

Estimating Educational Effects using Analysis of Covariance with Measurement Error

S. Paul Wright
SAS Institute Inc.

Prepared for CREATE's 2008 National Evaluation Institute
October 9-11, 2008
Wilmington, NC

Abstract. A popular value-added model (VAM) for estimating educational effects is the analysis of covariance (ANCOVA) model with the response variable (Y) being the current-year score, the covariate (X) being the previous-year (or beginning-of-the-year) score, and the grouping variable being the educational effect of interest (e.g., district, school, teacher). Since X is measured with error, its slope is biased toward zero. The estimated educational effects are also biased; and, being incompletely adjusted for prior schooling, the effects also tend to be correlated with demographic/socioeconomic variables. Attempts to correct for the demographic correlations by including demographic variables in the analysis may actually exacerbate the bias in the educational effects. This will be demonstrated via simulation, and less biased alternatives to the simple ANCOVA model will be presented.

Introduction

The purpose of this paper is to point out the danger of using a simple analysis of covariance (ANCOVA) model for value-added assessment of education providers (districts, schools, teachers). Such models, often implemented as a form of hierarchical linear model (HLM), have become popular, in part because of their perceived transparency. However, as applied in practice, these models have serious short-comings that are often unrecognized or ignored. In particular, the use of socioeconomic variables in the model to achieve "fairness" can be notably unfair.

Value-added models (VAMs) are useful and popular tools for evaluating education providers. They are based on the premise that education providers should be assisting students in the acquisition of new knowledge and skills. Therefore, it makes sense to assess how much students have learned (using a value-added model) rather than just how much they know (using a status model as required, for example, by the No Child Left Behind legislation). The value-added approach is especially attractive since students' status on entering a class is highly variable and not under the teacher's control.

The appeal of the value-added approach is such that a number of different value-added models have been developed, and the literature on value-added modeling is large and growing. The choice of a VAM unfortunately involves a fundamental conflict of interests. On the one hand, there is demand for "transparent" models that are easy to explain to educators and, usually, easy to compute. On the other hand, model results are expected to be accurate and stable. The bad news is that, in settings that often occur in practice, the simple, transparent models tend to be inaccurate ("biased" in statistical terminology) and/or unstable (having a large "variance" in

statistical terminology). An important source of the bias and variability in these models is the presence of measurement error in the tests scores that constitute the data for VAMs.

This paper will explore the consequences of measurement error on simple ANCOVA models by means of simulation and will consider several alternative approaches that may (or may not) address those consequences.

The ANCOVA model

The statistical model for a simple analysis of covariance is

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \tau_j + \varepsilon_{ij}.$$

$j = 1, \dots, m$ identifies a teacher.

$i = 1, \dots, N$ identifies a student.

Y_{ij} is a post-test score (e.g., spring of current year).

X_{ij} is a pre-test score (e.g., previous spring or current fall).

τ_j is a teacher “effect” (random or fixed).

The fitted equation, with unknown parameters replaced by their estimates, is written

$$\hat{Y}_{ij} = b_0 + b_1 X_{ij} + t_j = \bar{Y} + b_1 (X_{ij} - \bar{X}) + t_j.$$

When X is measured with error (as is always the case with test scores), then

- the estimated slope (b_1) is biased toward zero;
- the estimated teacher effects (t_j) are biased toward “unadjusted,” i.e., toward the value they would have in a status model rather than a value-added model;
- the estimated teacher effects tend to be highly correlated with their students’ socioeconomic status.

These three consequences are related in that, as the slope approaches zero, the fitted equation approaches $\bar{Y} + t_j$; that is, the resulting teacher effect measures the average deviation of the scores of this teacher’s students from the mean score of all students – a measure of status rather than value-added. It has long been observed that socioeconomically disadvantaged students tend to have lower test scores than socioeconomically advantaged students. Consequently, when teacher effects measure status rather than value-added, those teacher effects reflect the socioeconomic status of the teacher’s students. It is this third consequence which, understandably, receives the most attention, since the resulting teacher effects are clearly unfair.

Addressing the Consequences of Measurement Error in ANCOVA

A Simple “Fairness” Adjustment. The observed correlation between estimated teacher effects and student socioeconomic status has often been addressed by introducing socioeconomic indicators into the ANCOVA model.

Let

P_{ij} = a student poverty status indicator (1=yes, 0=no) such as the student's free/reduced-price lunch eligibility;

W_j = a classroom level poverty indicator, such as the percentage of students eligible for free/reduced-price lunch.

A revised version of the (fitted) model is then

$$\hat{Y}_{ij} = b_0 + c_1 P_{ij} + c_2 W_j + b_1 X_{ij} + t_j.$$

This model adjusts the line (represented by $b_0 + b_1 X_{ij}$) upward or downward depending on the values of P_{ij} and W_j . The teacher effects are then based on deviations from this adjusted line.

The main purpose of this paper is to demonstrate how this attempt at fairness can go terribly wrong.

Other approaches. A straightforward extension of the simple ANCOVA model is to include multiple pre-test scores. In the simple ANCOVA, if Y is a 5th grade math score, then X will typically be a 4th grade math score from the previous year. Additional pre-test scores that might be included are 4th grade scores in other subjects as well as test scores (math and other subjects) from even further back. The fitted model would then look like this:

$$\hat{Y}_{ij} = b_0 + b_1 X_{1ij} + b_2 X_{2ij} + b_3 X_{3ij} + K + t_j$$

where each X is a pre-test score.

Another approach, popular with econometricians, is the instrumental variables model in which the observed X is replaced by a "predicted X " which, if properly constructed, avoids the biasing effects of measurement error. This model can be represented by two equations (and is therefore sometimes called two-stage least squares):

$$\hat{X}_{1ij} = c_0 + c_1 X_{2ij} + t_j;$$

$$\hat{Y}_{ij} = b_0 + b_1 \hat{X}_{1ij} + t_j.$$

X_2 in the first-stage equation is commonly a test score (in the same subject as Y and X_1) from two years previously. E.g., if Y is 5th grade math, then X_1 is 4th grade math and X_2 is 3rd grade math.

The most complex (and least transparent) approach is a multivariate, longitudinal, mixed model. These models have been described and discussed, for example, in Sanders, et al. (1997) and McCaffrey, et al. (2004). They will not be discussed further here.

The Simulated Data

To assess the consequences of measurement error in the test scores, one thousand data sets were generated then analyzed with a variety of VAMs, mostly ANCOVA models. Each data set consisted of 300 students and 12 teachers (25 students per teacher). For each teacher, a true teacher effect (τ_j) was generated. For each student, a true pre-test score (ξ_i) was generated (with no measurement error). From these true values several observed pre-test scores and an observed post-test score were generated by adding measurement error. In addition, each student was assigned a free/reduced-price lunch status. Specifically,

$$\begin{aligned}\tau_j &\sim \text{normal}(0, 12^2) = 12 \Phi^{-1}[(j - 0.5) / m], \quad j = 1, \dots, m=12 \text{ teachers;} \\ \xi_i &\sim \text{normal}(50, 21^2) = 50 + 21 \Phi^{-1}[(i - 0.5) / N], \quad i = 1, \dots, N=300 \text{ students.}\end{aligned}$$

Φ is the standard normal distribution function; and Φ^{-1} is its inverse, the standard normal quantile function. These two equations indicate that the true teacher effects and the true pre-test scores were normally distributed, but they were not generated randomly. Rather, they were generated systematically so that each of the 1000 data sets started with the same teachers and the same students. The observed X and Y scores differed in the different data sets because of measurement error.

$$\begin{aligned}\varepsilon_{kij} &\sim \text{normal}(0, 14^2), \quad k = 0, 1, \dots, 8 \text{ measurement errors;} \\ Y_{ij} &= \xi_i + \tau_j + \varepsilon_{0ij}; \\ X_{1ij} &= \xi_i + \varepsilon_{1ij}; \\ X_{2ij} &= \xi_i + \varepsilon_{2ij}; \\ &\text{etc.}\end{aligned}$$

In all, eight observed pre-test scores were generated along with the one post-test score. Admittedly, this simulation is unrealistic in that all the observed pre-test scores were derived from the same true pre-test score whereas, in practice, multiple pre-test scores would come from different subjects and grades with different underlying true scores. However, for the purpose of showing the impact of measurement error on estimated teacher effects, the added complexity of generating multiple true scores seemed unnecessary since it would not alter the conclusions.

