School Finance 101

Data and thoughts on public and private school funding in the U.S.

If it's not valid, reliability doesn't matter so much! More on VAM-ing & SGP-ing Teacher Dismissal

By Bruce Baker
Posted on April 28, 2012

This post includes a few more preliminary musings regarding the use of value-added measures and student growth percentiles for teacher evaluation, specifically for making high-stakes decisions, and especially in those cases where new statutes and regulations mandate rigid use/heavy emphasis on these measures, as I discussed in the previous post.

========

The recent release of New York City teacher value-added estimates to several media outlets stimulated much discussion about standard errors and statistical noise found in estimates of teacher effectiveness derived from the city's value-added model. But lost in that discussion was any emphasis on whether the predicted value-added measures were valid estimates of teacher effects to begin with. That is, did they actually represent what they were intended to represent – the teacher's influence on a true measure of student achievement, or learning growth while under that teacher's tutelage. As framed in teacher evaluation legislation, that measure is typically characterized as "student achievement growth," and it is assumed that one can measure the influence of the teacher on "student achievement growth" in a particular content domain.

A brief note on the semantics versus the statistics and measurement in evaluation and accountability is in order.

At issue are policies involving teacher "evaluation" and more specifically *evaluation of teacher effectiveness*, where in cases of dismissal the evaluation objective is to identify particularly ineffective teachers.

In order to "evaluate" (assess, appraise, estimate) a teacher's effectiveness with respect to student growth, one must be able to "infer" (deduce, conjecture, surmise…) that the teacher affected or could have affected that student growth. That is, for example, given one year's bad rating, the teacher had sufficient information to understand how to improve her rating in the following year. Further, one must choose measures that provide some basis for such inference.

Inference and *attribution* (ascription, credit, designation) are not separable when evaluating teacher effectiveness. To make an *inference* about teacher effectiveness based on student achievement growth, one must *attribute* responsibility for that growth to the teacher.

In some cases, proponents of student growth percentiles alter their wording [in a truly annoying & dreadfully superficial way] for general public appeal to argue that:

1. SGPs are a measure of student achievement growth.
2. Student achievement growth is a primary objective of schooling.
3. Therefore, teachers and schools should obviously be held *accountable* for student achievement growth.

Where *accountable* is a synonym for *responsible,* to the extent that SGPs were designed to separate the measurement of student growth from *attribution of responsibility for it*, then SGPs are also invalid on their face for holding teachers *accountable.* For a teacher to be *accountable* for that growth it must be *attributable* to them and one must be using a method which permits such *inference*.

Allow me to reiterate this quote from the authors of SGP:

"The development of the Student Growth Percentile methodology was guided by Rubin et al's (2004) admonition that VAM quantities are, at best, descriptive measures." (Betebenner, Wenning & Briggs, 2011)

I will save for another day a discussion of the nuanced differences between statistical causation and inference and causation and inference as might be evaluated more broadly in the context of litigation over determination of teacher effectiveness. The big problem in the current context, as I have explained in my previous post, is created by legislative attempts to attach strict timelines, absolute weights and precise classifications to data that simply cannot be applied in this way.

Major Validity Concerns

We identify[at least] 3 categories of significant compromises to inference and attribution and therefore accountability for student achievement growth:

1. The value-added estimate (or SGP) was influenced by something other than the teacher alone
2. The value-added (or SGP) estimate given one assessment of the teacher's content domain produces a different rating than the value-added estimate given a different assessment tool
3. The value-added estimate (or SGP) is compromised by missing data and/or student mobility, disrupting the *link* between teacher and students. [the actual data link required for attribution]

The first major issue compromising *attribution of responsibility* for or inference regarding teacher effectiveness based on student growth is that some other factor or set of factors actually caused the student achievement growth or lack thereof. A particularly bothersome feature of many value-added models is that they rely on annual testing data. That is, student achievement growth is measured from April or May in one year to April or May in the next, where the school year runs from September to mid or late June. As such, for example, the 4th grade teacher is

assigned a rating based on children who attended her class from September to April (testing time), or about 7 months, where 2.5 months were spent doing any variety of other things, and another 2.5 months were spent with their prior grade teacher. Let alone the different access to resources each child has during their after school and weekend hours during the 7 months over which they have contact with their teacher of record.

