirical Illustration of Seniority v. Measures of Effectiveness. *Education Finance and Policy* 6(3), 439-54.
- Boyd, Donald, Pam Grossman, Hamilton Lankford, Susanna Loeb and Jim Wyckoff. 2008. Measuring Effect Sizes: The Effect of Measurement Error. CALDER Working Paper No. 19.
- Briggs, Derek and Ben Domingue. 2011. Due Diligence and the Evaluation of Teachers. National Education Policy Center Report.
- Brown, Michael L. 1982. Robust Line Estimation with Errors in Both Variables. *Journal of the American Statistical Association* 77, 71-9.
- Carroll, Raymond J. and P.P. Gallo. 1982. Some Aspects of Robustness in the Functional Errors-in-Variables Model. *Communications in Statistics, Series A* 11, 2573-2585.
- Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2011. The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. National Bureau of Economic Research Working Paper No. 17699.
- Corcoran, Sean P. 2010. Can Teachers be Evaluated by their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Report for the Annenberg Institute for School Reform, Education Policy for Action Series.
- Fuller, Wayne A. 1987. *Measurement Error Models*. John Wiley & Sons, Inc.
- Goldhaber, Dan and Duncan Chaplin. 2011. Assessing the Rothstein Falsification Test: Does It Really Show Teacher Value-Added Models Are Biased? CEDR Working Paper 2011-5.
- Goldhaber, Dan and Michael Hansen. 2010. Using Performance on the Job to Inform Teacher Tenure Decisions. *American Economic Review (PE&P)* 100(2), 250-255.
- Hambleton, Ronald K., H. Swaminathan and H. Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

- Hanushek, Eric A. 2009. Teacher Deselection, in *Creating a New Teaching Profession* eds. Dan Goldhaber and Jane Hannaway. Urban Institute, Washington, DC.
- . 2011. The Economic Value of Higher Teacher Quality. *Economics of Education Review* 30(3), 466-479.
- Hanushek, Eric A. and Steven G. Rivkin. 2010. Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review (P&P)* 100(2), 267-271.
- Harris, Douglas. 2011. *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press: Cambridge, MA.
- Harris, Douglas and Tim R. Sass. 2006. Value-Added Models and the Measurement of Teacher Quality. Unpublished Manuscript, Florida State University.
- Hill, Heather. 2009. Evaluating Value-Added Models: A Validity Argument Approach. *Journal of Policy Analysis and Management* 28(4), 700-708.
- Isenberg, Eric and Heinrich Hock. 2010. Measuring School and Teacher Value Added for IMPACT and TEAM in DC Public Schools: Final Report. Unpublished report, Mathematica Policy Research.
- Kane, Thomas J. and Douglas O. Staiger. 2002. The Promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives* 16(4), 91-114.
- . 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. National Bureau of Economic Research Working Paper No. 14607.
- Kane, Thomas J., Jonah E. Rockoff and Douglas O. Staiger. 2008. What Does Certification Tell us About Teacher Effectiveness? Evidence from New York City. *Economics of Education Review* 27(6), 615-631.
- Kinsler, Joshua. 2011. Assessing Rothstein's Critique of Teacher Value-Added Models. Unpublished Manuscript, University of Rochester.
- Koedel, Cory and Julian R. Betts. 2011. Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy* 6(1), 18-42.
- Lockwood, J.R., Thomas A. Louis and Daniel F. McCaffrey. 2002. Uncertainty in Rank Estimation: Implications for Value-Added Modeling Accountability Systems. *Journal of Educational and Behavioral Statistics* 27(3), 255-270.
- Lord, Frederic M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