Finally, each student was assigned a poverty status (free/reduced-price lunch or FRPL status) with half the students in each category. Also, each student was assigned to a particular teacher. The assignment of FRPL status and assignment of students to teachers was done using three different scenarios. These are described below along with the simulation results for each scenario.

Scenario #1

In this scenario

- $\text{Prob}(P_{ij}=1) = 0.50$; half of the students were free/reduced-price lunch eligible.
- The P_{ij} values (0, 1) were assigned randomly to students.
- Students were randomly assigned to teachers.

Figures 1 to 3 verify that the above assignments worked as intended. Each figure contains 12000 points, one for each of the 12 teachers in each of the 1000 data sets. Figure 1 verifies that the classroom %FRPL was uncorrelated with the classroom mean true score (subsequently identified as $\bar{\xi}_j$). Figures 2 and 3 show that the 12 true teacher effects ranged from -21 to +21. These teacher effects were uncorrelated with either $\bar{\xi}_j$ (Figure 2) or %FRPL (Figure 3).

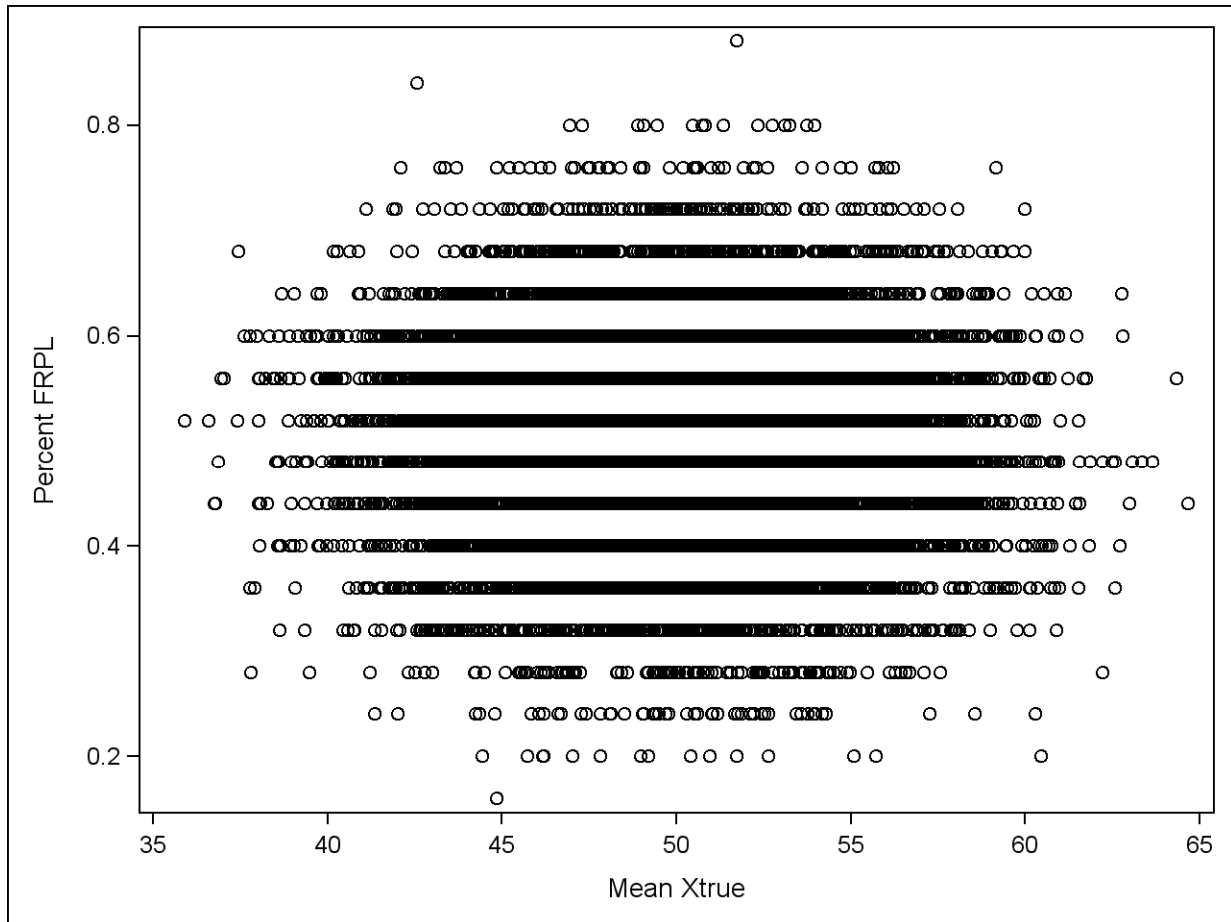


Figure 1. Classroom percent free/reduced-price lunch versus classroom mean true pre-test score for 12 teachers over 1000 simulated data sets using Scenario #1.

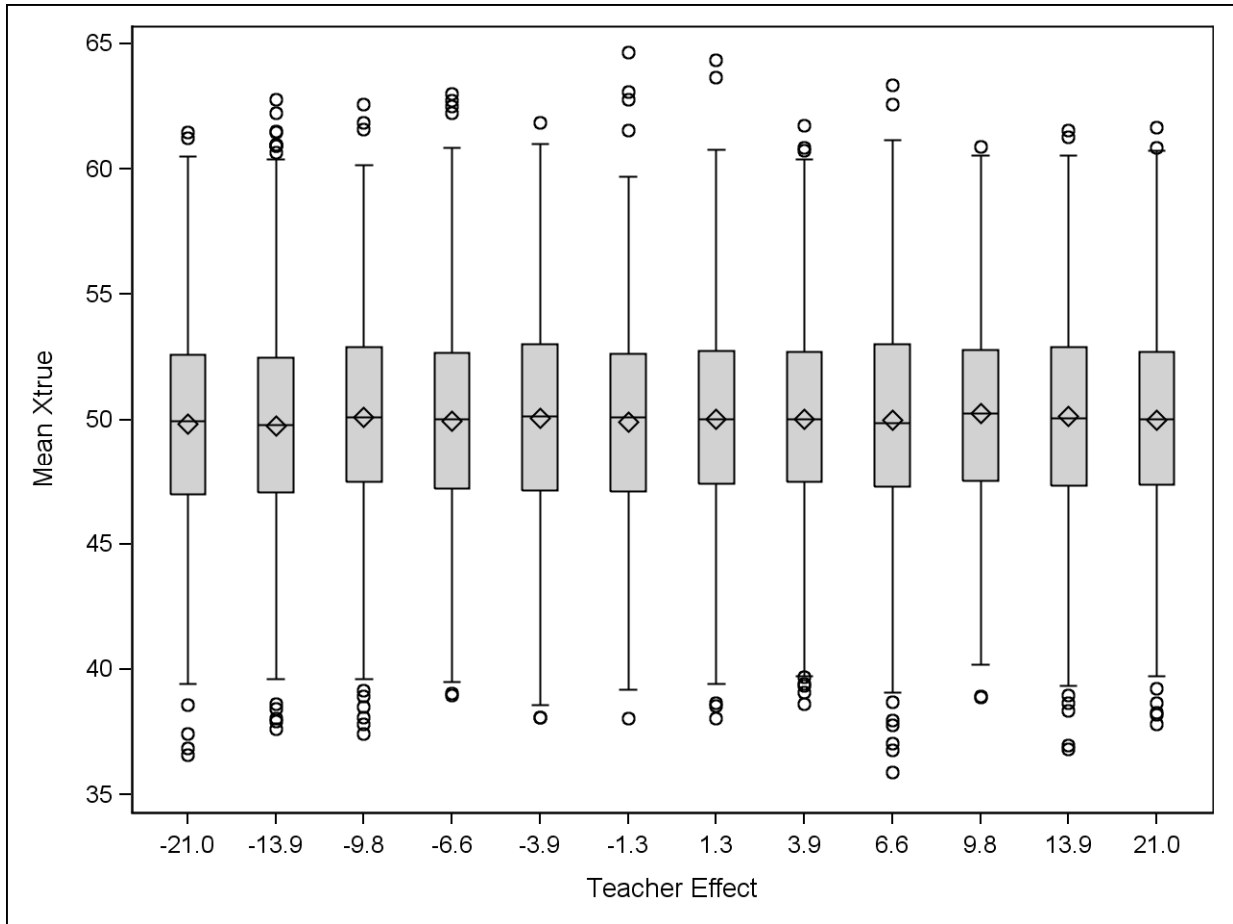


Figure 2. Classroom mean true pre-test score versus true teacher effect for 12 teachers over 1000 simulated data sets using Scenario #1.

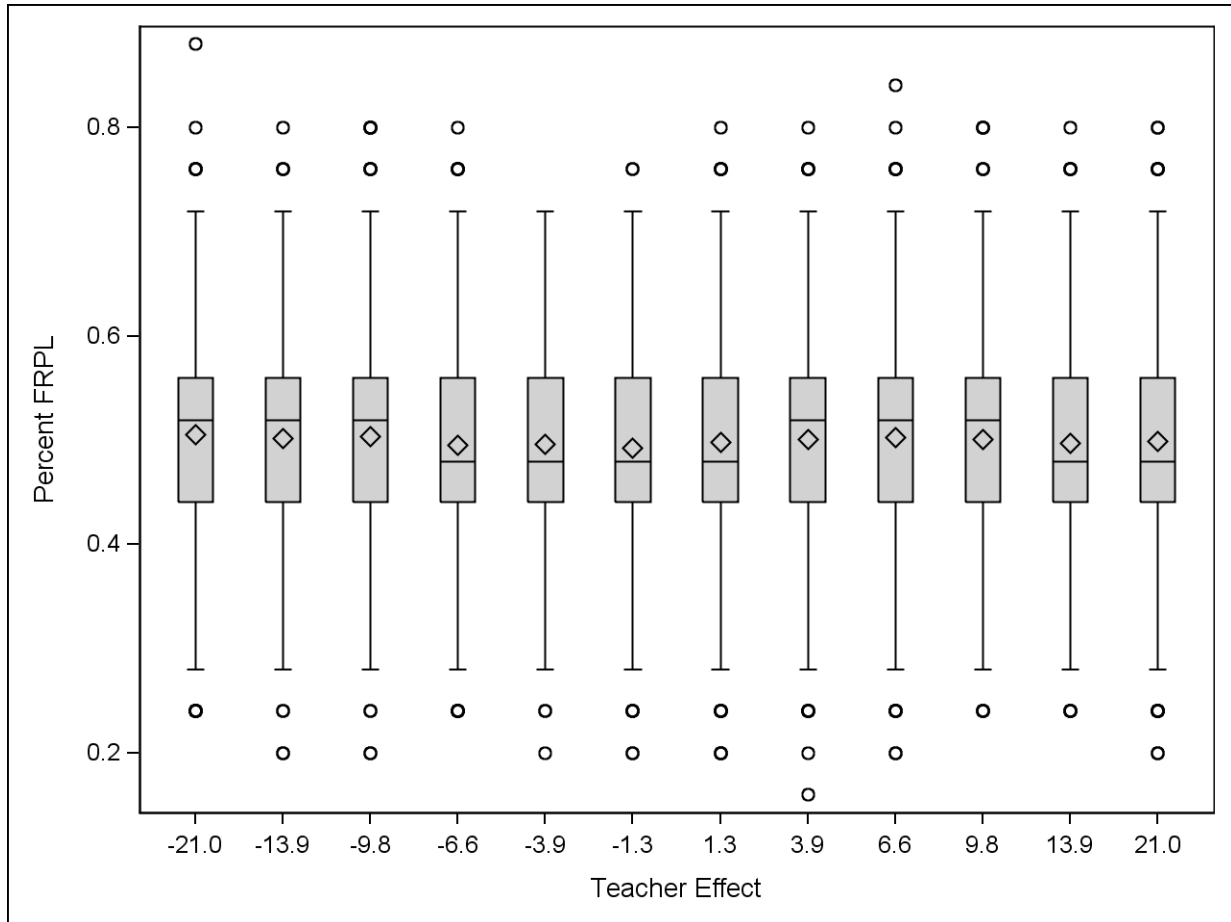


Figure 3. Classroom percent free/reduced-price lunch versus true teacher effect for 12 teachers over 1000 simulated data sets using Scenario #1.

Comparing Models. Since the objective of value-added modeling is to estimate teacher effects, the modeling results will be evaluated by comparing the true teacher effects (τ_j) to the estimated teacher effects (t_j) averaged over the 12 teachers in the 1000 data sets. The comparison will be made using the mean square error (MSE), defined as

$$\text{MSE} = \text{Avg}[(t_{jr} - \tau_j)^2], \quad j = 1, \dots, 12, \quad r = 1, \dots, 1000.$$

Using *squared* differences has two desirable effects. First, squaring the differences removes the sign of the difference; positive and negative differences are treated equally and do not cancel one another. Second, the MSE can be separated into two components which add together to produce the overall MSE. (Another popular criterion for comparison, the mean absolute error, removes the sign but lacks the additive decomposition that the MSE has.) That is,

$$\text{MSE} = \text{Avg}[(t_{jr} - \bar{t}_j)^2] + \text{Avg}[(\bar{t}_j - \tau_j)^2] = \text{variance} + \text{bias}^2$$

where \bar{t}_j is the average estimated effect for teacher j over the 1000 datasets. In words, bias indicates how far “off-target” the estimates are, on average; variance indicates how scattered the estimates are around their average value (ignoring the fact that the average may be off-target). In the context of the requirement that estimates should be accurate and stable, bias measures their inaccuracy and variance measures their instability.

Figure 4 shows, for Scenario #1, the MSEs (with bars subdivided into bias² and variance components) for a number of VAMs which do *not* employ any fairness adjustments. Each model is identified by an abbreviation that is constructed as follows.

a_bb_ccc:

a = 1: model with only pre-test scores (Xs) as covariates

a = 2: model with Xs and student FRPL status (P)

a = 3: model with Xs, P and classroom %FRPL (W)

bb = RE: model with random teacher effects

bb = FE: model with fixed teacher effects

bb = IV: instrumental variables model (with fixed teacher effects)

ccc = T: model using *true* pre-test score as covariate

ccc = kx: model using k observed pre-test scores as covariates ($k = 1,2,3,4,5,8$)

ccc = AOG: model using gain ($Y-X_1$) as dependent variable with no covariates

In Figure 4, the first two bars, for models that use the true pre-test scores, are shown for reference. They represent the best case, the case in which the X is measured without error. They also show the typical result for fixed-effects versus random-effects models: Random-effects models are slightly biased but have smaller variance and smaller MSE than fixed-effects models. In statistics, this is known as the bias-variance trade-off. This same pattern holds for the subsequent ANCOVA models that use observed X s having measurement error; it also holds for the two AOG (analysis of gains) models.

For both the random-effects and fixed-effects ANCOVAs, the MSE decreases as more pre-test scores are included in the model, showing that this is an effective method for reducing the effects of measurement error. The two AOG models have noticeably higher MSEs than the ANCOVA models, especially the ANCOVAs with multiple pre-test scores. The IV model (which uses fixed teacher effects) is nearly indistinguishable from the fixed-effects AOG model. This is because the post-test scores in these simulations were constructed in such a way that the true intercept and slope are $\beta_0=0$ and $\beta_1=1$, and the AOG model is mathematically equivalent to an ANCOVA model in which the slope is constrained to equal 1. To the extent that the IV model unbias the estimated slope so that it is close to one, the IV model would be expected to reproduce the fixed-effects AOG model results.