Students with different access to summer and out-of-school time resources may not be randomly assigned across teachers within a given school or across schools within a district. And students who had prior year teachers who may have *checked out* versus the teacher who delved into the subsequent year's curriculum during the post-testing month of the prior year may also not be randomly distributed. All of these factors go unobserved and unmeasured in the calculation of a teacher's effectiveness, potentially severely compromising the validity of a teacher's effectiveness estimate. Summer learning varies widely across students by economic backgrounds (Alexander, Entwisle & Olsen, 2001) Further, in the recent Gates MET Studies (2010), the authors found: "The norm sample results imply that students improve their reading comprehension scores just as much (or more) between April and October as between October and April in the following grade. Scores may be rising as kids mature and get more practice outside of school." (p.)

Numerous authors have conducted analyses revealing the problems of *omitted variables bias* and the non-random sorting of students across classrooms (Rothstein, 2011, 2010, 2009, Briggs & Domingue, 2011, Ballou et al., 2012). In short, some value-added models are better than others, in that by including additional explanatory measures, the models seem to correct for at least some biases. Omitted variables bias is where any given teacher's predicted value is influenced partly by factors other than the teacher herself. That is, the estimate is higher or lower than it should be, because some other factor has influenced the estimate. Unfortunately, one can never really know if there are still additional factors that might be used to correct for that bias. Many such factors are simply unobservable. Others may be measurable and observable but are simply unavailable, or poorly measured in the data. While there are some methods which can substantially reduce the influence of unobservables on teacher effect estimates, those methods can typically only be applied to a very small subset of teachers within very large data sets.[2] In a recent conference paper, Ballou and colleagues evaluated the role of omitted variables bias in value-added models and the potential effects on personnel decisions. They concluded:

"In this paper, we consider the impact of omitted variables on teachers' value-added estimates, and whether commonly used single-equation or two-stage estimates are preferable when possibly important covariates are not available for inclusion in the value-added model. The findings indicate that these modeling choices can significantly influence outcomes for individual teachers, particularly those in the tails of the performance distribution who are most likely to be targeted by high-stakes policies." (Ballou et al., 2012)

A related problem is the extent to which such biases may appear to be a wash, on the whole, across large data sets, but where specific circumstances or omitted variables may have rather severe effects on predicted values for specific teachers. To reiterate, these are not merely issues of instability or error. These are issues of whether the models are estimating the teacher's effect on student outcomes, or the effect of something else on student outcomes. Teachers should not

be dismissed for factors beyond their control. Further, statutes and regulations should not require that principals dismiss teachers or revoke their tenure in those cases where the principal understands intuitively that the teacher's rating was compromised by some other cause. [as would be the case under the TEACHNJ Act]

Other factors which severely compromise inference and attribution, and thus validity, include the fact that the measured value-added gains of a teacher's peers – or team members working with the same students – may be correlated, either because of unmeasured attributes of the students or because of spillover effects of working alongside more effective colleagues (one may never know) (Koedel, 2009, Jackson & Bruegmann, 2009). Further, there may simply be differences across classrooms or school settings that remain correlated with effectiveness ratings that simply were not fully captured by the statistical models.

Significant evidence of bias plagued the value-added model estimated for the Los Angeles Times in 2010, including significant patterns of racial disparities in teacher ratings both by the race of the student served and by the race of the teachers (see Green, Baker and Oluwole, 2012). These model biases raise the possibility that Title VII disparate impact claims might also be filed by teachers dismissed on the basis of their value-added estimates. Additional analyses of the data, including richer models using additional variables mitigated substantial portions of the bias in the LA Times models (Briggs & Domingue, 2010).

A handful of studies have also found that teacher ratings vary significantly, even for the same subject area, if different assessments of that subject are used. If a teacher is broadly responsible for effectively teaching in their subject area, and not the specific content of any one test, different results from different tests raise additional validity concerns. Which test better represents the teacher's responsibilities? [must we specify which test counts/matters/represents those responsibilities in teacher contracts?] If more than one, in what proportions? If results from different tests completely counterbalance, how is one to determine the teacher's true effectiveness in their subject area? Using data on two different assessments used in Houston Independent School District, Corcoran and Jennings (2010) find:

[A]mong those who ranked in the top category (5) on the TAKS reading test, more than 17 percent ranked among the lowest two categories on the Stanford test. Similarly, more than 15 percent of the lowest value-added teachers on the TAKS were in the highest two categories on the Stanford.

