Carnegie Foundation for the Advancement of Teaching

WHAT DO WE KNOW ABOUT USING VALUE-ADDED TO COMPARE TEACHERS WHO WORK IN DIFFERENT SCHOOLS?

STEPHEN W. RAUDENBUSH
UNIVERSITY OF CHICAGO

CARNEGIE KNOWLEDGE NETWORK
What We Know Series:
Value-Added Methods and Applications

ATIL ADVANCING TEACHING - IMPROVING LEARNING

HIGHLIGHTS

- Bias may arise when comparing the value-added scores of teachers who work in different schools.
- Some schools are more effective than others by virtue of their favorable resources, leadership, or organization; we can expect that teachers of similar skill will perform better in these more effective schools.
- Some schools have better contextual conditions than others, providing students with more positive influences – peers who benefit from safe neighborhoods and strong community support. These conditions may facilitate instruction, thus will tend to increase a teacher's value-added score.
- Value-added models statistically control for student background and previously demonstrated student ability. But these controls tend to be ineffective, and possibly even misleading, when we compare teachers whose classrooms vary greatly by these factors.
- There are several methods for checking the sensitivity of value-added scores to school variation in contextual conditions and student backgrounds.
- If value-added scores are sensitive to these factors, we can revise the analysis to ensure that the classrooms being compared are similar on measures of student background and school composition, thus reducing the risk of bias.

INTRODUCTION

This brief considers the problem of using value-added scores to compare teachers who work in different schools. My focus is on whether such comparisons can be regarded as fair, or, in statistical language, "unbiased." An unbiased measure does not systematically favor teachers because of the backgrounds of the students they are assigned to teach, nor does it favor teachers working in resource-rich classrooms or schools. A key caveat: A measure that is *unbiased* does not mean the measure is *accurate.* An unbiased measure could be imprecise – thus inaccurate – if, for example, it is based on a small sample of students or on a test with too few items. I will not consider the issue of statistical precision here, having considered it in a previous brief.[1] This brief focuses strictly on the bias that may arise when comparing the value-added scores of teachers who work in different schools.

CHALLENGES THAT ARISE IN COMPARING TEACHERS WHO WORK IN DIFFERENT SCHOOLS

In a previous brief, Goldhaber and Theobold showed that how teachers rank on value-added can depend strongly on whether those teachers are compared to colleagues working in the same school or to teachers working in different schools.[2] This discrepancy by itself does not mean that between-school comparisons are biased, or that comparisons between teachers working in the same school are unbiased. However, previous literature identifies three unique challenges that arise in comparing teachers who work in different schools, and each brings a risk of bias.

First, some schools are more effective than others by virtue of their favorable resources, leadership, or organization. We can expect that teachers of similar skill will perform better in these more effective schools. Second, some schools have more favorable contextual conditions than others, providing a student with more favorable peers – those who benefit from strong community support and neighborhood safety. These contextual conditions may facilitate instruction, thus tending to increase a teacher's value-added score. Third, value-added models use statistical controls to make allowances for students' backgrounds and abilities. These controls tend to be ineffective, and possibly misleading, when we compare teachers whose classrooms vary greatly in the prior ability or other characteristics of their students. This problem can be particularly acute when we compare teachers in different schools serving very different populations of students. It can also arise when we compare teachers who work in the same school but who serve very different sub-populations, as with teachers in high schools in which students are tracked by ability.[3]

After describing each of these challenges, I consider ways to check the sensitivity of value-added scores to variations in schools' contextual conditions and students' background. If value-added scores are sensitive to these factors, we can revise the analysis to ensure that the classrooms being compared are similar on measures of student background and school composition, thus reducing the risk of bias. In this revised analysis, the aim is to compare teachers who work with similar students in similar schools. While policymakers may debate the utility of such comparisons, such comparisons are better supported by the available data than are comparisons between teachers who work with very different subsets of students and under different conditions. Thus they are less vulnerable to bias.

1. Variation in school effectiveness.

Emerging evidence suggests that some schools are more effective than others in managing resources, creating cultures for learning, and providing instructional support. An effectively organized school may provide benefits to all teachers working in that school. If so, teachers assigned to such schools will look better on value-added measures than will teachers who work in less effective schools.

Social scientists have for years debated whether schools vary substantially in their effectiveness, and if so, why. In his landmark 1966 report "Equality of Educational Opportunity," sociologist James S. Coleman suggested that socio-economic segregation of schools contributed to variation in learning but that factors such as facilities and spending mattered little.[4] The Coleman Report cast a long shadow over the proposition that giving schools more resources would improve education. However, Coleman's and other early studies were based on cross-sectional data. Early value-added modeling of schools based on longitudinal data suggested that students with similar backgrounds experience very different growth rates depending on the schools they attend.[5] These and more recent longitudinal studies raised the question of whether the internal life of schools, and in particular differences in leadership and collegial support, are more important than student composition in promoting learning.[6]

Recent randomized experiments provide compelling evidence that schools have highly varied effects. These studies capitalize on the fact that new charter schools are often oversubscribed: more students apply than can be admitted. By law, applicants to these charter schools are offered admission on the basis of a randomized lottery. Researchers are now following the outcomes of winners and losers of these lotteries. A study of randomized lotteries in 36 charter schools found that being admitted to a charter school made little difference in outcomes, on average. However, the *variation* in the impact of being so assigned was substantial. This result gives strong causal evidence not that charter schools *per*

se are particularly effective, but that some schools are substantially more effective than others.[7] Other researchers developed a model to predict this variation and found that five policies, including "frequent feedback to teachers, the use of data to guide instruction, high-dosage tutoring, increased instructional time, and high expectations," explain approximately 50 percent of the variation in school effectiveness.[8] Leaders in effective charter schools take pains to ensure that teachers follow school-wide procedures and norms. These randomized studies corroborate work showing that effective school leadership, professional work communities, and even school safety during a base year predict changes in the value that a school adds to learning.[9]

Separating the contribution of school leadership and resources from the average contribution of teacher skill is challenging, however. A critic of the review above might reasonably argue that what makes a school effective is nothing more than the average skill level of its teachers. However, several recent studies provide evidence against this criticism. In two of these studies, experimenters randomly assigned whole schools to innovative school-wide instructional curricula, revealing substantial positive effects on student learning. In these cases, the teaching force remained stable, yet the introduction of a new school-wide curriculum created added value that cannot be attributed simply to the aggregate quality of the teaching force.[10] Another recent study followed the value-added scores of teachers as they moved from one school to another. This study provided evidence that a teacher's value-added score will tend to improve when that teacher moves to a school in which other teachers have high value-added scores. This evidence suggests that teachers learn from high-skill peers.[11] Teacher collaboration and peer learning may also augment the impact of school-level factors such as the coherence of the curriculum, the availability of instructional materials, and the length of the school day or year.[12] In sum, comparisons between the value-added scores of teachers who work in different schools confound teacher skill and school effectiveness. Current value-added technology provides no means by which to separate these influences, a fact that defines a challenge to future research. Thus we have good reason to suspect that school effectiveness biases comparisons of the value-added scores of teachers working in different schools.

2. Variation in peers.

Within a district, schools tend to serve quite different sets of students. High-achieving students tend to be clustered in schools in which peers are highly motivated, parents are committed to the success of the school, and the surrounding neighborhood is safe.[13] These are presumably favorable conditions for instruction. Relatedly, sociological research has established that teachers tend to calibrate the content and pacing of their instruction according to the average prior achievement of their students,[14] implying that these well-prepared students will learn faster in classes with high-ability peers. Moreover, there is evidence that teachers themselves believe they are more effective when teaching higher-ability students than when teaching lower-ability students.[15] We have good reason, then, to think that favorable peer composition facilitates school and teaching effectiveness. The problem at hand is that, in principle, standard value-added models cannot isolate the impact of school organization or teacher expertise if those factors are correlated with peer motivation, parent commitment, neighborhood safety, and other local conditions.[16] The reason is that although value-added models may include measures of peer composition such as average prior ability or average family socioeconomic status, the value-added of the teacher or school is unobserved. If the peer composition and value-added are correlated – as suggested by the research – we have no way of isolating value-added.[17] Indeed, an attempt to control for peer composition when estimating value-added may introduce extra bias into value-added scores.[18]

The connection between peer composition and instructional effectiveness likely plays out very differently at the elementary and secondary levels. Elementary schools draw students from local neighborhoods that tend to be quite segregated with respect to family income and race/ethnicity. As a result, elementary schools tend to be comparatively internally homogeneous with respect to student background. In contrast, large, comprehensive public secondary schools draw students from multiple elementary schools and thus tend to be internally more heterogeneous than are elementary schools. In response, high schools typically assign students to classrooms based on their perceived ability. This process of "tracking" can generate large differences among classrooms.[19] Harris considers the special problems that arise in studying teacher value-added within secondary schools that use tracking,[20] and reasons that problems of peer composition are more pronounced within high schools than they are within elementary schools. These effects are likely to be particularly important when we compare teachers who work in different schools, even elementary schools.

3. "Common support" and statistical adjustment for student background.