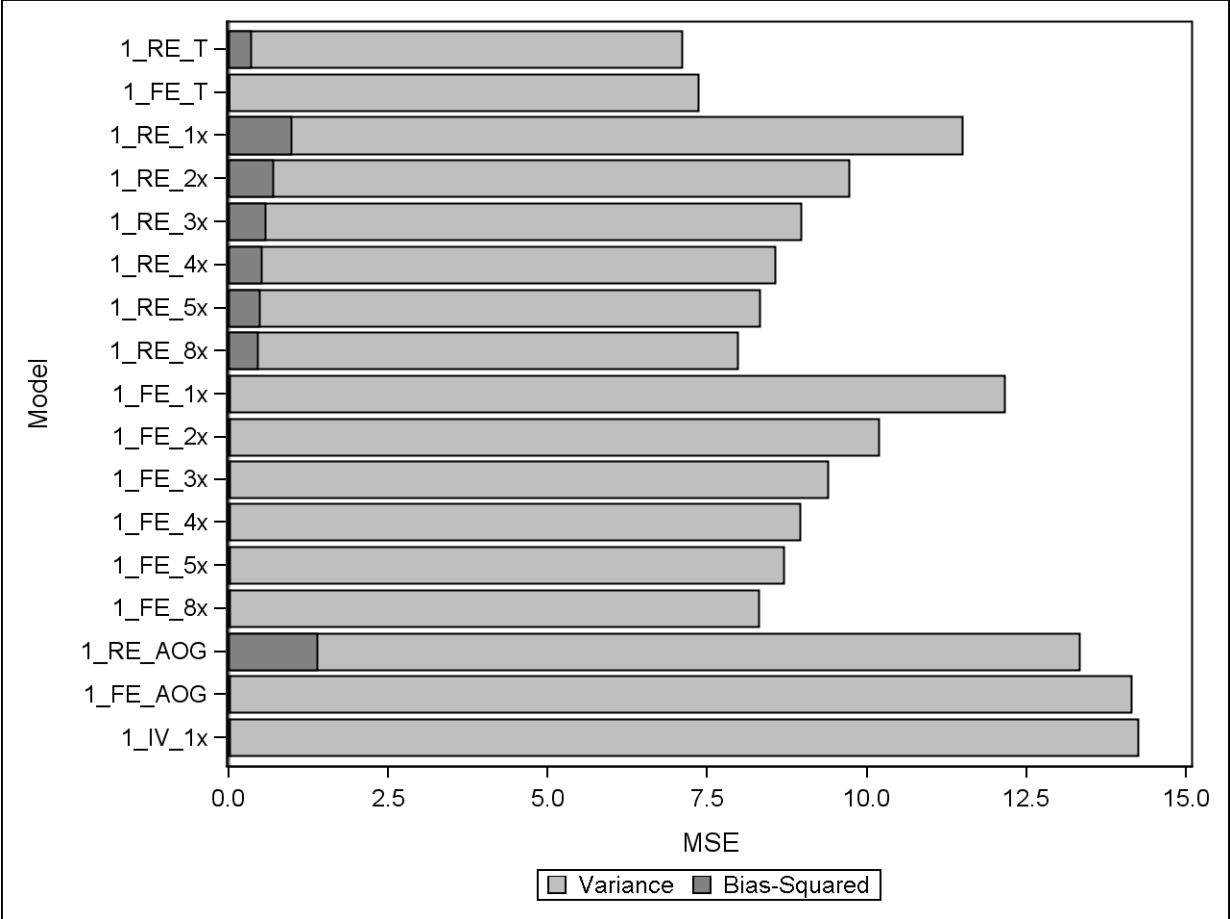


Figure 4. Mean square error (bias-squared plus variance) from models that contained no fairness adjustments for 12 teachers in 1000 simulated data sets using Scenario #1.

Figure 5 shows, for Scenario #1, the MSEs for random-effects ANCOVAs both with and without the fairness adjustments. Results for fixed-effects models (not shown) were similar, but with larger MSEs. There are two messages here. First, including the student-level FRPL indicator has virtually no impact on the MSEs; the results are indistinguishable from those with no fairness adjustment. Second, including the classroom %FRPL makes the results much worse. This is the case even when the pre-test score has no error (3rd bar from the top in Figure 5). Including more pre-test scores in the model improves the results slightly, but the MSEs for the classroom-level adjusted models are still much larger than for the unadjusted models.

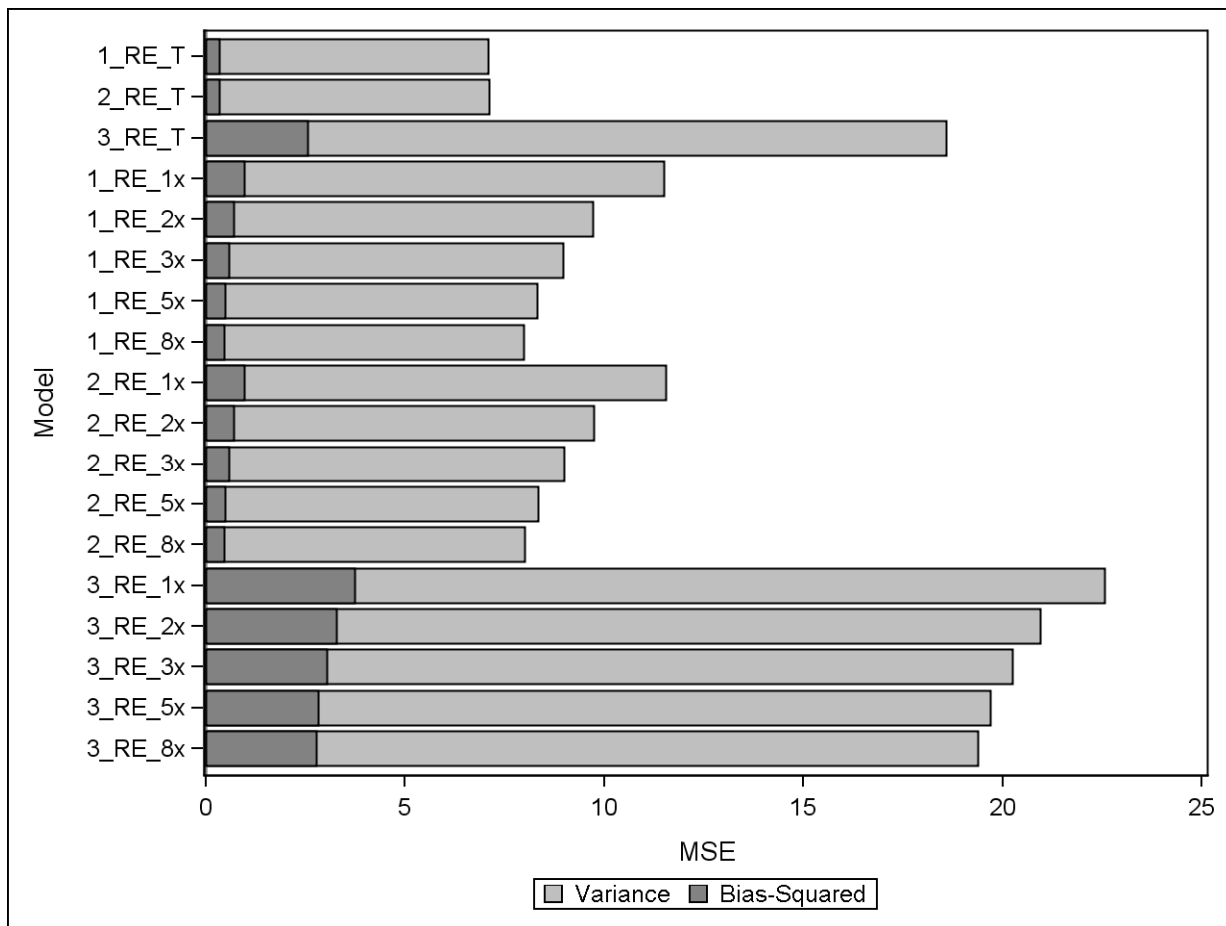


Figure 5. Mean square error (bias-squared plus variance) from selected models, with and without fairness adjustments, for 12 teachers in 1000 simulated data sets using Scenario #1.