- . 1984. Standard Error of Measurement at Different Ability Levels. *Journal of Educational Measurement* 21(3), 239-243.
- McCaffrey, Daniel F., J.R. Lockwood, Daniel M. Koretz and Laura S. Hamilton. 2003. *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica, CA: The RAND Corporation.
- Neal, Derek and Gadi Barlevy (forthcoming). Pay for Percentile. *American Economic Review*.
- Podgursky, Michael J. and Mathew G. Springer. 2007. Teacher Performance Pay: A Review. *Journal of Policy Analysis and Management* 26(4), 909-949.
- Rockoff, Jonah. 2004. The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review (P&P)* 94(2), 247-252
- Rothstein, Jesse. 2010. Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics* 125(1), 175-214.
- Staiger, Douglas O. and Jonah E. Rockoff. 2010. Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives* 24(3), 97-118.
- The New Teacher Project. 2009. The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. TNTP Policy Report.
- Thompson, Tony D. 2008. Growth, Precision, and CAT: An Examination of Gain Score Conditional SEM. Unpublished Research Report. Pearson Publishing.
- Thorndike, Robert L. 1951. Reliability. In *Educational Measurement* edited by Everett F. Lindquist. Washington, DC: American Council on Education.
- Todd, Petra E. and Kenneth I. Wolpin. 2003. On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal* 113, F3-F33
- Value-Added Research Center. 2010. NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model. Unpublished report, Wisconsin Center for Education Research, University of Wisconsin-Madison.
- Zamar, Ruben H. 1989. Robust Estimation in the Errors-in-Variables Model. *Biometrika* 76(1), 149-160.

Figure 1. Conditional Standard Errors of Measurement for Student Level Scores from the Missouri Data. Grade-5 Students, 2009.

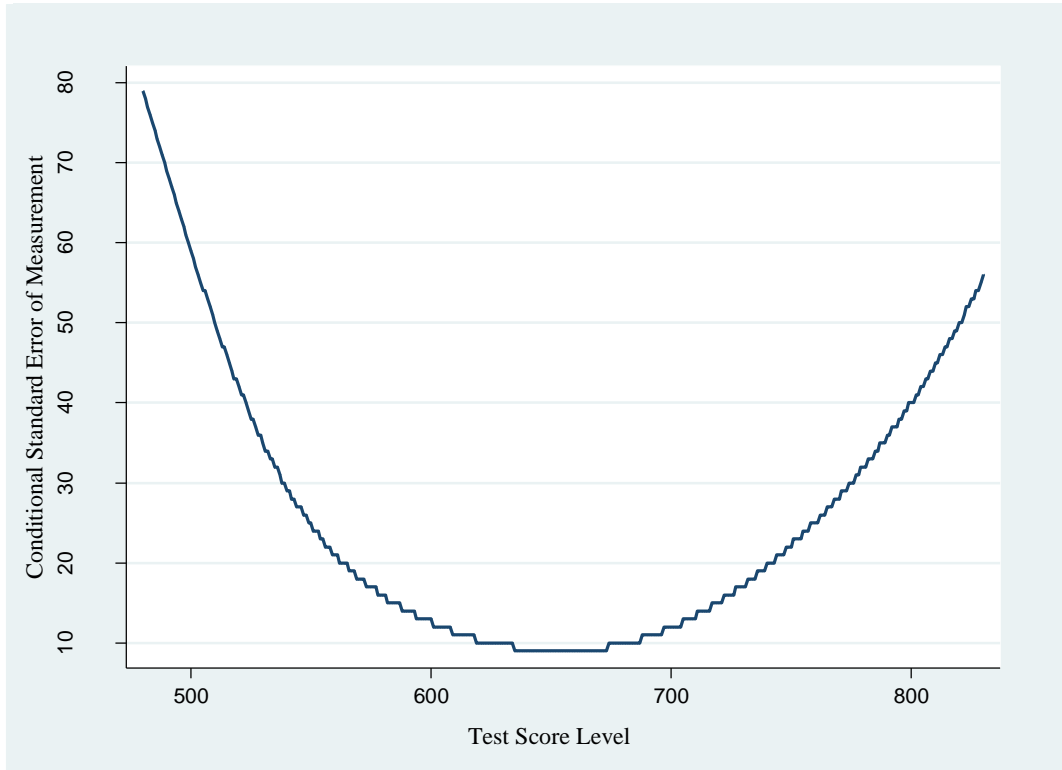


Table 1. Simulation Results. Random Assignment of Students to Teachers

<u>Class Size</u>	Unweighted VAM	Weighted VAM
<i>Correlation between teacher-effect estimates and actual values</i>		
5	0.40	0.46
10	0.52	0.59
20	0.66	0.72
40	0.78	0.83
<i>Share of teachers assigned to the wrong quintile</i>		
5	0.70	0.69
10	0.67	0.65
20	0.61	0.59
40	0.55	0.52
<i>Share of teachers with “major” error in quintile assignment</i>		
5	0.32	0.31
10	0.28	0.26
20	0.22	0.19
40	0.14	0.10

Notes: The simulations are performed 1,000 times. Averages across simulations are reported in the cells in the table. A “major” error in quintile assignment occurs when a teacher’s VAM estimate suggests that she belongs in a quintile that is more than one quintile away from her actual placement (e.g., a 15th percentile teacher being placed in the 43rd percentile based on her VAM estimate).

Table 2. Data Details.

	<u>2008-2009</u>	<u>2009-2010</u>
Number of students	35,027	35,394
Number of teachers	1,657	1,657
Number of schools	662	662
Average class size	21.1	21.4
<u>Sample means for students</u>		
Minority share	0.17	0.17
Female share	0.49	0.49
Free/reduced-price lunch share	0.42	0.44
Limited English proficiency share	0.02	0.02
Share of mobile students	0.01	0.01
Share of students with individualized education program	0.13	0.13

Notes: The data include all grade-5 students in Missouri with current and lagged MAP scores who can be linked to a valid classroom teacher. Teachers are required to have taught 10-or-more students in the same school for each year of the data panel.

Table 3. Year-to-Year Stability of Teacher Effects from Missouri Gainscore Models.

	Unweighted	CSEM-Weighted
Year-to-year correlation between teacher-effect estimates	0.454	0.475
Share of teachers with any change to their quintile assignment from year-to-year	0.688	0.684
Share of teachers with major change to their quintile assignment from year-to-year	0.313	0.306

Note: A “major” change in quintile assignment occurs when a teacher moves more than one quintile in the teacher rankings from year to year (e.g., from the 15th to 43rd percentiles).

Table 4. Year-to-Year Stability of Teacher Effects from Missouri Lagged-Score Models.

	Unweighted	Correct for Constant Lagged- Score TME Variance, Unweighted	Correct for Constant Lagged-Score TME Variance, Weight by Current-Score CSEM	Weight by Gainscore CSEM	Weight by Modified Gainscore CSEM
Year-to-year correlation between teacher-effect estimates	0.466	0.448	0.471	0.487	0.487
Share of teachers with any change to their quintile assignment from year-to- year	0.691	0.698	0.690	0.685	0.686
Share of teachers with major change to their quintile assignment from year-to- year	0.307	0.316	0.313	0.299	0.300

Notes: A “major” quintile assignment change occurs when a teacher moves more than one quintile in the teacher rankings from year to year (e.g., from the 15th to 43rd percentiles). The four adjustments to the VAM are described in more detail in the text (see the end of Section V).

Table 5. Benchmarking Improvements in Model Performance from the CSEM Adjustments Using Changes to Within-Teacher Sample Sizes.

	<u>Gainscore Model</u>		<u>Lagged-Score Models</u>		
	Weight by Gainscore CSEM	Correct for Constant Lagged-Score TME Variance, Unweighted	Correct for Constant Lagged-Score TME Variance, Weight by Current-Score CSEM	Weight by Gainscore CSEM	Weight by Modified Gainscore CSEM
Unadjusted VAM “target” correlation (from Tables 3 and 4)	0.454	0.466	0.466	0.466	0.466
CSEM-adjusted VAM correlation, full sample (from Tables 3 and 4)	0.475	0.448	0.471	0.487	0.487
<u>CSEM-adjusted VAM correlations:</u>					
Drop one student per teacher per year	0.469	N/A	0.464	0.480	0.480
Drop two students per teacher per year	0.461	N/A	0.456	0.472	0.473
Drop three students per teacher per year	0.451	N/A	0.446	0.463	0.463
Drop four students per teacher per year	0.442	N/A	0.437	0.453	0.453
Approximate per-teacher sample size gain from TME model adjustment	2-3 students	N/A	0-1 students	2-3 students	2-3 students

Notes: The average within-teacher sample size in the full dataset is approximately 21 (see Table 2). At each restricted sample-size threshold (e.g., drop 1 student, drop 2 students, etc.), we repeat the procedure by which students are randomly dropped from teachers’ classrooms 10 times. The average correlation across the 10 iterations is reported in each cell. The correlations are very stable across iterations.

Table 6. Benchmarking Improvements in Model Performance from the CSEM Adjustments Using Changes to Within-Teacher Sample Sizes. Shrunk Teacher-Effect Estimates.

	<u>Gainscore Model</u>	<u>Lagged-Score Models</u>	
	Weight by Gainscore CSEM	Weight by Gainscore CSEM	Weight by Modified Gainscore CSEM
Unadjusted shrunken VAM “target” correlation	0.457	0.467	0.467
CSEM-adjusted shrunken VAM correlation, full sample	0.477	0.488	0.488
<u>CSEM-adjusted VAM correlations:</u>			
Drop one student per teacher per year	0.470	0.481	0.482
Drop two students per teacher per year	0.464	0.475	0.475
Drop three students per teacher per year	0.454	0.465	0.465
Drop four students per teacher per year	0.446	0.456	0.456
Approximate per-teacher sample size gain from TME model adjustment	2-3 students	2-3 students	2-3 students

Notes: Same as in Table 5.

Appendix A Alternative Simulation Exercise

In this appendix we consider the case where students are perfectly sorted into classrooms by the student component of their test-score gains, then assigned to teachers at random. We call this the “ability grouping” scenario.²⁵ Recall that the TME metrics in the simulations are assigned to students such that the largest TME variance is assigned to the largest gainscore in absolute value. This means that in the ability-grouping scenario some teachers will have classrooms full of students with noisily-measured scores, and others will have classrooms full of students with precisely-measured scores. This is the key distinction between the estimates presented in this appendix and those reported in Table 1.

We relegate these results to the appendix because, frankly, they are not very interesting. The real mileage from the TME-adjusted VAMs comes from the shift in weighting toward more-precisely measured student test scores *within teachers*. That is, we take teachers who teach students with some noisy and some precise scores and put more weight on the precise scores. When a teacher has students with all noisy or all precise scores, the benefit of the weighting is limited because the reliability of the information from individual students is similar within teachers. This fact is reflected in our results in Table A.1 – the weighting has essentially no effect in any of the class-size scenarios that we consider under this alternative sorting scheme.

We also note that as a practical matter most of the variation in student test scores occurs within schools (also see Kane and Staiger, 2002). This means that in reality, data conditions are such that most teachers have students with scores from across the distribution (which, in turn means that teachers are exposed to students with a wide range of TME in their scores). This further supports our random-assignment simulations as being more-relevant than the ability-grouping simulations.

Finally, we briefly comment on the fact that the estimates in Table A.1 are consistently less-accurate than what we find when we assign students to teachers at random (Table 1). This, of course, is because ability grouping introduces bias into the teacher effects by construction. Teachers’ actual effects, as well as the bias, are both highlighted by the weighting.

²⁵ In practice, of course, ability grouping would be more likely to occur along the dimension of test score levels than gains, but the approximation here is sufficient to make our point.

Table A.1. Simulation Results. Ability Grouping Scenario.

<u>Class Size</u>	Unweighted VAM	Weighted VAM
<i>Correlations between teacher-effect estimates and actual values</i>		
5	0.29	0.29
10	0.31	0.32
20	0.27	0.27
40	0.36	0.36
<i>Shares of teachers assigned to the wrong quintile</i>		
5	0.72	0.72
10	0.72	0.71
20	0.72	0.72
40	0.76	0.76
<i>Shares of teachers with “major” error in quintile assignment</i>		
5	0.35	0.35
10	0.34	0.34
20	0.36	0.36
40	0.32	0.32

Note: The simulations are performed 1,000 times. Averages across simulations are reported in the cells in the table. A “major” error in quintile assignment occurs when a teacher’s VAM estimate suggests that she belongs in a quintile that is more than one quintile away from her actual placement (e.g., a 15th percentile teacher being placed in the 43rd percentile based on her VAM estimate).