The problem of statistically adjusting for student background is distinct from that of isolating peer effects or differences in school effectiveness. Even if peer effects were negligible and all schools were equally effective, classroom composition could bias value-added. For example, two classrooms taught by equally skilled teachers might display different learning rates simply because one classroom had more able students. Random assignment of students to teachers would solve this problem, but random assignment doesn't happen in practice, so statisticians have invented adjustments to control for student background factors that predict future achievement. The problem is that, in general, we cannot rely on standard methods of statistical adjustment to work well when the backgrounds of children attending different classrooms vary substantially. Statisticians call this the failure of "common support," and it is more likely to occur when we compare teachers in different elementary schools than when we compare teachers in the same elementary school. This problem is also likely to arise in comparisons of teachers within high schools that use tracking.

To understand how statistical adjustments work, consider the comparisons of two teachers, A and B, who teach students having different prior average achievement. Statistical adjustment is based on a statistical model that predicts how Teacher A's students would have done if they had been assigned to Teacher B and how Teacher B's students would have done if they had been assigned to Teacher A. If the statistical model is based on good background information, such as prior test scores that strongly predict future test scores, this may work very well. In particular, if the two groups of students overlap considerably in their background, the data will have good predictive information about how each set of students would do in either classroom. This is a case in which the two groups have good "common support" for the model.

However, if these two distributions do not overlap, we have a problem – a failure of common support. Suppose, in the worst case, that all of Teacher A's students have higher prior achievement than any of Teacher B's students. In this case the data have no information about how A's students would do in B's class or how well B's students would do in A's class; there is simply no valid comparison group for either teacher. In this case, the value-added score will simply be an extrapolation – a guess based on the analyst's belief about whether the relationship between prior background and future test score is linear or, in some known way, non-linear.[21] In essence, the comparison of value-added scores is not based on the data but is entirely based on the analyst's assumptions about the model. This extreme case – no overlap in the two distributions – is unlikely to arise in practice. The key point is that the smaller the

overlap in the two distributions of prior background, the less information the data can provide about the comparative effectiveness of the two teachers. This problem is especially acute if there is any reason to suspect that teachers are differentially effective for students of different backgrounds. In that case, it is essential to compare teachers serving similar children to draw any valid causal conclusions.[22] Central to our discussion is the fact that, at the elementary school level, a lack of common support is more likely to occur in comparisons between teachers working in different schools than in comparisons between teachers working in the same school. A failure of common support is also likely to arise in comparisons among high school teachers working in the same school if that school tracks students on the basis of ability. It is possible and useful to check comparability in any value-added analysis, a topic to which I return in the concluding section.

Exacerbating any failure of common support is the inherent uncertainty in achievement test scores. Suppose one school draw students from the upper end of the achievement distribution while another school draws students from the lower end. Suppose that students in these schools make, on average, a five-point gain in achievement. To say that these gains represent an equivalent amount of learning requires unwarranted assumptions about the achievement test. We can speak confidently of our measurements of things like time and distance, but measuring gains in cognitive skill is much more difficult. On some tests, low-achieving students can easily make comparatively large score gains simply because the test has a large number of easy items. In this case, teachers working in low-scoring schools would produce inflated value-added scores. On other tests, it will be comparatively easy for high-achieving students to make large gains, biasing value-added in favor of teachers working in those schools. Simply put, our current technology for constructing tests does not allow us to make strong claims about the relative gains of students who start from very different places in the achievement distribution. Our testing technology works much better when are comparing gains made by students of similar background and prior skill.

In sum, efforts to compare teachers working in schools that serve children from widely varied backgrounds are vulnerable to bias. In a typical school district about 15-20 percent of the total variation in students' average incoming achievement lies between schools.[23] This means that students attending a high-achieving school will tend to score around 1.5 standard deviations higher, on average, than students attending a low-achieving school. To compare teachers working in such a wide range of schools may include a risk of failure of common support, as well as introduce substantial peer effects, increasing the risk of bias when we compare the value-added scores of teachers working in different schools.

EMPIRICAL EVIDENCE OF BIAS

As we have seen, researchers have found that teachers rank quite differently when they are compared to colleagues in the same school than when they are compared to teachers in other schools. This finding leads us to predict that comparing teachers in different schools produces more bias than does comparing teachers in the same school. However, we need to see whether empirical evidence supports such predictions. What do we know about the magnitude of bias that arises in each case?

Comparing teachers within schools

Several randomized experiments lend support to the idea that value-added scores are approximately unbiased. In one large-scale study, students were randomly assigned to teachers. No statistical

adjustments were needed to correct for bias, so a comparison between classrooms within the same school can be regarded as a comparison of "true value-added." The analysts found that the variation in such value-added scores was quite large, indicating that teachers do, in fact, vary substantially in their effectiveness. Moreover, the variation in value-added in this experiment was similar in magnitude to the variation in value-added typically found in conventional non-randomized value-added analysis.[24] Two more recent experiments assessed the bias of value-added scores more directly. Value-added scores were computed conventionally for one year using standard methods of statistical adjustment. During the next year, pairs of teachers within schools were randomly assigned to student rosters, enabling the researchers to compute unbiased value-added scores with no statistical adjustment. Comparisons between the conventional and experimental value-added scores provided some evidence that the conventional scores were approximately unbiased.[25]

All of these encouraging studies used experimental evidence to investigate the bias of using conventional value-added scores to compare teachers working in the same school. A key question is whether such encouraging results can be found for comparing teachers in different schools. Here the research base is sparser. Perhaps the most important study of this type followed 2.5 million children in grades 3-8 into adulthood. The researchers found that students assigned to high value-added teachers had higher educational attainment, earnings, and wealth as adults.[26] The researchers tested the potential bias of the value-added scores in two ways. First, using parental tax data, they compared students who had experienced high value-added teachers to those who had experienced low value-added teachers. They found no association between teacher value-added and these socioeconomic measures, providing evidence against the claim that unmeasured family characteristics had biased the value-added scores. Secondly, they compared a school's average achievement before and after a "high-value-added" teacher had left the school. They found that when a school loses a teacher with a high value-added score, the school's achievement tends to decrease. This finding is important, because it supports the claim that the value-added score has causal content, and it supports the finding that within-school differences in teacher value-added reflect real differences in effectiveness. Nevertheless, the teacher value-added scores computed in this study, despite reflecting differences in teacher effectiveness, are vulnerable to bias. At least part of the variation in teacher value-added may have reflected differences in school organizational effectiveness or differences in community and peer effects. The researchers did not consider the possibility that high-value-added teachers work in schools that are effectively managed, and that attending such an effective school is key to students' future success. Nor did they test whether omitted school-level variables were associated with teacher value-added. Instead, the authors assumed implicitly that any differences between schools must reflect differences in the individual skill of teachers working in those schools. A re-analysis of these data could estimate the contribution of school value-added to adult success to assess whether an alternative explanation based on school differences is plausible.

CONCLUSIONS AND RECOMMENDATIONS

This brief has considered sources of potential bias when we use value-added scores to compare teachers working in different schools. A growing body of evidence suggests that schools can vary substantially in their effectiveness, potentially inflating the value-added scores of teachers assigned to effective schools. Schools also vary in contextual conditions such as parental expectations, neighborhood safety, and peer influences that may directly support learning or that may contribute to school and teacher effectiveness. Moreover, schools vary substantially in the backgrounds of the students they serve, and conventional

statistical methods tend to break down when we compare teachers serving very different subsets of students.

Although we know that there is great potential for bias when we compute value-added scores for teachers working in different schools, we do not yet know the extent to which this potential is actually realized. Nor do we have conclusive evidence regarding the extent to which teachers are particularly effective or ineffective for particular kinds of students. This uncertainty poses a challenge to those who wish to interpret teacher value-added scores. One way to address this challenge is to check the sensitivity of value-added results to variations across schools in effectiveness and student composition. Several approaches come to mind.

First, following Goldhaber and Theobold,[27] one can compute value-added scores two ways: by comparing teachers within schools and by comparing teachers without regard to their school assignment. If the rankings are consistent, we have little reason to favor the within-school comparisons. My colleagues and I did this, not with value-added scores but with student perceptions of teaching quality using seven indicators of teacher effectiveness based on the Tripod Survey Assessments of Ronald Ferguson from Harvard University.[28] Our results, taken from data on a large urban district, were highly convergent. Correlations between indicators computed in these two different ways ranged from .91 to .96 across the seven dimensions, with a mean of .94. This was not surprising, because school differences accounted for little of the variation in Tripod: Only 2-7 percent of the variation in these indicators lay between schools. Given these convergent results, there is little reason to believe that school differences were adding extra bias in these indicators based on student perceptions. (As a further check, I recommend computing the percentile rank of teachers under the two procedures; that is, comparing teachers who work in the same school and comparing teachers without reference to the school in which they work).[29] These findings are not surprising: Students are likely to base their perceptions of teaching quality on experiences with teachers in the same or similar schools, which likely explains why the fraction of variation between schools in student perceptions is comparatively small. In contrast, as mentioned, value-added scores tend to vary considerably between schools.

What if results are not convergent? A sensible strategy is to divide schools into subsets that serve rather similar students. One might then use value-added scores (or other indicators) to compare teachers who work in the same *subset* of schools. To check the sensitivity of those measures, one might again check convergence between two sets of estimates: those that compare teachers within schools and those that compare teachers working in different schools but in the same subset of schools. If these are convergent, we can assume that the decision to compare teachers working in different schools (within the same *subset* of schools) has not contributed bias to the value-added scores.[30]

Shavelson and Wiley suggest a refined version of this approach: For each school, select a subset of schools that match that school in terms of student characteristics. Call this the reference set for a particular school. Ideally, each school is located in the middle of its reference set with respect to the distribution of expected achievement gains.[31]

An additional check is to compute the “contextual effect” of student composition, as is common in research in educational sociology. This indicator, along with a measure of variation between schools in school-mean prior achievement, is diagnostic of the bias that plausibly arises from school heterogeneity.[32] The same procedure can be used to assess whether classrooms within a school are too heterogeneous in student background to support unbiased value-added. I would recommend using this

procedure (see appendix), particularly in the case of secondary schools that track students to classrooms on the basis of ability.

These sensitivity checks, and the possible stratification of teachers into sub-groups serving similar students, complicate value-added analysis and may not be congruent with policymakers' wish to compare all teachers in a district. However, these steps may win the approval of teachers who want to be sure that comparisons between themselves and other teachers are free of bias. Moreover, I recommend this modified approach as scientifically responsible, for it limits us to answering questions that our data can actually answer.

APPENDIX

Raudenbush and Willms (1995) wrote down a fairly general linear model for value-added, then used a very simple model for illustration. The illustration is informative for this brief. We have an outcome Y_{ij}, a linear function of a covariate X_{ij} defined for each student *i* within classroom *j* $(i = 1,...,n_j;\ j = 1,...J)$. We can orthogonally decompose the regression within and between classrooms as

$$Y_{ij} = \beta_0 + \beta_w(X_{ij} - \overline{X}._j) + \beta_b(\overline{X}._j - \overline{X}..) + U_j + e_{ij} \qquad \text{(A1)}$$

where $\overline{X}._j$ is the classroom mean of the covariate, $\overline{X}..$ is the overall ("grand") mean; β_w characterizes the association between X and Y within classrooms; and β_b characterizes the association between and the classroom mean of the covariate and the classroom mean outcome. By re-centering the covariate, we obtain a re-parameterization of (A1) known in the sociology literature as the "contextual effects model."

$$Y_{ij} = \beta_0 + \beta_w(X_{ij} - \overline{X}..) + \beta_c(\overline{X}._j - \overline{X}..) + U_j + e_{ij} \qquad \text{(A2)}$$

Here $\beta_c = \beta_b - \beta_w$ is the "contextual effect" (the association between $\overline{X}._j$ and the mean outcome controlling for person-level X_{ij}); U_j and e_{ij} are random effects at the classroom and student levels respectively; and we assume

Ignorable student-level error $E(e_{ij} \mid X) = E(e) = 0$ Assumption (i)

We allow for confounding between, but do not, however, assume $E(U_j \mid \overline{X}) = E(U_j) = 0$. Let us now average the outcome within classrooms to obtain

$$\overline{Y}._j = \beta_0 + \beta_w(\overline{X}._j - \overline{X}..) + \beta_c(\overline{X}._j - \overline{X}..) + U_j + \overline{e}._j \qquad \text{(A3)}$$

Type A and Type B Effects

Now we can define two effects. The first is the Type A effect, of interest to parents choosing classrooms, that is

$$\beta_c(\overline{X}._j - \overline{X}..) + U_j$$

This is composed of two components. The first is the contextual component $\beta_c(\overline{X}._j - \overline{X}..)$, attributable to the composition of the classroom with respect to the covariate $\overline{X}._j$. This might represent the impact of peer composition, the quality of the resources available to the teacher, or the supportiveness of the parents. The second is the random effect U_j, attributable to the effectiveness of teacher *j*'s practice.

The Type B effect, of interest to teacher evaluators, is thus the random effect U_j itself. That is,

$$B_j \equiv U_j$$

We shall assume that this key estimand – the true value added for teacher j, has zero mean and variance

$$Var(U_j) = \tau^2 \qquad \text{(A4)}$$

Identification

Under Assumption (i) in connection with A2, we can identify β_w, and hence the Type A effect by subtraction

$$\begin{aligned} A_j &\approx \overline{Y}._j - \beta_0 - \beta_w(\overline{X}._j - \overline{X}..) = \beta_c(\overline{X}._j - \overline{X}..) + U_j + \overline{e}._j \\ &= A_j + \overline{e}._j \end{aligned} \qquad \text{(A5)}$$

Equation (A5) is what economists would call the estimated "fixed effect."

Under an additional strong identification assumption $Cov(U_j, \overline{X}._j) = 0$, we can identify the Type B effect by subtraction as

$$\begin{aligned} B_j &\approx \overline{Y}._j - \beta_0 - \beta_w(\overline{X}._j - \overline{X}..) - \beta_c(\overline{X}._j - \overline{X}..) \\ &= \overline{Y}._j - \beta_0 - \beta_b(\overline{X}._j - \overline{X}..) = U_j + \overline{e}._j \end{aligned} \qquad \text{(A6)}$$

However, if $Cov(U_j, \overline{X}._j) \neq 0$, we will estimate β_b with bias, that is, the least squares estimator $\hat{\beta}_b$ will have expectation

$$E(\hat{\beta}_b) = \beta_b + \beta_{u\bar{x}} \qquad \text{(A7)}$$

where $\beta_{u\bar{x}} = Cov(U_j, \overline{X}._j)/\sigma^2_{\bar{x}}$ and $\sigma^2_{\bar{x}}$ is the variance of the sample mean of the covariate $\overline{X}._j$ (Note this variance will be non-zero even if n_j is infinite, assuming classrooms vary in composition). As a result,

if $Cov(U_j, \overline{X}._j) \neq 0$, (6) does not approximate the true type B effect but instead estimates a quantity we might called the "prima facie" effect, that is

$$B_{PFj} = U_j - \beta_{u\bar{x}}(\overline{X}._j - \overline{X}..) \quad \text{(A8)}$$

A Way Forward

We can put a *bound* on how biased our analysis will be using the following result Raudenbush and Willms (1995):

$$Var(B_{PFj}) \leq Var(B_j) \leq Var(A_j)$$
$$or \quad \text{(A9)}$$
$$(1-\rho^2_{u\bar{x}})\tau^2 \leq \tau^2 \leq (\beta_c + \beta_{u\bar{x}})^2\sigma^2_{\bar{x}} + (1-\rho^2_{u\bar{x}})\tau^2$$

We can consistently estimate $Var(B_{PFj})$ and, if Assumption (i) is plausible, we can consistently estimate $Var(A_j)$. We can therefore put bounds around the variance of the true value-added effects. Moreover, it follows from (A9) that the variance of the bias is

$$0 \leq \rho^2_{u\bar{x}}\tau^2 = (\beta_c + \beta_{u\bar{x}})^2\sigma^2_{\bar{x}} \quad \text{(A10)}$$

If (A5) is reasonable, we can estimate the upper bound as

$$(\beta_c + \beta_{u\bar{x}})^2\sigma^2_{\bar{x}} \approx \hat{\beta}^2_c\sigma^2_{\bar{x}} \quad \text{(A11)}$$

If the estimated contextual effect is null or if the classroom mean $\overline{X}._j$ varies little, we need not worry. If their product is large, our biases are large in magnitude. We might then back away from the universal "Type B" effect and instead block on classrooms with similar $\overline{X}._j$. This would narrow the scope of the reference population for teacher *j* to include other teachers whose classrooms have similar composition.

Note schools are not represented in our model. Hence the Type B effect is in fact the sum of the teacher value-added plus the effectiveness of the school.

ENDNOTES

[1] For a detailed discussion of precision, see Raudenbush, S.W. and M. Jean. Carnegie Knowledge Network, "How Should Educator Interpret Value-Added Scores," (2012). http://www.carnegieknowledgenetwork.org/briefs/value-added/interpreting-value-added/

[2] Goldhaber, D. and R. Theobold. Carnegie Knowledge Network, "Do Different Value-Added Models Tell Us the Same Things?" Last updated April 2013. (2012). http://www.carnegieknowledgenetwork.org/briefs/value-added/different-growth-models/

[3] See the discussion by Harris (2013).

[4] Coleman, et al., "Equality of educational opportunity: Summary report. Vol. 2." US Department of Health, Education, and Welfare, Office of Education, 1966.

[5] Bryk, Anthony S. and Stephen W. Raudenbush, "Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model," *American Journal of Education* (1988): 65-108.

[6] Lauen, Douglas Lee and S. Michael Gaddis, "Exposure to classroom poverty and test score achievement: Contextual effects of selection," *American Journal of Sociology* 118, no. 4, (2013): 943-979. http://www.stevenmichaelgaddis.com/Lauen%20and%20Gaddis%202012%20-%20Classroom%20Poverty%20AJS%20Forthcoming.pdf

[7] Gleason et al., "The Evaluation of Charter School Impacts: Final Report. NCEE 2010-4029." National Center for Education Evaluation and Regional Assistance (2010). http://www.mathematica-mpr.com/publications/pdfs/education/charter_school_impacts.pdf

[8] Dobbie, Will and Roland G. Fryer, Jr., "Getting beneath the veil of effective schools: Evidence from New York City. National Bureau of Economic Research, (Working paper #w17632, 2011). http://www.nber.org/papers/w17632

[9] Bryk et al., *Organizing schools for improvement: Lessons from Chicago*, University of Chicago Press, 2010.

[10] Borman et al., "Final reading outcomes of the national randomized field trial of Success for All." *American Educational Research Journal* 44, no. 3 (2007): 701-731.; Borman, G.D., M.M. Dowling, and C. Schneck, "A multisite cluster randomized field trial of Open Court Reading." *Educational Evaluation and Policy Analysis* 30, no. 4 (2008): 389-407.

[11] Jackson, Clement Kirabo and Elias Bruegmann, "Teaching students and teaching each other: The importance of peer learning for teachers." *American Economic Journal: Applied Economics* 1, no. 4 (2009): 85-108. http://digitalcommons.ilr.cornell.edu/workingpapers/77/

[12] Raudenbush, Stephen W. "Can School Improvement Reduce Racial Inequality?" Research on Schools, Neighborhoods, and Communities: Toward Civic Responsibility (2012): 233.

[13] Harding (2010) provides a vivid portrayal of these differences. Harding, David J., *Living the drama: Community, conflict, and culture among inner-city boys,* University of Chicago Press, 2010.

[14] Dreeben, Robert and Rebecca Bar, "Educational Policy and the Working of Schools" (1983); Gamoran, Adam, "Tracking and inequality: New directions for research and practice," The Routledge international handbook of the sociology of education (2010): 213-228.; Raudenbush, Stephen W., Brian Rowan, and Yuk Fai Cheong, "Higher order instructional goals in secondary schools: Class, teacher, and school influences," *American Educational Research Journal* 30, no. 3 (1993): 523-553.

[15] Raudenbush, Rowan, and Cheong (1992) compared the self-efficacy of a teacher when teaching high-ability high-school classes to the self-efficacy score of the same teacher when teaching low-ability high-school class. They found large differences, particularly in mathematics. Raudenbush, Stephen W., Brian Rowan, and Yuk Fai Cheong, "Contextual effects on the self-perceived efficacy of high school teachers," *Sociology of Education* (1992): 150-167.

[16] Raudenbush, Stephen W. and J. Douglas Willms "The estimation of school effects," *Journal of educational and behavioral statistics* 20, no. 4 (1995): 307-335.

[17] One might reason that a statistical model should estimate and remove the effect of peer composition, thus isolating value added. However, if the peers and value added are correlated, the failure to measure and control value added will cause bias in the estimation of peer composition. As a result, the estimate of value added based on removal of this biased peer composition effect will also be biased.

[18] McCaffrey, D.F. Carnegie Knowledge Network, "Do Value-added models level the playing field for teachers?" Last updated June 2013 (2012). http://www.carnegieknowledgenetwork.org/briefs/value-added/level-playing-field/

[19] See Gamoran's (2010) review.

[20] Harris, D. Carnegie Knowledge Network, "Does Value Added Work Better in Elementary than in Secondary Schools?" (2013). http://www.carnegieknowledgenetwork.org/briefs/value-added/grades/

[21] If the distribution of prior test scores in one classroom does not overlap with the distribution of prior test scores in the other classroom, a comparison between the two classrooms depends entirely on the functional form of the

statistical model and not at all on the data. In this case different functional forms – linear, quadratic, logarithmic, or exponential, for example – for the relationship between prior achievement and post achievement will yield different value-added scores; the data will provide no information about which functional form is preferable.

[22] Reardon and Raudenbush (2009) show that to compare the effectiveness of teachers who teach very different students requires another strong assumption, that teachers who are comparatively effective for one type of student are also comparatively effective for other types of students. This assumption is required unless value-added models include, for each teacher, separate estimates of value-added for subsets of students who vary in background. Most value-added models do not, and cannot, do so because we can assess a teacher's value-added only for the students she is assigned to teach. Also the amount of data needed multiplies rapidly. Teachers who work in affluent schools do not give us information about how effective they might be might teach low-income schools, for example. More research is needed on the extent to which teacher effectiveness varies for students of highly varied background. However, this problem – that some teachers may be more effective at working with some kinds of students than others – will not cause bias in value-added analysis that compare individuals teaching very similar students. Reardon, Sean F. and Stephen W. Raudenbush, "Assumptions of value-added models for estimating school effects." *Education* 4, no. 4 (2009): 492-519.

[23] Hedges, Larry V. and E.C. Hedberg "Intraclass correlations for planning group randomized experiments in rural education." *Journal of Research in Rural Education* 22, no. 10 (2007): 1-15.

[24] Nye, Barbara, Spyros Konstantopolous, and Larry V. Hedges "How large are teacher effects?" Educational *evaluation and policy analysis 26*, no. 3 (2004): 237-257. http://steinhardt.nyu.edu/scmsAdmin/uploads/002/834/127%20-%20Nye%20B%20%20Hedges%20L%20%20V%20%20%20Konstantopoulos%20S%20%20(2004).pdf

[25] Kane, Thomas J. and Douglas O. Staiger, "Estimating teacher impacts on student achievement: An experimental evaluation," National Bureau of Economic Research, (Working paper No. w14607, 2008). http://www.nber.org/papers/w14607; Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project," Bill & Melinda Gates Foundation (2013). http://www.rand.org/pubs/external_publications/EP50156.html

[26] Chetty, Raj, John N. Friedman, and Johan E. Rockoff, "The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood," National Bureau of Economic Research (Working paper No. w17699, 2011). http://www.nber.org/papers/w17699. This study is consistent with an earlier study showing that random assignment to effective kindergarten teachers produced favorable effects on educational attainment and adult economic status (Chetty et al., "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *The Quarterly Journal of Economics* 126, no. 4 (2011): 1593-1660).

[27] See note 2.

[28] Raudenbush, Stephen and Marshall Jean. Carnegie Knowledge Network, "How Should Educators Interpret Value-Added Scores?" (2012). http://www.carnegieknowledgenetwork.org/briefs/value-added/interpreting-value-added/

[29] To remove school differences, one adds school fixed effects to the conventional regression model. This is equivalent to a random effects analysis that centers all covariates around the school mean (Raudenbush, 2009). The latter analysis can readily be elaborated to study whether some teachers are better than others at teaching particular kinds of students.

[30] This strategy is not perfect, because it might be that such schools in the same subset vary in organizational effectiveness.

[31] Personal communication with Richard Shavelson and David Wiley.

[32] The "contextual effect" is not really a causal effect but rather the partial association between school-average prior achievement and student achievement controlling for all student background characteristics as in value added. Based on Raudenbush and Willms (1995), we can show that the squared contextual coefficient multiplied by the school mean prior achievement sets an upper bound for the variance of the biases of the value-added score. The key assumption is that the *within-school value added score* is unbiased. See also K. Castellano and S. Rabe-Hesketh (2013), *Composition, Context, and Endogeneity in School and Teacher Comparisons*, Berkely CA Working Paper: University of California at Berkeley.

AUTHOR

Stephen Raudenbush, Ed.D. is the Lewis-Sebring Distinguished Service Professor in the Department of Sociology, Professor at the Harris School of Public Policy Studies and is Chairman of the Committee on Education at the University of Chicago. He received an Ed.D. in Policy Analysis and Evaluation Research from Harvard University. He is a leading scholar on quantitative methods for studying child and youth development within social settings such as classrooms, schools, and neighborhoods. He is best known for his work on developing hierarchical linear modes, with broad applications in the design and analysis of longitudinal and multilevel research. He is currently studying the development of literacy and math skills in early childhood with implications for instruction, and methods for assessing school and classroom quality. He is a member of the American Academy of Arts and Sciences and the recipient of the American Educational Research Association Award for distinguished contributions to educational research.

ABOUT THE CARNEGIE KNOWLEDGE NETWORK

The Carnegie Foundation for the Advancement of Teaching has launched the Carnegie Knowledge Network, a resource that will provide impartial, authoritative, relevant, digestible, and current syntheses of the technical literature on value-added for K-12 teacher evaluation system designers. The Carnegie Knowledge Network integrates both technical knowledge and operational knowledge of teacher evaluation systems. The Foundation has brought together a distinguished group of researchers to form the *Carnegie Panel on Assessing Teaching to Improve Learning* to identify what is and is not known on the critical technical issues involved in measuring teaching effectiveness. Daniel Goldhaber, Douglas Harris, Susanna Loeb, Daniel McCaffrey, and Stephen Raudenbush have been selected to join the Carnegie Panel based on their demonstrated technical expertise in this area, their thoughtful stance toward the use of value-added methodologies, and their impartiality toward particular modeling strategies. The Carnegie Panel engaged a User Panel composed of K-12 field leaders directly involved in developing and implementing teacher evaluation systems, to assure relevance to their needs and accessibility for their use. This is the first set of knowledge briefs in a series of Carnegie Knowledge Network releases. Learn more at **carnegieknowledgenetwork.org**.

CITATION

Stephen Raudenbush. Carnegie Knowledge Network, "What Do We Know About Using Value-Added to Compare Teachers Who Work in Different Schools." Last modified August 2013. URL = <http://carnegieknowledgenetwork.org/briefs/value-added/different-schools/>

Carnegie Foundation
for the Advancement of Teaching

Carnegie Foundation for the Advancement of Teaching
51 Vista Lane
Stanford, California 94305
650-566-5100

Carnegie Foundation for the Advancement of Teaching seeks to vitalize more productive research and development in education. We bring scholars, practitioners, innovators, designers, and developers together to solve practical problems of schooling that diminish our nation's ability to educate all students well. We are committed to developing networks of ideas, expertise, and action aimed at improving teaching and learning and strengthening the institutions in which this occurs. Our core belief is that much more can be accomplished together than even the best of us can accomplish alone.
www.carnegiefoundation.org

We invite you to explore our website, where you will find resources relevant to our programs and publications as well as current information about our Board of Directors, funders, and staff.

Knowledge Brief 10
AUGUST 2013

carnegieknowledgenetwork.org

Funded through a cooperative agreement with the Institute for Education Science. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.