The Gates Foundation Measures of Effective Teaching Project also evaluated consistency of teacher ratings produced on different assessments of mathematics achievement. In a review of the Gates findings, Rothstein (2010) explained:

The data suggest that more than 20% of teachers in the bottom quarter of the state test math distribution (and more than 30% of those in the bottom quarter for ELA) are in the top half of the alternative assessment distribution.(p. 5)

And:

In other words, teacher evaluations based on observed state test outcomes are only slightly better than coin tosses at identifying teachers whose students perform unusually well or badly on assessments of conceptual understanding.(p. 5)

Finally, student mobility, missing data, and algorithms for accounting for that missing data can severely compromise inferences regarding teacher effectiveness. Corcoran (2010) explains that the extent of missing data can be quite large and can vary by student type:

Because of high rates of student mobility in this [Houston] population (in addition to test exemption and absenteeism), the percentage of students who have both a current and prior year test score – a prerequisite for value-added – is even lower (see Figure 6). Among all grade four to six students in HISD, only 66 percent had both of these scores, a fraction that falls to 62 percent for Black students, 47 percent for ESL students, and 41 percent for recent immigrants." (Corcoran, 2010, p.20- 21)

Thus, many teacher effectiveness ratings would be based on significantly incomplete information, and further, the extent to which that information is incomplete would be highly dependent on the types of students served by the teacher.

One statistical resolution to this problem is imputation. In effect, imputation creates pre-test or post-test scores for those students who weren't there. One approach is to use the average score for students who were there, or more precisely for otherwise similar students who were there. On its face imputation is problematic when it comes to attribution of responsibility for student outcomes to the teacher, if some of those outcomes are statistically generated for students who were not even there. But not using imputation may lead to estimates of effectiveness that are severely biased, especially when there is so much missing data. Howard Wainer (2011) esteemed statistician and measurement expert formerly with Educational Testing Service (ETS) explains somewhat mockingly how teachers might game imputation of missing data by sending all of their best students on a field trip during fall testing days, and then, in the name of fairness, sending the weakest students on a field trip during spring testing days.[3] Clearly, in such a case of gaming, the predicted value-added assigned to the teacher as a function of the average scores of low performing students at the beginning of the year (while their high performing classmates were on their trip), and high performing ones at the end of the year (while their low performing classmates were on their trip), would not be correctly attributed to the teacher's actual teaching effectiveness, though it might be attributable to the teacher's ability to game the system.

In short, validity concerns are at least as great as reliability concerns, if not greater. If a measure is simply not valid, it really doesn't matter whether it is reliable or not.

If a measure cannot be used to validly infer teacher effectiveness, cannot be used to attribute responsibility for student achievement growth to the teacher, then that measure is highly suspect as a basis for high stakes decisions making when evaluating teacher (or teaching) effectiveness or for teacher and school accountability systems more generally.

References & Additional Readings

Alexander, K.L, Entwisle, D.R., Olsen, L.S. (2001) Schools, Achievement and Inequality: A Seasonal Perspective. *Educational Evaluation and Policy Analysis* 23 (2) 171-191

Ballou, D., Mokher, C.G., Cavaluzzo, L. (2012) Using Value-Added Assessment for Personnel Decisions: How Omitted Variables and Model Specification Influence Teachers' Outcomes. Annual Meeting of the Association for Education Finance and Policy. Boston, MA. http://aefpweb.org/sites/default/files/webform/AEFP-Using%20VAM%20for%20personnel%20decisions_02-29-12.docx

Ballou, D. (2012). Review of "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-long-term-impacts

Baker, E.L., Barton, P.E., Darling-Hammong, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., Shepard, L.A. (2010) Problems with the Use of Student Test Scores to Evaluate Teachers. Washington, DC: Economic Policy Institute. http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf

Betebenner, D., Wenning, R.J., Briggs, D.C. (2011) Student Growth Percentiles and Shoe Leather. http://www.ednewscolorado.org/2011/09/13/24400-student-growth-percentiles-and-shoe-leather

Boyd, D.J., Lankford, H., Loeb, S., & Wyckoff, J.H. (July, 2010). *Teacher layoffs: An empirical illustration of seniority vs. measures of effectiveness*. Brief 12. National Center for Evaluation of Longitudinal Data in Education Research. Washington, DC: The Urban Institute.

Briggs, D., Betebenner, D., (2009) Is student achievement scale dependent? Paper presented at the invited symposium Measuring and Evaluating Changes in Student Achievement: A Conversation about Technical and Conceptual Issues at the annual meeting of the National Council for Measurement in Education, San Diego, CA, April 14, 2009. http://dirwww.colorado.edu/education/faculty/derekbriggs/Docs/Briggs_Weeks_Is%20Growth%20in%20Student%20Achievement%20Scale%20Dependent.pdf

Briggs, D. & Domingue, B. (2011). *Due Diligence and the Evaluation of Teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the* Los Angeles Times. Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/publication/due-diligence.

Budden, R. (2010) *How Effective Are Los Angeles Elementary Teachers and Schools?,* Aug. 2010, *available at* http://www.latimes.com/media/acrobat/2010-08/55538493.pdf.