What is happening is this. Even though, over the 1000 data sets, there is no correlation between the true teacher effects and classroom %FRPL, within individual data sets there will be a correlation, just by chance. Sometimes better teachers (or poorer teachers, or average teachers) will have a lower percentage of FRPL students; sometimes they will have a higher percentage, again just by chance. Including %FRPL in the model removes that correlation. In doing so, it

removes part of the true teacher effect. The model is no longer estimating the true teacher effect; it is estimating that portion of the true teacher effect which is unrelated to classroom %FRPL.

Scenario #2

In this scenario

- Overall, half of the students were free/reduced-price lunch eligible.
- For individual students, $\text{Prob}(P_{ij}=1)$ decreased with increasing true pre-test score (ξ_i).
- Students were randomly assigned to teachers.

Figure 6 shows the negative correlation between the classroom %FRPL and the classroom mean true score ($\bar{\xi}_j$).

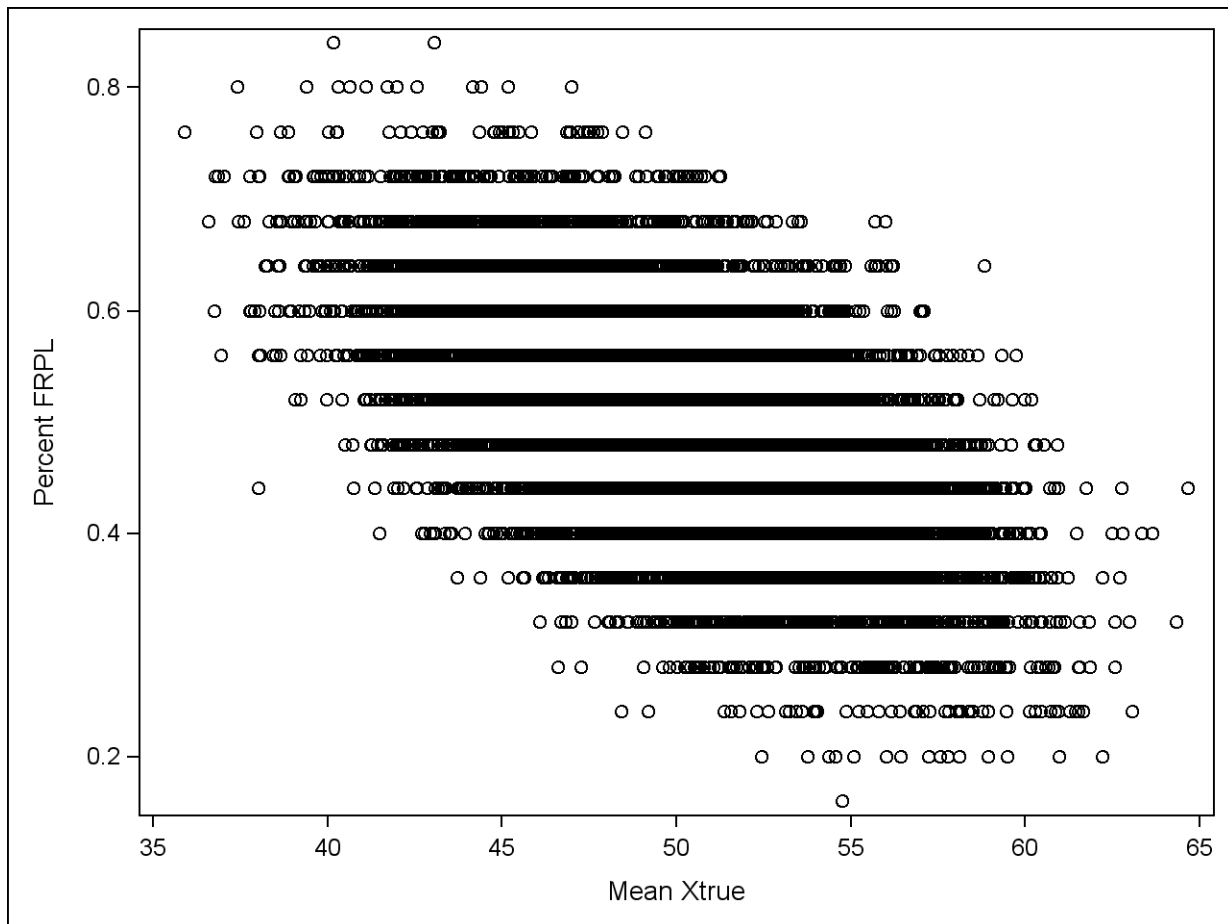


Figure 6. Classroom percent free/reduced-price lunch versus classroom mean true pre-test score for 12 teachers over 1000 simulated data sets using Scenario #2.

It turned out that all other results for Scenario #2 were indistinguishable from Scenario #1, so they have been omitted. The common feature of the two scenarios, of course, was random assignment of students to teachers. In the real world, this rarely happens. Scenario #3 was designed to be more realistic. It has often been noted that higher poverty schools tend to have a higher proportion of inexperienced teachers and teachers teaching outside their credentialed area, and that these teachers are estimated to be less effective on average. Scenario #3 attempts to portray this situation.

Scenario #3

In this scenario

- Overall, half of the students were free/reduced-price lunch eligible.
- For individual students, $\text{Prob}(P_{ij}=1)$ decreased with increasing true pre-test score (ξ_i).
- Lower performing students (lower ξ_i) were more likely to be assigned to a poorer teacher.

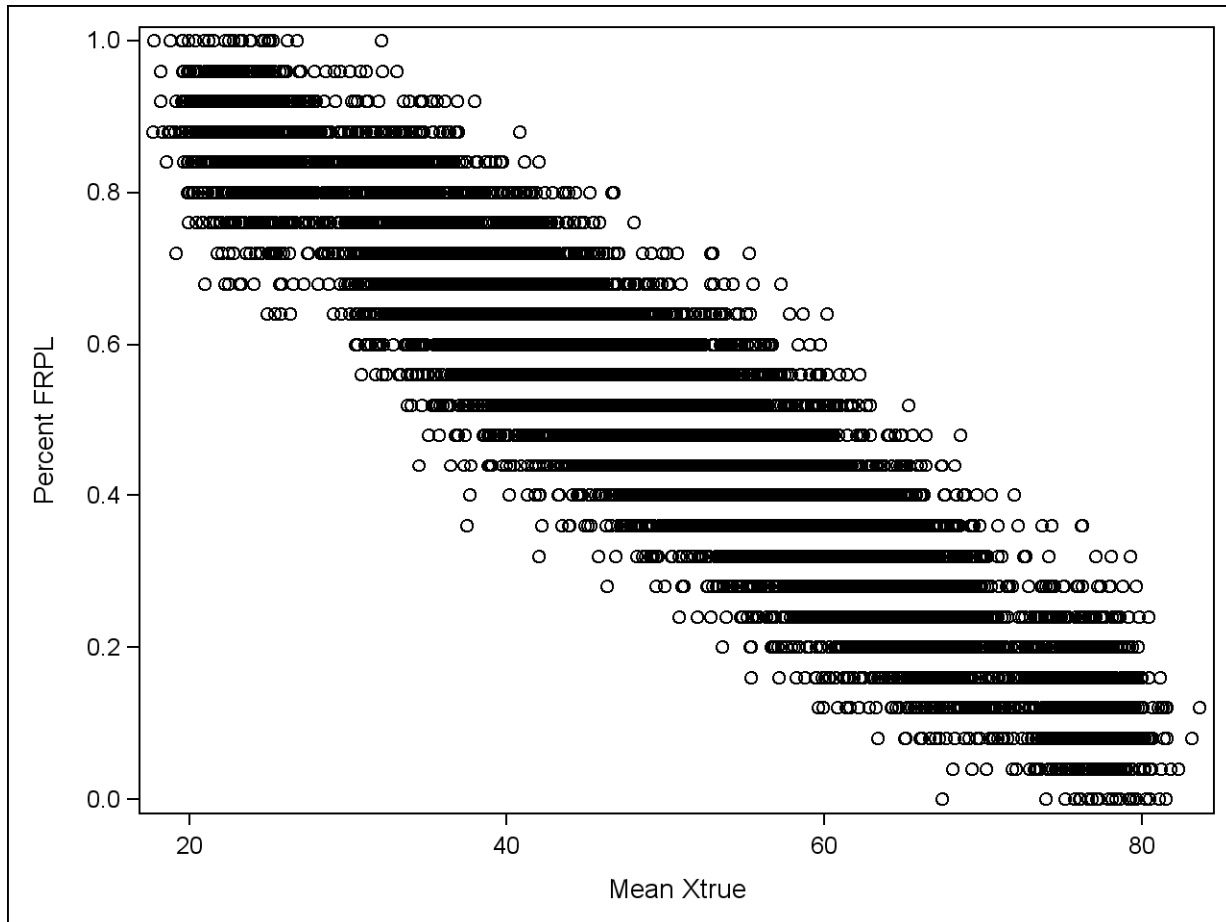


Figure 7. Classroom percent free/reduced-price lunch versus classroom mean true pre-test score for 12 teachers over 1000 simulated data sets using Scenario #3.

Figure 7 shows the negative correlation between the classroom %FRPL and the classroom mean true score ($\bar{\xi}_j$). The correlation is stronger than in Scenario #2 because students are no longer randomly assigned to teachers. Figure 8 shows the positive correlation between the teacher effect and the classroom mean true score ($\bar{\xi}_j$). Figure 9 shows the negative correlation between the teacher effect and the classroom %FRPL.

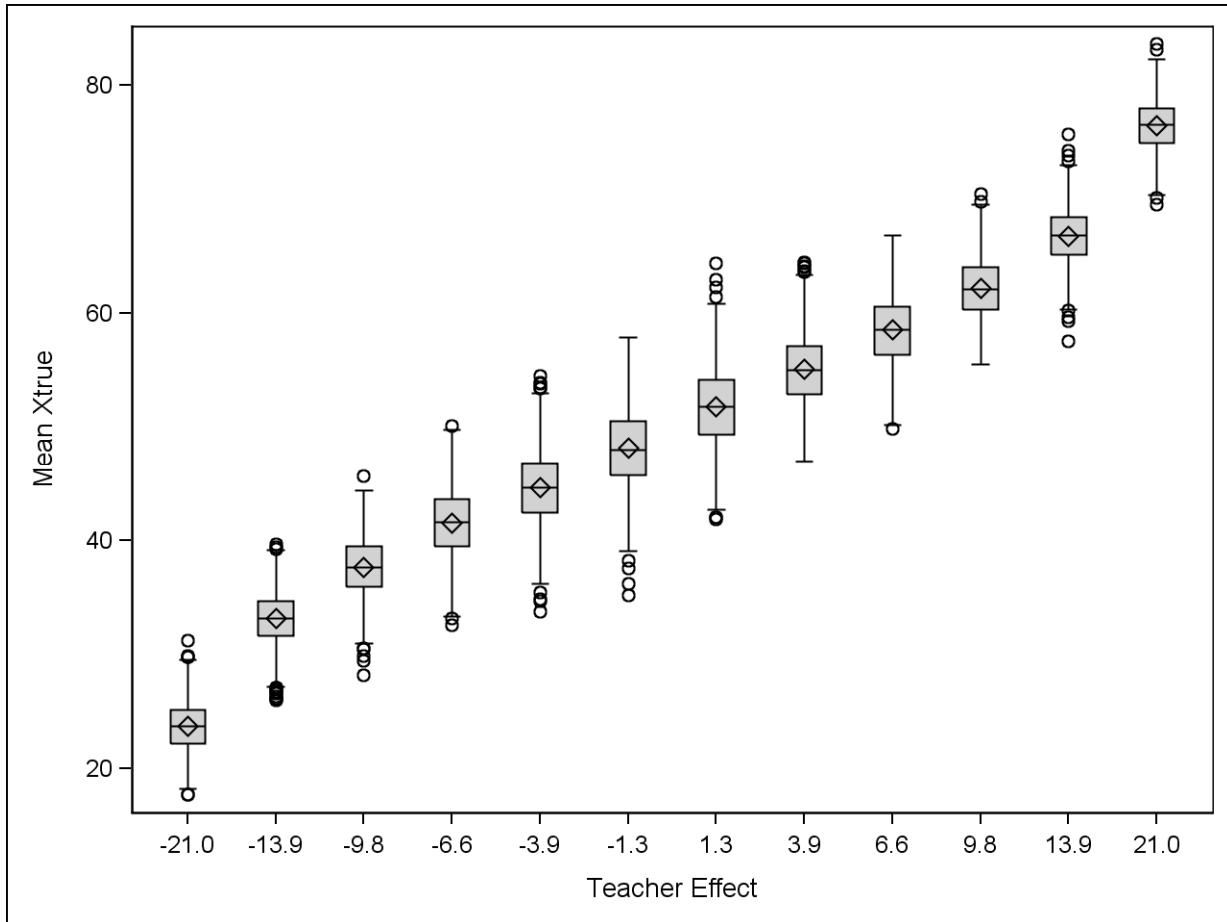


Figure 8. Classroom mean true pre-test score versus true teacher effect for 12 teachers over 1000 simulated data sets using Scenario #3.

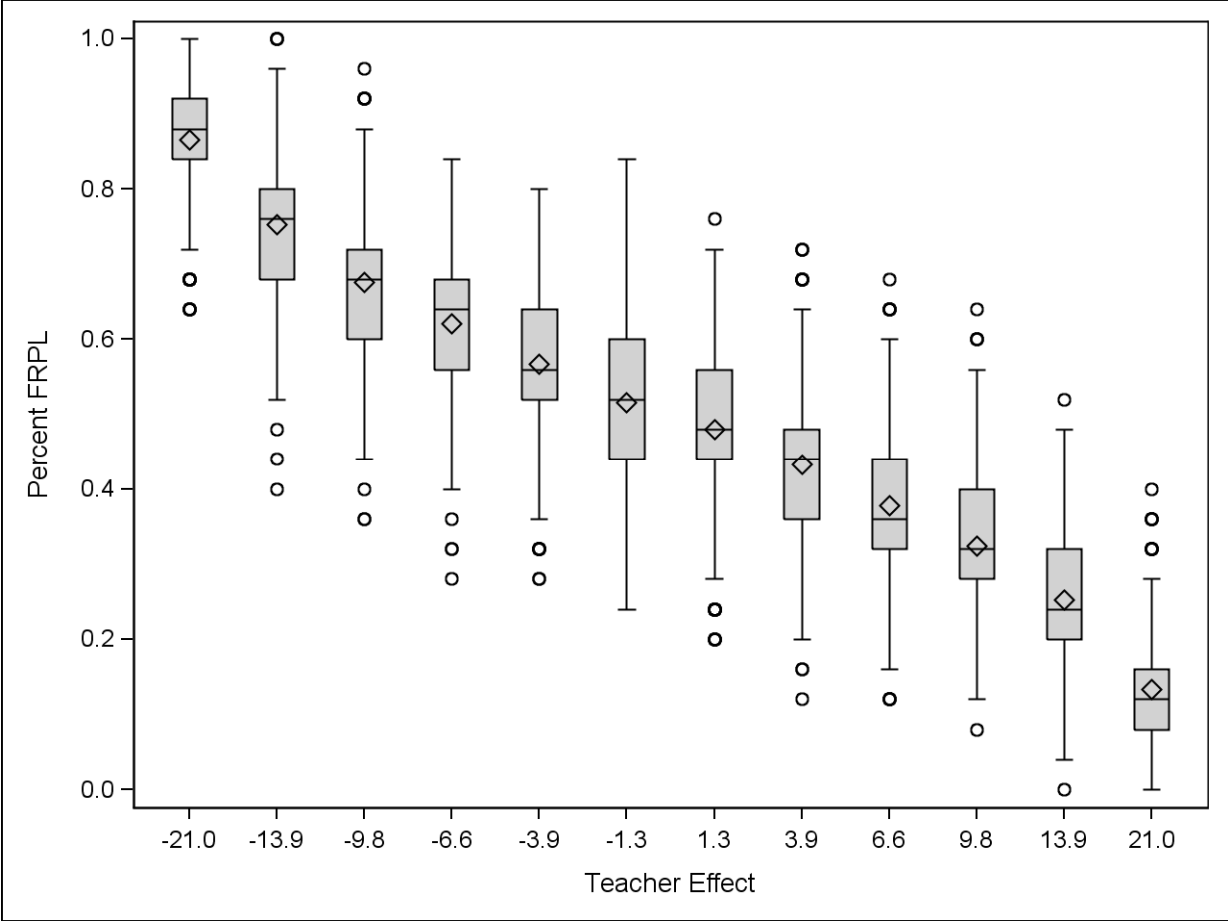


Figure 9. Classroom percent free/reduced-price lunch versus true teacher effect for 12 teachers over 1000 simulated data sets using Scenario #3.

Figure 10 shows, for Scenario #3, the MSEs for VAMs which do not employ any fairness adjustments. Bias is a much more prevalent problem for the ANCOVA models in Scenario #3 than was the case in Scenario #1 (compare Figure 4). Furthermore, in contrast to Scenario #1, bias is even more of a problem for the fixed-effects models than for the random-effects models! (In Scenario #1 the fixed-effects estimates were essentially unbiased.) However, including multiple pre-test scores, especially in the random-effects ANCOVA, effectively reduces the bias due to measurement error. The AOG and IV estimates, as in Scenario #1, avoid the bias due to measurement error, but continue to have more variance, and a larger MSE, than those ANCOVA models which have a sufficient number of pre-test scores (three or more) to dampen the bias due to measurement error.

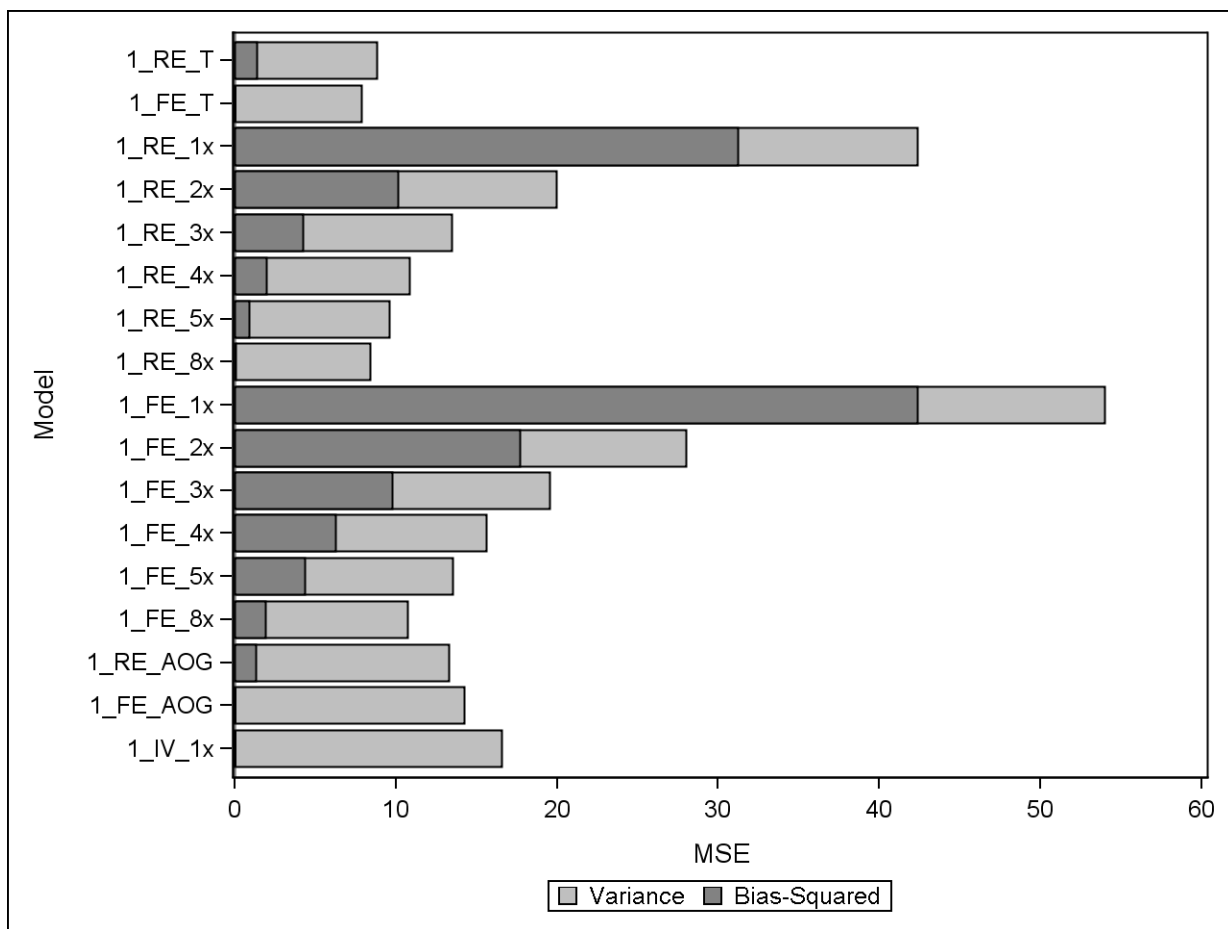


Figure 10. Mean square error from models that contained no fairness adjustments for 12 teachers in 1000 simulated data sets using Scenario #3.

Figure 11 shows, for Scenario #3, the MSEs for the random-effects ANCOVAs both with and without the fairness adjustments. Results are similar to Scenario #1 (Figure 5), but the bias produced by including the *classroom-level* adjustment is very much larger. Again, this bias exists even in the model with no measurement error in the pre-test score. Again, adding more pre-test scores to the model fails to remove the bias; in fact, in this scenario the amount of bias increases. As before, the bias is due to the fact that the true teacher effect is correlated with %FRPL; with %FRPL in the model, it is not the true teacher effects that are being estimated, but only that part of the true teacher effects that are uncorrelated with %FRPL.

Unlike the Scenario #1 results, including a *student-level* adjustment in the model (but not the classroom-level adjustment) does substantially reduce the bias compared to the unadjusted models with only one or two pre-test scores. But with three or more pre-test scores, the student-level adjusted model is negligibly better than the unadjusted model.

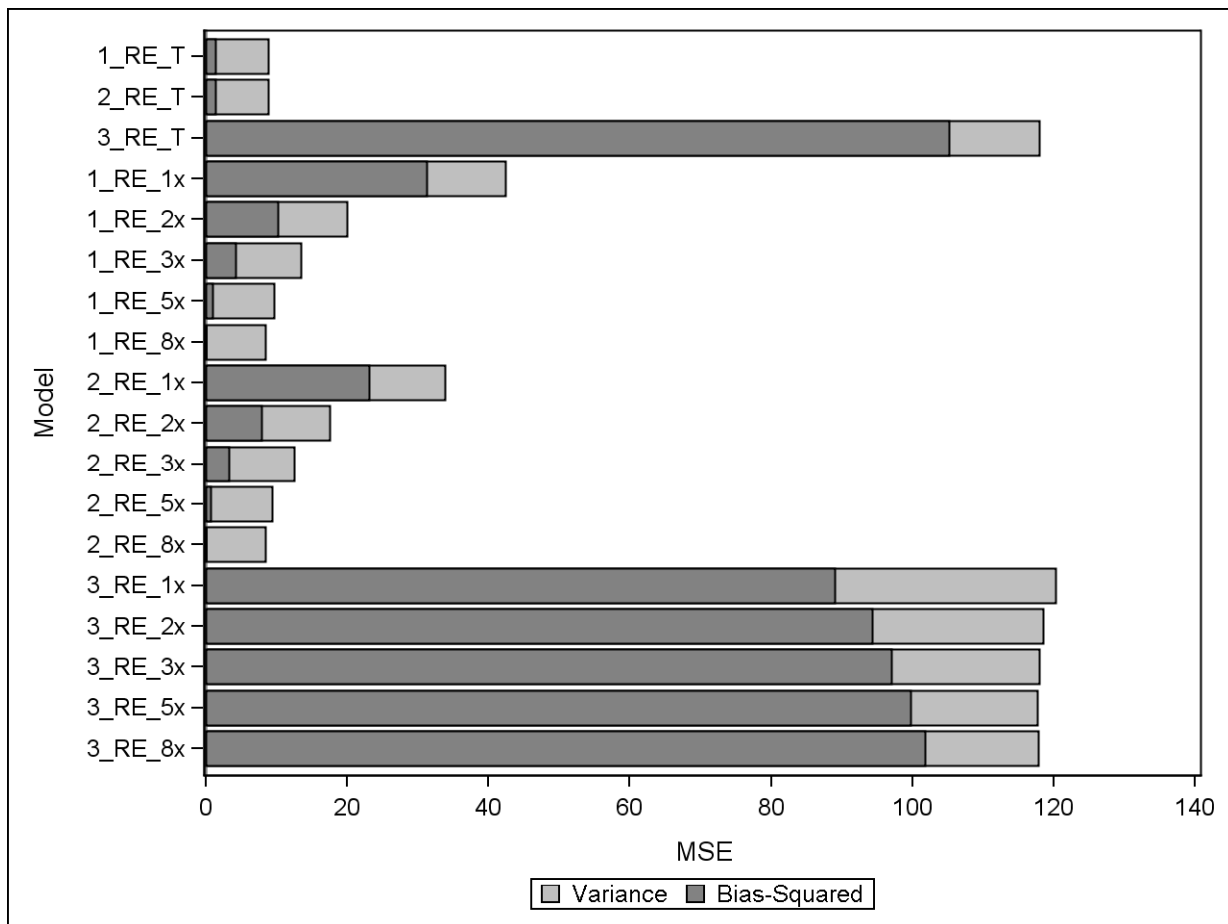


Figure 11. Mean square error from selected models, with and without fairness adjustments, for 12 teachers in 1000 simulated data sets using Scenario #3.

Conclusions

For the scenarios presented here, especially Scenario #3 in which the least capable teachers were disproportionately assigned to the highest poverty classrooms:

- adjusting for *classroom* %FRPL produced seriously biased teacher effect estimates;
- adjusting for *student* poverty level had little impact in models with a sufficient number of pre-test scores;
- using multiple pre-test scores (at least 3, preferably more) seemed to offer the best solution for the detrimental effects of measurement error;
- other solutions (analysis of gains, instrumental variables) were noisier, and in practice they will likely perform even worse than in these simulations (see “Caveats”).

Caveats: Missing Data. In order to focus more effectively on the problem of measurement error, which is the main focus of this paper, the simulations were kept as simple as possible – unrealistically simple in several respects. For example, as already mentioned, the pre-test scores were all generated by adding measurement error to the *same* true score rather than using different true scores for different pre-test subjects and grades. A more serious lack of realism is the lack of missing data in the simulations. In the real world, missing test scores are ubiquitous. Their impact on value-added modeling would include:

- Missing data increase the instability (variance) of estimates from all models.
- When data are not missing at random, as is usually the case, missing data increase the bias of estimates from all models; this includes, in particular, analysis of gains and instrumental variables models.
- Including multiple pre-test scores in ANCOVA models can still be effective in counteracting the effects of measurement error, but only if special procedures are used to avoid omission of all students having missing test scores.
- A multivariate, longitudinal, mixed model seems to offer the best alternative, not only for handling missing data and measurement error, but for other problems as well.

Wright(2004) used simulation to explore how missing data degrade estimates and how the multivariate, longitudinal approach leads to better estimates. McCaffrey, et al. (2003, 2004, 2008) cover many of the issues surrounding the choice of a VAM. Specifically, McCaffrey, et al. (2008) find that the multivariate approach outperforms the simpler approaches.

Other classroom level measures. In the simulations, a single socioeconomic variable (free/reduced-price lunch status) was used to demonstrate the consequences of adjusting for such variables at both the student and classroom level. Those who use hierarchical linear models for educational value-added modeling often include multiple demographic, socioeconomic, and other contextual variables at multiple levels (student, classroom, school, neighborhood, etc.). One classroom-level measure deserves particular mention. The simulations have shown that ANCOVA models using pre-test scores can effectively estimate value-added teacher effects (provided there are a sufficient number of pre-test scores to counteract the effects of measurement error). One might be tempted to think that it would be appropriate to include

classroom-level pre-test scores (e.g., the average pre-test score for the class) as well. In fact, this is commonly done in hierarchical linear modeling. However, a moment's thought should make it clear that, since a classroom average pre-test score would be highly correlated with classroom socioeconomic status (e.g., %FRPL), inclusion of such a classroom-level pre-test score will have the same detrimental effect on the estimated teacher effects. For confirmation, additional simulations, not reported here, were done using classroom average pre-test score rather than classroom %FRPL. As expected, the detrimental effects were similar.

The general principle regarding the use of classroom-level measures in ANCOVA models for value-added analysis is this: *Inclusion of any classroom-level measure which is correlated with the true teacher effects will produce biased estimates of those teacher effects.* Again, this is because the effects being estimated by the model are no longer the true teacher effects.

Parting thoughts. This final set of bullets is intended to focus attention on the most important conclusions of this paper.

- Do not blindly include demographic/socioeconomic variables in VAMs, especially at the classroom level, since this may produce serious bias.
- Use multiple pre-test scores (at least 3, preferably more) in ANCOVA models to counteract the effects of measurement error.
- In choosing a VAM, accuracy and stability trump transparency.

References

- McCaffrey, D. F., Han, B., and Lockwood, J. R. (2008). From Data to Bonuses: A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of Their Students' Progress. Paper presented at the 2008 national conference on Performance Incentives: Their Growing Impact on American K-12 Education, Peabody College of Vanderbilt University, Nashville, TN, February 28-29, 2008. Working paper available online at http://www.performanceincentives.org/data/files/directory/ConferencePapersNews/McCaffrey_et_al_2008.pdf.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., and Hamilton, L. (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics*, vol. 29, no. 1, pp. 67-101.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., and Hamilton, L. S. (2003). Evaluating Value-Added Models for Teacher Accountability. Santa Monica, CA: RAND. Available online at <http://www.rand.org/pubs/monographs/MG158>.
- Sanders, W. L., Saxton, A. M., and Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A Quantitative, Outcomes-Based Approach to Educational Assessment. In J. Millman (ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* (Chapter 13, pp. 137-162). Thousand Oaks, CA: Corwin Press. Available online at <http://www.sas.com/govedu/edu/sanderssaxtonhorn.pdf>.
- Wright, S. P. (2004). Advantages of a Multivariate Longitudinal Approach to Educational Value-Added Assessment Without Imputation. Paper presented at CREATE's 2004 National

Evaluation Institute, Colorado Springs, CO, July 8-10, 2004. Available online at <http://www.wmich.edu/evalctr/create/2004/Wright-NEI04.pdf> .