Braun, H, Chudowsky, N, & Koenig, J (eds). (2010) *Getting value out of value-added. Report of a Workshop*. Washington, DC: National Research Council, National Academies Press.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service. Retrieved February, 27, 2008.

Chetty, R., Friedman, J., Rockoff, J. (2011) The Long Term Impacts of Teachers: Teacher Value Added and Student outcomes in Adulthood. NBER Working Paper # 17699 http://www.nber.org/papers/w17699

Clotfelter, C., Ladd, H.F., Vigdor, J. (2005) Who Teaches Whom? Race and the distribution of Novice Teachers. *Economics of Education Review* 24 (4) 377-392

Clotfelter, C., Glennie, E. Ladd, H., & Vigdor, J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics 92*, 1352-70.

Corcoran, S.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. Annenberg Institute for School Reform. http://annenberginstitute.org/pdf/valueaddedreport.pdf

Corcoran, S.P. (2011) Presentation at the Institute for Research on Poverty Summer Workshop: Teacher Effectiveness on High- and Low-Stakes Tests (Apr. 10, 2011), *available at* https://files.nyu.edu/sc129/public/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf.

Corcoran, Sean P., Jennifer L. Jennings, and Andrew A. Beveridge. 2010. "Teacher Effectiveness on High- and Low-Stakes Tests." Paper presented at the Institute for Research on Poverty summer workshop, Madison, WI.

D.C. Pub. Sch., IMPACT Guidebooks (2011), *available at* http://dcps.dc.gov/portal/site/DCPS/menuitem.06de50edb2b17a932c69621014f62010/?vgnextoid=b00b64505ddc3210VgnVCM1000007e6f0201RCRD.

Education Trust (2011) Fact Sheet- Teacher Quality. Washington, DC. http://www.edtrust.org/sites/edtrust.org/files/Ed%20Trust%20Facts%20on%20Teacher%20Equity_0.pdf

Hanushek, E.A., Rivkin, S.G., (2010) Presentation for the American Economic Association: Generalizations about Using Value-Added Measures of Teacher Quality 8 (Jan. 3-5, 2010), *available at* http://www.utdallas.edu/research/tsp-erc/pdf/jrnl_hanushek_rivkin_2010_teacher_quality.pdf

Working with Teachers to Develop Fair and Reliable Measures of Effective Teaching. MET Project White Paper. Seattle, Washington: Bill & Melinda Gates Foundation, 1. Retrieved December 16, 2010, from http://www.metproject.org/downloads/met-framing-paper.pdf.

Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project. MET Project Research Paper. Seattle, Washington: Bill & Melinda Gates Foundation. Retrieved December 16, 2010, from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf.

Jackson, C.K., Bruegmann, E. (2009) Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers. *American Economic Journal: Applied Economics 1(4): 85–108*

Kane, T., Staiger, D., (2008) *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*. NBER Working Paper #16407 http://www.nber.org/papers/w14607

Koedel, C. (2009) An Empirical Analysis of Teacher Spillover Effects in Secondary School. 28 (6) 682-692

Koedel, C., & Betts, J. R. (2009). *Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique*. Working Paper.

Jacob, B. & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*. 26(1), 101-36.

Sass, T.R., (2008) The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy. National Center for Analysis of Longitudinal Data in Educational Research. Policy Brief #4. http://eric.ed.gov/PDFS/ED508273.pdf

McCaffrey, D. F., Lockwood, J. R, Koretz, & Hamilton, L. (2003). Evaluating value-added models for teacher accountability. RAND Research Report prepared for the Carnegie Corporation.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. Journal of Educational and Behavioral Statistics, 29(1), 67.

Rothstein, J. (2011). Review of "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project." Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, 4(4), 537–571.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*, 125(1), 175–214.

Sanders, W. L., Saxton, A. M., & 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 measure? (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Sanders, William L., Rivers, June C., 1996. Cumulative and residual effects of teachers on future student academic achievement. Knoxville: University of Tennessee Value- Added Research and Assessment Center.

Sass, T.R. (2008) The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy. *Urban Institute* http://www.urban.org/UploadedPDF/1001266_stabilityofvalue.pdf

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

McCaffrey, D.F., Lockwood, J.R. (2011) Missing Data in Value Added Modeling of Teacher Effects. *Annals of Applied Statistics* 5 (2A) 773-797

Reardon, S. F. & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. Education Finance and Policy, 4(4), 492–519.

Rubin, D. B., Stuart, E. A., and Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. Journal of Educational and Behavioral Statistics, 29(1):103–116.

Schochet, P.Z., Chiang, H.S. (2010) Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains. Institute for Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf.