

Learning from Recent Advances in Measuring Teacher Effectiveness

August 9, 2012 Washington, DC

Two-page Briefs Prepared by:

**Damian Betebenner, National Center for the
Improvement of Educational Assessment**

Henry Braun, Boston College

Sean Corcoran, New York University

Linda Darling-Hammond, Stanford University

John Friedman, Harvard University

Daniel Goldhaber, University of Washington

Andrew Ho, Harvard Graduate School of Education

Thomas Kane, Harvard University

Helen Ladd, Duke University

Robert Pianta, University of Virginia

Jonah Rockoff, Columbia Business School

Jesse Rothstein, University of California, Berkeley

Damian Betebenner, National Center for the Improvement of Educational Assessment

“Validity of the Use of Large-Scale Assessments for Teacher Evaluation”

Background

Recent policy initiatives including Race to the Top and the ESEA waiver program have rapidly advanced the use of large-scale assessment results in analyses directed toward the evaluation of teacher quality. The vast majority of research conducted on the use of large-scale assessment outcomes for teacher evaluation has focused on the technical characteristics of the indicators (often referred to as teacher “value-added” effects). Though of great importance, the technical characteristics of value-added/student growth statistics goes only part way toward addressing larger issues associated with the development and implementation of complex indicator systems. The trajectory in the development of value-added/student growth and its use for teacher evaluation is not unlike the early development of large-scale assessments where technical considerations dominated the discussions about the tests and the ultimate (ab)use of the results was not as well mapped out in advance.

In a recent piece entitled *Problems with the use of test scores to evaluate teachers* issued by the Economic Policy Institute (EPI, 2010), the distinguished authors document many of the lessons learned over the past two decades with the use of large-scale assessment. The authors highlight the importance of creating a comprehensive evaluation system that supports increased teacher performance. They point out that the use of large-scale test results is not orthogonal to the development of such an evaluation system, but fear that an over-reliance on such indicators will have many unintended negative consequences.


Currently, my primary interests center on the development of complex indicator systems that support, in some cases, teacher evaluation systems. The biggest challenge in this development is to strike the perfect balance between issues associated with technical adequacy (e.g., reliability/precision, accuracy/bias, validity) and the creation of a system that has the potential to increase the efficacy of the education system. To that end, my interests are in exploring real world design issues associated with implementing complex indicator systems for reporting and accountability purposes.

Criteria for evaluating evaluation systems

What would be considered a good teacher evaluation system—a component of which is a value-added/student growth metric? Put more colloquially, how would we know a good use of value-added/student growth for educator evaluation if we ran into one? In his 2008 presentation at the Reidy Interactive Lecture Series (Braun, 2008), Professor Henry Braun argued that the ultimate criterion by which to judge the validity of an accountability system is by the consequences.

Assessment practices and systems of accountability are systemically valid if they generate useful information and constructive responses that support one or more policy goals (Access, Quality, Equity, Efficiency) within an education system, without causing undue deterioration with respect to other goals.

For indicator systems to yield the outcomes we intend—that is, be systematically valid—it is critical to map out these outcomes together with all the intermediate steps that would lead to their realization at the beginning. This map that details how data derived from value-



added/student growth analyses is turned into knowledge and, ultimately, education changing actions is sometimes referred to as a “theory of action.” Very often there is no detailed theory of action specified with regard to value-added/student growth data use; and perhaps worse, there has yet to be written (to my knowledge) a great paper/book on data use in education detailing how this resource can be utilized to transform education.

References

Braun, H.I. (2008). *Vicissitudes of the validators*. Presentation made at the 2008 Reidy Interactive Lecture Series, Portsmouth, NH, September, 2008.

Economic Policy Institute. (2010). *Problems with the use of student test scores to evaluate teachers* (Tech. Rep.). Washington, DC. (Downloaded August 30, 2010 from http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf)

Henry Braun, Boston College

“Learning from Recent Advances in Measuring Teacher Effectiveness”

Research on the use of VAMs and related models (e.g., Student Growth Percentiles) as indicators of teacher effectiveness continues at a furious pace, with no general consensus on the horizon. Perhaps the closest the community has come to a consensus is documented in two recent National Research Council publications: *Getting Value Out of Value-added* (2010) and *Incentives and Test-based Accountability in Education* (2011). The former documents the proceedings of a workshop and offers a compendium of the measurement and analytic issues that must be addressed in providing support for the high-stakes use of VAM results. Most participants had serious concerns about such use, particularly if VAM scores were heavily weighted, but some favored moving forward in view of the state of teacher evaluation today (typically dismal) and the poor operating characteristics of current practice-based indicators. The latter publication summarizes an extensive review of the literature on test-based incentives and concludes that there is little empirical support for the expectation that the implementation of test-based accountability will result in substantial improvements in learning. Its recommendations center on the need for systematic research on the design of incentive systems. My view on the need for caution and careful analysis of ongoing and forthcoming implementations of test-based accountability is consistent with the thrust of these reports.

Turning to peer-reviewed research, the blockbuster on the positive side is the article by Chetty et al. (2011) that reports on a monumental analysis of value-added that makes a strong case for (i) value-added scores being relatively unbiased, (ii) that teacher’s value-added scores are predictive of their future students’ “growth,” and (iii) that a student’s exposure to a single high value-added teacher is statistically associated with various positive distal outcomes. Although some commentators have appropriately argued against over-interpretation of these findings (Ballou, 2012; Harris, 2012), this study offers strong evidence that, in the aggregate and at the individual level, value-added scores contain useful information. A less well-known paper (Sanders et al., 2009) demonstrates that teachers with high value-added obtained in lower poverty schools tend to maintain their relative ranking when they move to higher poverty schools. Although the number of such teachers was small (54) and the result is somewhat at odds with others’ findings on the volatility of value-added scores (see below), it does add to the argument that value-added scores may not be as dependent on local context as some contend. Clearly, more such studies are called for.

On the negative side, the blockbusters are papers by J. Rothstein (2009, 2010) that demonstrate the presence of substantial bias in value-added scores, in part due to the dynamic allocation of students to teachers, as well as the effects of reversion to the mean. Rothstein also provides estimates of the magnitude of the bias. In this regard, it is important to note that the estimates of uncertainty that usually accompany the value-added scores are derived from the models that generate the scores and, hence, do not incorporate the contributions of bias (which could be substantial, as Rothstein’s work shows) to the mean squared error. The article by Reardon and Raudenbush (2009) presents a comprehensive review of the plausibility of the various assumptions that undergird all VAMs. It makes for sobering reading.

That said, there have been some responses to Rothstein. Koedel and Betts (2010) replicate Rothstein’s results with different data but assert that averaging value-added scores over three cohorts mitigates the bias. Though it is mentioned only in passing in both articles, the analyses are done for within-school estimates of teacher effects. District-wide estimates are likely to suffer from greater problems of bias. Goldhaber and Chaplin (2012) take a different tack in undercutting Rothstein’s argument. However, their conclusion seems to be highly dependent on

the correctness of the model—an assumption that cannot be taken for granted. For me, the upshot is a reinforcement of the need for caution.

The article by Newton et al. (2010) offers further evidence on the sensitivity of the value-added results to the choice of model and their volatility over time. In their critique of Buddin's analysis of teachers' value-added in LAUSD, Briggs and Dominigue (2011) also demonstrate this sensitivity. These and related concerns are also documented by Corcoran (2010). It is noteworthy that the latter two reports employ data from large urban districts. Some current work (Braun et al., 2011) shows that two quite different approaches to evaluating value-added yield reasonably consistent results, even with a single cohort of data. Thus, the evidence on volatility is somewhat mixed. (Note that investigations of consistency across models do not directly address the issue of bias in the value-added estimates.)

Although most researchers have focused on VAMs, at least as many states have adopted Student Growth Percentiles (SGPs) as the test-based indicator of choice. SGPs have some technical advantages over VAMs—principally that they require no metric assumptions about the score scales. On the other hand, conditioning on only one or two test scores in computing a conditional percentile results in considerable uncertainty. In general, as with VAMs, the use of SGPs raises issues of drawing accurate causal inferences from observational data.

Hill et al. (2011) argue that obtaining value-added estimates (of teachers' relative effectiveness) is a type of measurement process and, hence, should be subject to the same guidelines for good practice that every measurement activity should follow. This approach is further developed in Braun (2013). The bottom line, at least for me, is that test-based indicators do have a role to play in educator accountability but that we must invest more in data QC and that extreme rankings derived from a value-added analysis should be carefully audited before they are incorporated into an overall evaluation. Protocols for such audits would have to be designed and implemented. We also need to build the infrastructure both to collect evidence regarding other valued outcomes of schooling and to monitor unintended consequences of the accountability system.

In that regard, we need to pay more attention to the design of the accountability system *qua* system, recognizing that test-based indicators constitute only one component of a complex set of components that interact with each other, as well as with the larger education system. The degree of success of the system in achieving its goals depends on the quality of each component and how well the different components work with each other. An interesting effort in this direction can be found in McCaffrey et al. (2009).

References

- Ballou, D. (2012). Review of *The Long-term Impacts of Teachers*. Retrieved 3/12/12 from <http://nepc.colorado.edu/thinktank/review-long-term-impacts>.
- Braun, H., Qu, Y., & Trapani, C. (2011). Evaluating School Effectiveness: Robustness to Scale and Model. [Under revision]
- Braun, H. (forthcoming, 2013). Magical Thinking and the Use of Value-added Models for Educator Accountability. *Applied Measurement in Education*.
- Briggs, D., & Dominigue, B. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of LAUSD teachers by the LA Times. Boulder, CO: National Education Policy Center.
- Chetty, R., Friedman, J.N., and Rockoff, J.E. (December, 2011). The Long-term Impacts of Teachers: Teacher Value-added and Student Outcomes in Adulthood. National Bureau of

- Economic Research. [See also “short version” in *Education Next* (2012, summer) and accompanying comments.]
- Corcoran, S. (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. Providence, RI: Annenberg Institute for School Reform.
- Goldhaber, D., and Chaplin, D. (January, 2012). Assessing the “Rothstein Test”: Does it Really Show Teacher Value-added Models are Biased? CALDER Working Paper #71.
- Harris, D. (2012). Implications for policy are not so clear. *Education Next*, 12(3).
- Hill, H., Kapitula, L., & Umland, K. (May 9, 2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*.
- Koedel, C., & Betts, J.R. (2010). Does student sorting invalidate VAMs of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*.
- McCaffrey et al. (2009). Turning student test scores into teacher compensation systems. In M.G. Springer (Ed.), *Performance Incentives: Their growing impact on American K-12 education* (chapter 6, pp. 113-148). Washington, DC: Brookings Institution Press.
- National Research Council. (2010). Getting value out of value-added. H. Braun, N. Chudowsky and J. Koenig (Eds.). Washington, DC: National Academies Press.
- National Research Council. (2011). Incentives and test-based accountability in education. M. Hout and S. Elliott (Eds.). Washington, DC: National Academies Press.
- Newton, X.A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, 18(23).
- Reardon, S.F., & Raudenbush, S.W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, 4(4).
- Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, 4(4).
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement,” *Quarterly Journal of Economics*, 125(1).
- Sanders, W.L., Wright, S.P., & Langevin, W.E. (2009). The performance of highly effective teachers in different school environments. In M.G. Springer (Ed.), *Performance Incentives: Their growing impact on American K-12 education* (chapter 8, pp. 171-189g). Washington, DC: Brookings Institution Press.

Sean Corcoran, New York University

Research on teacher labor markets and factors associated with teaching effectiveness has developed at an unprecedented pace over the past 10-15 years. This work has vastly advanced our understanding of, for example, the effectiveness of teachers entering through different pathways (Boyd et al., 2006; Harris & Sass, 2011), hiring practices (Loeb et al., 2011), retention (Goldhaber et al., 2011), teacher sorting (Hanushek et al., 2004), and the long-run impact of teachers (Chetty et al., 2011) to name only a few examples. The Measures of Effective Teaching (MET) project, currently underway, is one of the most ambitious efforts in history to systematically identify effective classroom teaching practices. This literature would not be where it is today without advances in value-added measurement and the careful linking of student-level achievement data to teachers over time.

That value-added measures (VAMs) have proven useful in research, however, does not imply they will be useful as on-the-job performance measures. Whereas empirical research relies on large samples of teachers and students over many years, personnel decisions are made in real time, often with limited information about any individual teacher. In research, statistical inferences about any one teacher are unimportant; we are only interested in relationships that hold on average. In practice, inferences about an individual teacher can end a career. If value-added measures are to be part of an evaluation system, one needs a relatively high level of confidence in their causal attribution and precision; I have serious concerns about both.

To be sure, performance evaluation in the teaching profession is sorely lacking, and collective bargaining agreements have often made it difficult to remove demonstrably underperforming teachers. And I do believe VAMs do contain some useful information. But in my view, the potential for these measures to vastly improve classroom performance and the quality of the teaching workforce is overblown. Below, I highlight some of my specific concerns about the use of value-added measures in practice.

- 1. There is not a single dimension of value-added.** Academic exercises simulating improvements in the workforce that would result from terminating the bottom x percent of teachers make the assumption teachers can be ranked according to their underlying “value-added,” a single dimension of effectiveness. Under this theory, policies that dismiss the “low value-added” teachers would produce big gains in performance. While in the abstract this makes sense, in practice value-added is measured using a specific test. Value-added scores can be generated for multiple subjects, and in some cases on multiple tests. Teachers receive VAM scores in reading and math, and presumably additional subject tests are forthcoming. Although VAMs across subjects are correlated, given their sampling variability, the likelihood a teacher will be identified as “ineffective” across several or all subjects, given a particular cut score or decision rule, is extremely low (for one example, see Corcoran, Jennings, & Beveridge, 2011). Without the ability to dismiss generically “low value-added” teachers, districts will need to devise decision rules that appropriately take into account performance across multiple subjects and tests. How one should do this optimally is not immediately obvious.
- 2. What ultimately matters is how value-added scores are mapped to ratings.** While there has been much discussion about the estimation and statistical properties of VAMs, few researchers or practitioners argue that specific point estimates should be the focus of performance evaluation. In other words, we rarely can make meaningful distinctions between teachers in, say, the 55th and 70th percentiles of value-added, and it makes little sense to invest much energy in drawing such comparisons. Instead, states are using VAMs to assign teachers to broad effectiveness categories. For example, in New York, teachers will be assigned to one of four groups: highly effective, effective, developing, and ineffective.

Consequences result for teachers in the ineffective, and to a lesser extent, developing categories. In practice, then, decision rules mapping VAMs to performance categories ultimately govern their practical impact.

The use of such decision rules raises several issues in my mind. First, given value-added scores are norm-referenced, category assignment is ultimately a political decision about what fraction of teachers will be considered ineffective each year. Second, rules will need to take into account both point estimates and uncertainty. New York has recognized this in its new system, requiring ineffective teachers to have both low VAMs and a minimum degree of precision; if VAMs are used in personnel decisions, such caution is appropriate. Third, these rules will need to address differences across subjects and tests (see #1). Fourth, by design, decision rules are focused almost entirely on the tail of the distribution. This is consistent with many researchers' view that VAMs are more informative in the tails, and that most teachers in the middle cannot be differentiated. But this implies a VAM-based evaluation system will only provide meaningful information to a small fraction of teachers—most will simply learn they cannot be differentiated from the average. Fifth, by identifying those in the tails, the system is more likely to flag teachers whose student population differs most from the norm (e.g., non-English speaking special education students, gifted students near the ceiling). Finally, the use of an appropriately designed decision rule begs the question of what VAMs can add beyond alternative modes of evaluation. If we design a system that conservatively identifies the worst of the worst (or best of the best), would these teachers not be so identified otherwise?

3. **VAMs are potentially biased.** Whether VAMs are an unbiased measure of the causal impact of an individual teacher on student outcomes has been a topic of much research and discussion (Kane & Staiger, 2008; Rothstein, 2010). I will leave it to others to elaborate. However, two potential sources of bias give me the most concern.
 - a. **Separating teacher from school effects.** The use of school effects in value-added models is common in research, but not in practice. This is for good reason; as Gordon, Kane, & Staiger (2006) have argued, teacher quality is unevenly distributed across schools, and within-school comparisons (via school fixed effects) ignore this important source of variation. At the same time, there is reason to believe that there are school-level inputs associated with teacher effectiveness, including school leadership, disciplinary policies, supplementary instruction (e.g., coaches), parent involvement, expenditures, and so on. Without school effects, teacher effects are easily confounded with these other inputs. The problem is compounded when teachers are compared across districts with quite different settings (and resources). Both approaches have their problems, and it is not a priori clear either is correct.
 - b. **Isolating teacher effects when there are multiple teachers.** When students receive instruction from multiple teachers during the course of a day or school year (e.g., middle and high school), the assumption that student achievement gains in one subject are entirely attributable to the teacher of that class subject is tenuous. I have not yet seen a statistical model that convincingly separates the contributions of multiple, simultaneous teacher inputs, and there is evidence of spillovers between teachers in the same school (e.g., Jackson & Bruegmann, 2009).
4. **Missing data.** In order for a student to contribute to a teacher's VAM estimate, he or she must have been tested in the prior year. In settings with mobile students (New York City and Houston, for example), a significant share are missing prior year test scores. Little is known as to whether the omission of these students from the teacher's value-added estimate biases that estimate enough to make a difference. But a performance evaluation system that does not credit teachers' work with students not tested in the prior year lacks face validity, and potentially creates perverse incentives to focus on students who count toward the VAM.

5. **The one-size-fits-all approach of VAMs may limit innovation.** VAMs work best when as much as possible is held constant across classrooms—student composition, curriculum, testing conditions, etc. At the same time, there is a push for schools to experiment with more innovative modes of instruction—providing more differentiation, greater use of technology, team teaching methods, nontraditional classroom structures, and so on. The assumptions behind VAMs are even less likely to hold in these settings. To the extent VAM becomes the predominant model for evaluating teacher effectiveness, it may discourage such innovation.
6. **Face validity.** Even if VAM-based performance evaluations are carefully and conservatively designed, they may still suffer from concerns of face validity. Economists may be comfortable with year-to-year correlations of 0.4–0.5, but teachers who observe large fluctuations in their value-added score (with no apparent connection to changes in their own practice) may be less so. By the same token, while it is correct for researchers to say that VAMs stabilize and become more precise after several years of classroom teaching, this is of less comfort to a principal or teacher who needs feedback in real time. Indeed, in some settings, many teachers leave their school or district before 2-3 years of results accumulate. (A study in New York City, for example, found that 57 percent of middle school teachers left their school within 3 years, with more than half of these exiting the district [Marinell, 2011]). Finally, perhaps the most important indicator of face validity in this context is the extent to which VAMs adequately capture the full range of teachers' job expectations. If the standardized tests on which VAMs are based represent only a fraction of the skills teachers are expected to cultivate in their students, but are given disproportionate weight in their evaluation, there will be a perceived misalignment between expectations and assessment (being rewarded for doing A, while being asked to do A, B, and C).

References

- Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement. *Education Finance and Policy*, 1(2), 176–216. Retrieved from <http://www.mitpressjournals.org/toc/edfp/1/2>
- Chetty, R., Friedman, J.N., and Rockoff, J.E. (2011). The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. National Bureau of Economic Research Working Paper #17699. <http://www.nber.org/papers/w17699.pdf>
- Corcoran, S., Jennings, J., and Beveridge, A. (2011). Teacher Effectiveness on High- and Low-Stakes Tests. Working Paper, New York University. https://files.nyu.edu/sc129/public/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf
- Goldhaber, D., Gross, B., & Player, D. (2011). Teacher career paths, teacher quality, and persistence in the classroom: Are public schools keeping their best? *Journal of Policy Analysis and Management*, 30(1), 57–87. Retrieved from <http://dx.doi.org/10.1002/pam.20549>
- Gordon, R., Kane, T.J., & Staiger, D.O. (2006). Identifying Effective Teachers Using Performance on the Job. Washington, DC: Brookings Institution. Retrieved from http://www3.brookings.edu/views/papers/200604hamilton_1.pdf
- Hanushek, E.A., Kain, J.F., & Rivkin, S.G. (2004). Why Public Schools Lose Teachers. *Journal of Human Resources*, 39(2), 326–354. Retrieved from <http://jhr.uwpress.org/content/XXXIX/2/326.short>
- Harris, D.N., & Sass, T.R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7-8), 798-812.

- 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.O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Cambridge, MA, NBER.
- Loeb, S., Kalogrides, D., & Béteille, T. (2011). Effective Schools: Teacher Hiring, Assignment, Development, and Retention. National Bureau of Economic Research Working Paper #17177. Retrieved from <http://www.nber.org/papers/w17177>
- Marinell, W. (2011). The Middle School Teacher Turnover Project: A Descriptive Analysis of Teacher Turnover in New York City's Middle Schools. New York, NY: Research Alliance for New York City Schools. Retrieved from: http://steinhardt.nyu.edu/scmsAdmin/media/users/jnw216/RANYCS/WebDocs/TTP_EXECUTIVE-SUMMARY-FINAL.pdf
- Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*, 125(1), 175–214. doi:10.1162/qjec.2010.125.1.175

Linda Darling-Hammond, Stanford University

Recent developments in the value-added literature raise a growing number of questions about the use of VAM for high-stakes purposes in teacher evaluation systems. There are now many studies that establish the instability of teacher effectiveness ratings based on VAM methods. This instability exists on many dimensions.

- Teachers' effectiveness ratings differ substantially from class to class, from year to year, and from one statistical model to the next.
- In some systems currently in use for teacher evaluation, the correlation in value-added ratings from one year to the next is near zero. Correlations for individual teachers' ratings from year to year in the published literature range from about 0.20 to 0.50, with a teacher's ratings in one year accounting for 4-25 percent of the variance in ratings in the second year, leaving most of the variance related to other factors.
- Instability appears to be worse at the tail end of distribution, where policymakers would most like to use VAM ratings to reward or dismiss teachers. Braun pointed this out in a review of research some years ago. Looking across five large urban districts, Sass found that, of teachers in the bottom quintile in one year, only about 20-30 percent would remain equally low scoring in the next year, and more would move to the top half of the distribution. The same trends were true for teachers in the top quintile in a given year with respect to movement to the bottom half.
- The difference in a teacher's VAM ratings across classes and years is significantly associated with classroom composition, even if prior test scores and student demographics such as poverty, parent education, and English learner status have been previously controlled. Adding classroom composition as an additional independent variable can reduce this effect somewhat, though not eradicate it, but current district models do not begin to reach this level of sophistication.
- Teachers teaching different classes typically receive very different value-added ratings and, when tested, the class turns out to be a stronger predictor of the rating than the teacher.
- Even though reliability can be ostensibly increased by averaging across years, this technique does not solve the reasons for the instability; it merely masks it. Rolling averages can disadvantage teachers as much as annual ratings due to the effects of one low VAM year (see, for example, cases in Houston when teachers have been dismissed after recent strong ratings due to a single low rating 2 or 3 years earlier when they had been assigned a class of newly mainstreamed English learners).

There is evidence of systematic bias in VAM ratings, as a function of both characteristics of students and characteristics of tests:

- Several studies have found that teachers are disadvantaged by having large numbers of new English learners and special education students in their classes, receiving lower VAM ratings than they do in other classes with different student populations. This may be in part because these students are often not validly assessed by traditional standardized tests and in part because of the effects of concentrations of such students on the functioning of the classroom.
- Several studies have also found that teachers are disadvantaged by teaching advanced or gifted and talented students, because it is difficult to show gains at the top of the distribution, especially on tests that have a low ceiling. Problems at the edges of the distribution are likely exacerbated by the No Child Left Behind (NCLB) requirement that current state tests measure grade-level standards only. The new assessment consortia have been instructed that they can create tests that measure more than a single grade level, but uncertainty about

NCLB and practical concerns about test costs and time are likely to limit advances in this arena.

Individual teachers' VAM ratings do not strongly correlate with other indicators of teacher effectiveness, even other test-based measures.

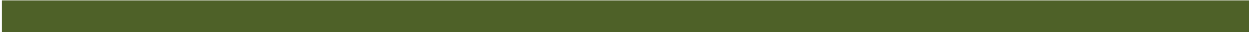
- VAM scores are correlated, though weakly, with principals' evaluations and with a variety of structured classroom observation tools that are reasonably related for research purposes but not strong enough to give confidence that the unstable VAM rating is a better indicator of effectiveness than the more stable ratings of practice.
- Teachers' VAM scores are noticeably different on different tests, including differences on tests measuring basic and higher order skills or performance skills. Correlations are typically in the range of 0.3 to 0.5—again, reasonably related for research purposes but far too different at the individual teacher level to justify the use of a single metric as the only measure of learning.

The use of a single VAM rating on the high-stakes state test is particularly problematic in the context of current policies that give this measure primacy in the ultimate judgment of teacher effectiveness. Recent analyses of data in New York City (where teachers must be rated ineffective overall and put on a path toward dismissal if their VAM rating is in the range pre-labeled as ineffective) have surfaced many examples of teachers with low VAM ratings who are highly rated by their principals or whose students score well on tests other than the state tests. For example, the “worst” teacher in the city by recent VAM ratings is a well-respected teacher of brand-new immigrant students who is highly rated by her principal and whose students show gains on local assessments of English proficiency progress.

The “worst” 8th grade math teacher in New York City illustrates another problem illuminated in an analysis by Columbia University professor Aaron Pallas: Carolyn Abbott, a teacher of a combined 7th/8th grade GATE class, taught in a school for the gifted. After a year in her classroom, her 7th grade students scored at the 98th percentile of New York City students on the 2009 state test. As 8th graders, they were predicted to score at the 97th percentile on the 2010 state test. However, their actual performance was at the 89th percentile of students across the city on material they had studied 2-3 years earlier. That shortfall placed Abbott at the bottom of 8th grade mathematics teachers in New York City. Meanwhile, Abbott was teaching her students the more advanced Regents high school curriculum and 100 percent of her honors section 8th graders passed the Integrated Algebra test (normally taken by 10th graders) in January—one-third of them with a perfect score. Despite this success, her principal, who rated Abbott very highly, could not guarantee that Abbott would receive tenure, which is based, first and foremost, on value-added measures in New York. Abbott—arguably one of the most successful math teachers in New York City—left teaching at the end of this year.

In Houston, where VAM ratings are unstable and have also been found to be related to proportions of new English learners on one hand and gifted students on the other, principals have been required to make their ratings of teachers conform to the VAM ratings. Several hundred teachers have been dismissed in Houston based on these ratings. The same policy (requiring principals to align their ratings to VAM scores) is currently being considered in Tennessee.

There are few studies that have looked at the accuracy of decisions about teacher tenure and continuation based on how student-learning evidence (including but not limited to VAM measures), observation evidence, and other measures (student surveys, peer reviews, etc.) are assembled, combined, and evaluated in relation to each other and the students being served.



This kind of research could be critically important as policy uses of VAMs are about to expand greatly in the coming months and years.

John Friedman, Harvard University

“Value-Added and Long-Term Outcomes”

The Long-Term Impact of Teachers: Teacher Value-Added and Student Outcomes in Adulthood by Chetty, Friedman, and Rockoff (2012) is one of the key papers on value-added and teacher effectiveness in the past few years. This memo summarizes the findings in that paper and the implications of those findings for policy.

The debate on value-added (VA) and teacher effectiveness stems from two fundamental questions among others. First, does VA accurately measure teachers' impacts on scores or does it unfairly penalize teachers who may systematically be assigned lower achieving students? Second, do high VA teachers improve their students' long-term outcomes or are they simply better at teaching to the test? Researchers have not reached a consensus about the accuracy and long-term impacts of VA because of data and methodological limitations.

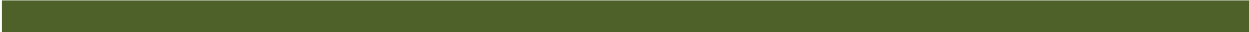
Our study addresses these questions by tracking one million children from a large urban school district from 4th grade to adulthood. We evaluate the accuracy of standard VA measures using several methods, including natural experiments that arise from changes in teaching staff. We find that when a high VA teacher joins a school, test scores rise immediately in the grade taught by that teacher; when a high VA teacher leaves, test scores fall. Test scores change only in the subject taught by that teacher, and the size of the change in scores matches what we predict based on the teacher's VA. These results establish that VA accurately capture teachers' impacts on students' academic achievement. Moreover, our methods provide a simple yet powerful technique to estimate the bias of VA models in any district.

In the second part of our study, we analyze whether high VA teachers also improve students' long-term outcomes. We find that students assigned to higher VA teachers are more successful in many dimensions. They are more likely to attend college, earn higher salaries, live in better neighborhoods, and save more for retirement. They are also less likely to have children as teenagers.

Teachers' impacts on students are substantial. Replacing a teacher whose true VA is in the bottom 5 percent with one of average quality would generate a cumulative earnings gain of \$52,000 per student or more than \$1.4 million for the average classroom; discounting at a 5 percent interest rate to age 12 yields a present value gain of more than \$250,000 per classroom. VA estimates are less reliable when they are based on data from fewer classes. However, even after observing teachers' impacts on test scores for 1 year, estimates of VA are reliable enough that such personnel changes would yield large gains on average.

Teachers have large impacts in all the grades we analyzed (4 to 8), suggesting that improving teacher quality is valuable throughout elementary school. Teachers' impacts on earnings are similar in percentage terms for students from low- and high-income families. This result shows that better teaching can be valuable even in environments where students may not have the best resources outside school.

In addition to aiding in the evaluation of individual teachers, VA also has great promise as a measure of teacher quality more generally. For instance, districts could use VA to measure the effectiveness of teacher development programs. Districts could also use VA to find observable characteristics of applicants that better predict future effectiveness in the classroom. Finally, districts could use VA to measure the distribution of teacher equality across grades or schools. In each of these applications, schools would be focusing on the average VA within a group of



teachers, which sharply reduces both the incentive for teaching to the test and the uncertainty from classroom-level test score fluctuations that are not driven by teachers.

More Research is Needed

More research is still needed to understand the implications of using VA for teacher evaluations. A key concern is that using VA in teacher evaluations could potentially induce counterproductive responses that make VA a poorer measure of teacher quality, such as teaching to the test or cheating. We can learn about this issue only by studying school districts that start using VA to evaluate teachers. The best policy will likely put some weight on VA measures and some weight on subjective evaluations (e.g., classroom observations by principals). The best weight is an important question for future research.

Federal Policy Implications

There are a number of federal policies that would encourage the take-up of VA and increase our understanding of VA as a policy tool. For instance, the federal government could:

1. Encourage districts to calculate VA and analyze the accuracy of the measure in their particular institution setting.
2. Incentivize the development and validation of performance measures in subjects and grades not currently tested.
3. Collect data on test scores, and principal and peer observation ratings in a standardized format.

The federal government could also take steps to improve teacher quality more broadly:

1. Help make teaching an elite, high-status profession through salary bonuses and public recognition for high VA teachers or schools. Promise such bonuses and recognition to attract top talent.
2. Give states incentives to reform tenure provisions and last-in-first-out dismissal rules. For instance, give grants to districts that adopt more rigorous tenure evaluations.

Dan Goldhaber, University of Washington

“The Ability to Act on Differences Between Teachers: Empirical Work Fueling the Debate Over the Use of Value-Added”

Five central empirical findings have fueled recent academic and policy debates about the high-stakes use of student growth measures as an input in teacher evaluation.¹ First, teacher effectiveness varies widely and the variation has educationally meaningful consequences for students. Second, a spate of new research affirmed earlier findings, from the last decade and before, that showed teacher effectiveness is not strongly related to the credentials typically used to determine employment eligibility and compensation. Third, we now know that there is little variation in in-service teacher performance evaluation ratings. Fourth, value-added measures have been found to be moderately reliable. And fifth, research suggests that value-added measures may be biased. In describing a few of the studies that have influenced this debate, I concentrate most of the discussion on the last two findings—that value-added measures *may* be biased and/or unreliable—believing that the first points are now largely uncontroversial.

The teacher-effects literature has grown concurrent with the availability of administrative data that links teachers and students. The literature typically finds teacher effect size estimates in the neighborhood of 0.10 and 0.25 standard deviations.² For some perspective on what this means in more concrete terms, the magnitude of these effect sizes suggest that having a highly effective teacher (84th percentile) rather than an average teacher (50th percentile) is estimated to make a difference of roughly 10-25 percent of a grade-level’s worth of achievement in elementary school, and/or could cut achievement gaps between black and white or economically advantaged and disadvantaged students down 10-35 percent.³ There may be some debate about the magnitude of estimated teacher effects, but few would debate the differences between individual teachers are meaningful. What is debatable is how or whether to use this information.

Knowing that teachers are important and acting on that knowledge are two different things.⁴ The effect size estimates presented above are, by definition, based on historical information, but acting on estimated differences between teachers requires a prediction based on past information. Several recent studies have shown teacher value-added effect estimates to be only moderately reliable/stable from year to year (e.g., Goldhaber and Hansen, forthcoming; McCaffrey et al., 2009), with adjacent year correlations in the neighborhood of 0.3 to 0.5.⁵

¹ I use the terms “student growth measures,” “value-added,” “teacher effectiveness,” and “teacher performance” interchangeably. For a good example of how this academic debate plays out amongst academics in outlets designed to be accessible to policymakers and practitioners, see Darling-Hammond et al. (2012) and Glazer et al. (2010).

² The estimates are typically in the neighborhood of 0.10 to 0.15 for within-school estimates and are 0.15 to 0.25 for estimates that include between-school differences in teacher effectiveness. See, for instance, Goldhaber and Hansen (forthcoming) and Hanushek and Rivkin (2010) for a more thorough discussion of the teacher effect size literature.

³ This calculation is based on the finding that as students move from one grade to the next, they typically gain about one standard deviation in math and reading achievement in the lower elementary grades (Bloom et al., 2008), and that the average gap between black and white students, or economically advantaged and disadvantaged students, is on the order of magnitude of 0.7 to 1.0 standard deviations (Hanushek and Rivkin, 2010).

⁴ I will not delve into it in any detail here, but there are a number of logistical issues (e.g., test timing, student teacher attribution, etc.) that have received little empirical attention, but have potential import for both the accuracy and use of value-added.

⁵ Research generally finds value-added estimates are not terribly sensitive to model specification, but does find evidence of sensitivity to the test employed—i.e., different tests administered to the same set of students result in somewhat different value-added estimates for their teachers (e.g., Lockwood et al., 2007; Papay, 2011).

Given reliability estimates of this magnitude, teacher performance estimates that relied on value-added *alone* (a bit of a red herring as I do not know of policymakers or researchers advocating this) would result in a non-trivial number of misclassifications of teachers (Schochet and Chiang, 2010).⁶ But, as Glazerman et al. (2010) point out, the magnitude of these stability estimates are not very different from what is observed in other occupations that use them for high-stakes purposes.

Research has also raised questions about the validity of value-added estimates. For instance, in an influential paper, Rothstein (2010) shows that in standard value-added models, teachers assigned to students in the *future* have statistically significant predictive power in predicting *past* student achievement. This finding clearly cannot be causal, rather it suggests that typical models do not adequately account for the selection process leading to the matching of students and teachers in classrooms; consequently, value-added estimates are likely to be biased. Other research, however, has come to a different conclusion about the likelihood that value-added models are biased, and/or the potential magnitude of bias, which clearly matters when we are thinking about using value-added.⁷ More importantly, I would venture that few would dispute whether there is useful information in value-added teacher performance estimates; disputes arise when it comes to discussions of how the information ought to be used.⁸

Perhaps a deeper issue than the potential that value-added is unreliable or biased is whether student test scores really ought to be the centerpiece of teacher accountability policies. If value-added is more a reflection of teachers focusing narrowly on test-taking skills to elicit short-term test gains at the expense of genuine learning, then teacher effects on students are likely to “fade out” over time, as has been found empirically (Jacob et al., 2010; Kane and Staiger, 2008). Thus, using value-added for high-stakes purposes would likely have unintended negative consequences. But recent research by Chetty et al. (2011) provides a measure of external validity of value-added estimates as it shows the estimates are statistically significant predictors of later life outcomes, such as college-going behavior and earnings. In other words, the Chetty study strongly suggests that value-added is, in fact, an important measure of true learning gains by students.

It is of course important to consider the counterfactual when assessing the risks of using value-added. As I stated at the beginning, it is relatively uncontroversial today to assert that, outside of early career teaching experience, the credentials (licensure, degree, and experience level) used for employment eligibility and compensation decisions in most school districts are, at best, only weakly related to teacher effectiveness.⁹ And, as was documented in *The Widget Report* (Weisburg et al., 2009), most teachers are at the top of whatever rating system school districts use, implying that in-service evaluations are not useful for informing personnel decisions. This

⁶ Schochet and Chiang, for instance, used simulated data that relies on plausible estimates of the signal to noise ratio in teacher effect estimates and conclude that, if three years of data are used for estimating teacher effectiveness, the probability of identifying an average teacher as being “exceptional” (Type I error)—defined by them as teachers who are roughly one standard deviation of teacher performance above (or below) the mean—is about 25 percent. Conversely, the probability that a truly exceptional teacher is not identified for special treatment (Type II error) is also roughly 25 percent.

⁷ See, for instance, discussion of other specification tests in Chetty et al. (2011) and Kane and Staiger (2008), and reasons why the Rothstein test may not work as intended in Goldhaber and Chaplin (2012) and Kinsler (2011). See Goldhaber and Chaplin (2012) and Rothstein (2009) for a discussion of the potential magnitude of bias.

⁸ Simulations (e.g., Boyd et al., 2010; Goldhaber and Hansen, 2010; Goldhaber and Theobald, 2011; Hanushek, 2009; Staiger and Rockoff, 2010) suggest that using value-added for consequential teacher workforce selection policies (tenure, layoffs, etc.) could have significant effects on the quality of the teacher workforce. These simulations, however, ignore the potential behavioral responses of teachers to the use of value-added (e.g., the potential that it could change the propensity to pursue a career in teaching).

⁹ Value-added effect estimates have high-predictive power for out-of-sample student achievement (Jacob and Lefgren, 2008; Jackson and Breugmann, 2009; Kane and Staiger, 2008) and have been shown to be a better predictor than licensure status, degree, and experience levels (Goldhaber and Hansen, 2010).

means that the existing counterfactual in most places is the use of a set of teacher credentials that we know are not associated with teacher effectiveness.

Ultimately, one's opinion about the risk inherent in using value-added is likely to be shaped by an assessment of either sticking with today's counterfactual strategy that ties teacher personnel policies to teacher credentials, or some other means of assessing teachers. There are certainly a variety of methods other than value-added that can be used to assess teachers (and I would personally advocate using multiple indicators to assess teacher performance), but it is important to recognize that some of the other methods that can be used to derive teacher performance estimates—such as classroom observations, student perceptions surveys, student learning objectives, etc.—are also relatively untested and likely suffer from some of the same shortcomings as value-added.¹⁰ If I am right that the central aspect of the debate over using value-added is not primarily centered on specific empirical findings, but rather the assessment of risks given what is known about teacher quality, then it is unlikely that evaluation of the existing evidence will do much to settle the question of whether value-added ought to be used.

References

- Bloom, H., Hill, C., Black, A., & Lipsey, M. (2008). *Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions*. New York: MDRC.
- Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2010). *Teacher Layoffs: An Empirical Illustration of Seniority v. Measures of Effectiveness*. CALDER working paper, July 20, 2010.
- Chetty, R., Friedman, J.N., & Rockoff, J. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. NBER.
- Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). *Evaluating teacher evaluation*. *Phi Delta Kappan*, March, 2012.
- Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Brookings Institution.
- Goldhaber, D., & Chaplin, D. (2012). *Assessing the "Rothstein Falsification Test." Does it really show teacher value-added models are biased?* CEDR Working Paper 2012-1.3. University of Washington, Seattle, WA.
- Goldhaber, D., & Hansen, M. (2010). *Using performance on the job to inform teacher tenure decisions*. *American Economic Review*, 100(2), 250-255.
- Goldhaber, D., & Hansen, M. (forthcoming). *Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance*. *Economica*.
- Goldhaber, D., & Theobald, R. (2011). *Managing the Teacher Workforce in Austere Times: The Implications of Teacher Layoffs*. CEDR Working Paper 2011-1.2. University of Washington, Seattle, WA.
- Hanushek, E. (2009). *Teacher Deselection*. In D. Goldhaber & J. Hannaway (Eds.), *Creating a New Teaching Profession*. Washington, DC: Urban Institute Press.
- Hanushek, E., & Rivkin, S. (2010). *Generalizations about using value-added measures of teacher quality*. *American Economic Review*, 100(2), 267-271.
- Jackson, C., and 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.

¹⁰ Untested in the sense that we do not know the degree to which they predict true teacher effectiveness/student learning.

- Jacob, B.A., Lefgren, L., & Sims, D.P. (2010). The Persistence of Teacher-Induced Learning. *Journal of Human Resources*, 45(4), 915-943.
- Kane, T., & Staiger, D.O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Cambridge, MA, NBER.
- Kinsler, J. (February, 2011). Assessing Rothstein's critique of teacher value-added models. Working paper.
- Lockwood, J.R., McCaffrey, D.F., Hamilton, L.S., Stecher, B.M., Le, V., & Martinez, F. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Educational Measurement*, 44(1), 47-67.
- McCaffery, 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.
- Papay, J.P. (2011). Different Tests, Different Answers. *American Educational Research Journal*, 48(1), 163-193.
- 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.
- Schochet, P., & Chiang, H. (2010). Error Rates In Measuring Teacher and School Performance Based on Student Test Score Gains. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Staiger, D.O., & Rockoff, J.E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, 24(3), 97–118.
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York, NY: The New Teacher Project.

Andrew Ho, Harvard Graduate School of Education

In the first half of this brief, I review three areas where I believe there have been significant advances since 2008: Sensitivity Studies, Normative Growth Models, and the Measures of Effective Teaching (MET) Project. In the second half, I review four areas that I believe hold significant promise for the future: *Shifting from “Symptoms” to “Treatment,” Properties of Accountability Indices, Auditing for Unintended Consequences, and Test Scaling.*

1) Sensitivity Studies: Over Time, Across Models, Across Tests

An explosion of recent research addresses the general question, “Would value-added teacher rankings be different had we used different X,” where X is time points, models, or tests. The 2010 EPI Briefing¹ reviewed much of this literature, including an oft-cited piece by McCaffrey, Sass, Lockwood, and Mihaly² on “intertemporal variability” (see also Goldhaber and Hansen³). In a 2012 *Science* article, Henry, Fortner, and Bastian⁴ attempt to explain some of this intertemporal variability as average growth for novice teachers over time.

Cross-model comparisons are more dated, including those by McCaffrey et al., and Tewke et al., in the 2004 JEBS special issue.⁵ It is worth noting that these have made it into reporting practice, for example, in the LA Times value-added reporting tool, where estimates from multiple models are readily available for each teacher.⁶ Recent developments have not made as substantial advances, as the reasons for cross-model discrepancies are generally well explained by the different questions each model is implicitly asking. Cross-test comparisons include those by Lockwood et al. in 2007⁷ and Papay in 2010.⁸ The best of these articles explain cross-X discrepancies with thoughtful hypotheses rather than treating all differences as “noise.”

2) Normative Growth Models and Usability

The rise of “normative growth models,” particularly in the form of Betebenner’s Student Growth Percentiles (SGPs),⁹ has had an immense impact on recent practice. The approach willingly sacrifices statistical rigor for an interpretable scale (percentile ranks) and transparent aggregation (with median SGPs, whereas rival VAMs have no straightforward individual-level “growth” statistics to aggregate). Although a recent JEBS article by Castellano and Ho notes strong similarities between SGPs with traditional regression alternatives,¹⁰ I believe that the advance that SGPs represent is less one of statistics than one of reporting, and it is no less an important advance for this effort.

¹ <http://www.epi.org/publication/bp278/>

² <http://www.mitpressjournals.org/doi/pdf/10.1162/edfp.2009.4.4.572>

³ [http://cedr.us/papers/working/CEDR%20WP%202010-3_Bad%20Class%20Stability%20\(8-23-10\).pdf](http://cedr.us/papers/working/CEDR%20WP%202010-3_Bad%20Class%20Stability%20(8-23-10).pdf)

⁴ <http://www.sciencemag.org/content/335/6072/1118>

⁵ <http://jeb.sagepub.com/content/29/1.toc>

⁶ <http://documents.latimes.com/buddin-white-paper-20110507/>

⁷ <http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2007.00026.x/full>

⁸ <http://aer.sagepub.com/content/48/1/163.full.pdf+html>

⁹ <http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.2009.00161.x/abstract>

¹⁰ <http://jeb.sagepub.com/content/early/2012/05/03/1076998611435413.full.pdf+html>

3) Measures of Effective Teaching

Given Tom Kane's attendance, I'll only say that these reports are a much needed step toward the multiple measures that a formative evaluation system requires. I expand on this in the next point. The remainder of this brief is dedicated to areas that I think hold particular promise for the future.

4) Advancing from Symptoms to Diagnosis and Treatment

The VAM field has been disproportionately focused on modeling and the reliable ranking of teacher effects. To use a medical analogy, this is akin to focusing on the "How much pain do you feel?" questionnaires, where doctors get an initial read on whether or not you are sick. But medicine (and education) is not only about symptoms (and even less so about unidimensional rankings of symptoms) but, far more critically, diagnosis and, ultimately, treatment. How can we use VAM results to improve teaching and/or the teacher corps? Current IES (Institute of Education Sciences) ventures into researcher-practitioner partnerships could help here, but so could an increased focus on score reporting and usability.

5) Differences that Matter: The Statistical Properties of Accountability Indices

VAM metrics rarely support decisions unilaterally. Instead, they are incorporated into increasingly complex indices whose calculations are driven by simplicity and political rhetoric. There are devils in these details. A current line of my work quantifies the variability that arises from differing interpretations of ambiguous statements like "50% of the teacher evaluation metric should be supported by VAM" and describing variability that arises from the popular approach of creating arbitrary categories, like "below average," "average," and "above average," at various stages of index construction. These decisions can contribute to far more variance than, for example, the difference between SGPs and mixed-effects models. We need to distinguish between "accountability metrics" and "accountability indices" and explicate properties of indices as often as we do for metrics.

6) Auditing for Unintended Consequences

In this day and age, unintended consequences are rarely unanticipated, and yet a science of auditing has not advanced far beyond NAEP (National Assessment of Educational Progress) comparisons. My colleague, Dan Koretz, is doing some work on developing "internal audits" for high-stakes tests. He should not be alone.

7) Scaling

Finally, as a psychometrician, I must mention scaling as an elephant in the room. My work with Sean Reardon and Ed Haertel develops nonparametric, "ordinal" measures for trends and gaps that are invariant to some scaling decisions. But we still lack thoughtful frameworks, somewhere between the "ordinal" and the "interval/cardinal," to describe the dependence of VAMs on scaling decisions.

Tom Kane, Harvard University

“Recent Advances in Understanding of Teacher Value-Added”

School districts and states are in the midst of reinventing the way they evaluate and provide feedback for teachers. As we have seen in Chicago, the changes are controversial. Although parents may find it bewildering and teachers wonder if they are being blamed for student failure, there are three findings which underscore the importance of that work.

First, it is very difficult to know who the effective teachers are going to be based on their preparation and characteristics before they enter the classroom. There is a longstanding literature which finds small or statistically insignificant difference in student learning outcomes between certified and uncertified teachers, those with advanced degrees and those without. However, the most striking evidence comes from the Teach for America (TFA) program, which provides an upper bound on how selective school districts could ever hope to be. Yet, even in the random assignment study of TFA posted on the website, the gains are 2 percentile points.

Second, once they enter the classroom, large differences between teachers become apparent. Some teachers are much more successful than others in promoting student achievement gains. I don't need to rehash that voluminous literature.

Third, most of the evidence suggests that teachers largely plateau in their effectiveness after a few years on the job.

Given these facts, we need a teacher evaluation system which can do two things well: (i) help identify the teachers who will be successful with future groups of students and (ii) provide feedback on specific instructional practices which will allow teachers to improve their practice.

In the first of these two goals, student achievement gains are fundamental. Despite volatility, a teacher's track record of value-added is the single best predictor of their future value-added. However, for the second goal—providing teachers with the feedback they need to continue to improve their practice—value-added measures are not very useful at all. Therefore, although value-added measures have a role to play in teacher evaluation systems, they should not be the sole measure used.

The literature on teacher effectiveness and value-added has been evolving rapidly in recent years. I summarize the highlights of recent advances, organized by topic, below.

1. **Causality:** Rothstein (2010) correctly pointed out that selection on unobserved student characteristics could lead to bias in “value-added” estimates as measures of causal effects of teachers on student achievement. However, the paper presents no evidence that there is sorting on the basis of unobservables which is generating bias. He simply did not have the data or the analytic design to test that proposition. So far, we have two papers which present direct tests of bias due to sorting on unobserved student characteristics: Kane and Staiger (2008) randomly assigned students (at the roster level, not the student level) within 78 pairs of teachers in Los Angeles. They could not reject the hypothesis of no bias in value-added estimates in predicting student achievement within a school and grade. The study design did not allow one to test for bias outside of a given school. Chetty, Friedman, and Rockoff (2011) studied changes in school-level average scores in various grades, following the movement of high and low value-added teachers across schools and grades. They could not reject the hypothesis of no bias.

Two upcoming reports will shed additional light. The MET project randomly assigned classroom rosters among clusters of teachers within school, grade and subject. Those results should be available by January 2013. Steve Glazerman is working on a study in which high value-added teachers were provided incentives to switch schools. While the former study will provide additional data on the bias in value-added measures in predicting differences within school, the latter study will focus on the bias between schools.

2. **Volatility:** Prior research on volatility has focused on year-to-year volatility. However, for most purposes, it is the correlation between a single year performance and the long-term or career measure that we should be most concerned with. For instance, in a tenure decision, it's a teacher's likely impact on students over the course of their later careers which should be the criterion. If a teacher's effectiveness is stable over long periods, then the year-to-career correlation is equal to the square root of the year-to-year correlation. So, if the year-to-year correlation in value-added is .49 (commonly found with math), then the correlation between a single year and career-value added would be much higher—.7.

Moreover, we have also learned that volatility on other measures of teaching effectiveness, such as classroom observations, is also a challenge. In the MET report, we averaged four observation scores, each by a different observer, to attain reliabilities of 0.6 or above. If a principal is doing the observation, the reliability challenge may not be visible from year to year, but there will be large shifts in ratings whenever a principal turns over.

3. **Long-term effects on students:** Chetty, Friedman, and Rockoff (2011) reported the relationship between a teacher's value-added and his or her long-term impact on earnings. Their research is a breakthrough in the debate over whether value-added on state tests is measuring anything we care about. It also suggests that the fade-out in teacher effects—which many have also noted recently—does not continue indefinitely.
4. **The nature of the test matters, especially in literacy:** Researchers have commonly found that a teacher's value-added varies by test and subject. For instance, among elementary teachers, gains in math and ELA tend to have a correlation of .60. In the second MET report, we found that a teacher's gains on state ELA tests are correlated roughly .4 with a teacher's gains on the Stanford 9 open-ended reading test. However, we also found that the gains as measured by the Stanford 9 were more correlated with other measures of teaching practice, such as the PLATO instrument, than the value-added on the state tests were. The lower external validity of the state tests in ELA raises serious concerns about the quality of the information yielded by the current ELA tests. One explanation is that they rely heavily on multiple choice questions of reading comprehension and often have very few constructed response items, even though ELA instruction after the early grades focuses elsewhere.
5. **External validity:** Jacob and Lefgren and Rockoff, Kane, and Staiger reported that principal perceptions are related to value-added. Kane, Taylor, and Tyler reported that value-added is related to formal classroom observations. Through the MET project, we have learned that value-added is related to formal observations, student surveys and tests of pedagogical content knowledge. In other words, value-added measures are not purely a statistical construct. There is an underlying construct of excellent instruction to which they are all related.
6. **Model specification:** Given the wide range of empirical methods used to estimate teacher effects, we have much to learn about the critical assumptions required for validity. Still, what we have managed to learn narrows down the range of options slightly.

- a. Student fixed effects lead to bias: For the most part, researchers have abandoned the inclusion of student fixed effects in estimating teacher effects. Rothstein (2010) pointed out that such models require assuming that students are sorted to teachers based on fixed characteristics, not variations in actual performance on state tests. If students are assigned to teachers partially on the basis of how they performed on recent state tests, the student fixed effect models will generate biased estimates of teacher effects. Kane and Staiger (2008) showed that teacher effects estimated with student fixed effects generated biased predictions of student achievement following random assignment.
- b. A single year of prior achievement yields estimates which are highly correlated with those which include more than 1 year of lagged achievement: Although Rothstein (2010) could reject the hypothesis that the coefficient on the 2-year lag was zero, the teacher effects estimated with and without more than 1 year of prior achievement were correlated .98. As a practical matter, it makes little difference whether one controls for more than 1 prior year of achievement.
- c. As long as baseline achievement is included, the inclusion of student demographics and free lunch status often makes little difference. For several years, policymakers have argued over whether or not to include controls for individual students' demographics—such as race or socioeconomic status. There is no easy conceptual resolution to the debate. On one hand, race and family income may serve as proxy measures of environmental influences outside a teacher's control and, as such, one could argue for their inclusion. On the other hand, including such measures may enshrine differing teacher expectations or average effectiveness of teachers assigned to students of different race or socioeconomic status. Empirically, however, the estimates with and without such controls tend to be highly correlated (>.95), making the debate over the right approach conceptually moot in most districts. (The nature of sorting could differ within each new district and so the above analysis should be replicated before setting the issue aside.)
- d. There is a sizeable error at the classroom-by-year level, which must be accounted for, above and beyond sampling variation. The variance in the classroom-by-year error—as would be caused, for instance, by a “dog barking in the parking lot on the day of the test”—is often nearly as large as the teacher-level signal variance. This has large implications for “shrinkage” or empirical Bayes predictions of future student achievement. The tests for bias—mentioned above—incorporate such estimates in the predictions and shrink the estimates accordingly. Failing to account for such error would lead to biased predictions.

In practice, one specification choice which “matters” (in the sense that it leads to different estimates at the teacher level) is the inclusion of mean peer characteristics of other students in the class. Some researchers include peer characteristics in their specifications, but many do not. Moreover, most states and districts are not including peer characteristics in their own empirical models. We need to learn whether or not such covariates are necessary to avoid bias.

References

- Chetty, R., Friedman, J.N., & Rockoff, J.E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. National Bureau of Economic Research Working Paper #17699. Retrieved from <http://www.nber.org/papers/w17699.pdf>
- Kane, T., & Staiger, D.O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Cambridge, MA, NBER.
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1), 175–214.
doi:10.1162/qjec.2010.125.1.175

Helen Ladd, Duke University

Based on my review of the literature on value-added models of teacher effects (Ladd, 2008) and various other research projects, I am skeptical of the usefulness and fairness of such estimates for high-stakes decisions about individual teachers and I am concerned about the harm that might result from their use. The skepticism comes from the sensitivity of the estimates to model specification, the small samples of test scores available for many teachers, the difficulties of separating classroom effects from teacher effects, the instability of such estimates from one year to the next, and the fact that they can be estimated for only a small portion of all teachers. My concern about harmful effects reflects the potential for any quantitative measure to corrupt the social processes it is intended to monitor (Campbell, 1979). Such corruption might come from narrow teaching to the test and reducing the appeal of teaching to current or future teachers who believe the purpose of teaching is broader than raising student test scores.

I continue to support the views expressed in the Economic Policy Institute policy brief of which I was a co-author (Baker et al, 2010). I will leave to others the task of summarizing the most recent technical research and instead will raise two specific policy-relevant concerns about their use in practice.

Value-added models are not well suited to comparing the effectiveness of teachers across schools.

A major challenge for any estimate of a teacher's value-added is that students are not randomly assigned to teachers either across schools or across classrooms within schools. Unless the analyst takes account of that basic fact, the resulting value-added measures could well be seriously biased because part of what is attributed to the teacher may in fact reflect the abilities and motivations of the students she teaches and the context in which she is teaching. The inclusion of student fixed effects and/or lagged student achievement addresses a large part of the problem but does not address the bias that arises because of differing contexts across schools or across classrooms within schools. That conclusion holds not only for straightforward value-added models but also for the data intensive mixed method models such as the TVAAS model (Lockwood and McCaffrey, 2007; Ballou, Sanders and Wright, 2004).

To address the problem of non-random sorting of teachers across schools that are often stratified by disadvantage, many value-added models control statistically for school characteristics, either fully with the use of school fixed effects or partially with the inclusion of measurable school characteristics. Either procedure means that the value-added models cannot be used to compare the effectiveness of teachers in one school relative to another. A model that includes fixed effects for individual schools, for example, removes all school-specific variation (both measurable and unmeasurable) from the analysis and generates teacher effects that should be interpreted as the effect of one teacher relative to another within each school. Even in models that eschew school fixed effects in favor of measurable characteristics of the schools, the problem remains. For example, if one controls statistically for school poverty rates or correlated measures, one cannot then make any statements about the effectiveness of teachers in high poverty schools relative to those in low poverty schools. There is no technically acceptable way to get around this problem.

Although I acknowledge that measures of the relative effectiveness of teachers within a school could potentially be useful for the managers of individual schools, their impact on the quality of teaching within the school is likely to be quite limited at best. Along with other information, the value-added measures may help school administrators advise teachers on the need for professional development or perhaps on the desirability of leaving the profession. Research by

Jacob and Lefgren, however, suggests that even without value-added measures, school leaders know which teachers are more or less effective within the school—at least at the extremes of the distribution which is all that the value-added models can distinguish in any case (Jacob and Lefgren, 2008). Although the apparent objectivity of value-added measures could potentially be useful in strengthening the case for dismissing a weak teacher, the concern is that they may not pass legal muster in due process hearings because of the large measurement error associated with them. In addition, value-added measures of teacher effectiveness may not be of much use to teachers themselves because of the complexity of the necessary calculations, the fact that they provide no information on what the teachers need to do to improve student outcomes, and the long lag time for value-added reports.

For higher level district or state policymakers, the within-school measures of teacher effectiveness, are not very useful for raising the level of the teacher labor force or improving the distribution of teacher quality across schools, which leads me to my next point.

- **The focus on value-added measures takes attention away from other potential policy levers for increasing the quality and improving the distribution of the teaching force.**

Based on my own research with Duke colleagues (primarily Charles Clotfelter and Jacob Vigdor), I put myself in the camp that believes teacher credentials have sufficiently strong predictive power for them to be policy relevant. That conclusion is based on extensive research—both cross-sectional and longitudinal—in North Carolina, investigating the relationship between teacher credentials at the elementary school level and also at the high school level and student achievement as measured by scores on state tests (Clotfelter, Ladd and Vigdor, 2006, 2007a, 2007b, and 2010). In particular, at the elementary school level we find that the difference between a teacher with a set of very weak credentials and one with average credentials is about 0.15 standard deviations in math and 0.10 standard deviations in reading. At the high school level, we find that the student achievement difference between teachers at the 10th and 90th percentiles of the predicted achievement distribution based on credentials alone is 0.23 standard deviations and that at least a fifth of the overall distribution in teacher quality as measured by student test scores is attributable to the variation in teacher credentials.

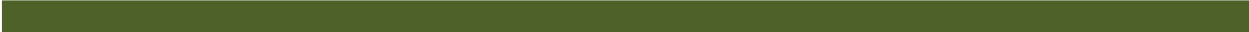
These findings are important for two policy-relevant reasons. The first is that extensive data from North Carolina and other states show that teachers are very unevenly distributed across schools to the detriment of disadvantaged students. At all three levels of schooling in North Carolina—elementary, middle, and high school—the schools serving higher proportions of disadvantaged students have higher proportions of teachers with weak credentials as measured by a wide range of credentials (Clotfelter, Ladd, Vigdor and Wheeler, 2007). Schools with more disadvantaged students, for example, have higher proportions of teachers with limited experience or with training from a less competitive college and lower average licensure test scores. In addition, they have lower proportions of board certified teachers and higher proportions of lateral entry teachers. Similar patterns emerge for other states. Given that teacher credentials are predictive of student achievement, this maldistribution of teachers across schools should be of major concern to policymakers, but it tends to be completely hidden in discussions based on value-added because those measures are typically within-school measures. These patterns imply that if policymakers wish to even out the distribution of effective teachers across schools, they will need to pursue policies designed to alter the distribution of teachers across schools (including, for example, by offering differential salaries or paying more attention to the working conditions, especially the quality of leadership in disadvantaged schools rather than just trying to make individual teachers within schools more effective).

The second reason to focus on teacher credentials is that credentials are potentially more useful than value-added measures as policy levers that policymakers can use to shape the teacher labor force. For example, in our research at both the elementary and high school levels in North Carolina, we find that teachers with higher licensure test scores are more effective in raising student test scores than those with lower licensure test scores. That raises the policy question of how to attract more high-scoring teachers into the profession. Moreover, at the high school level, we find consistent effects for various types of certification. The results indicate that subject specific teacher certification in math and English are predictive of higher test scores in math subjects (algebra and geometry) and English, respectively; that lateral entry teachers are less effective than regular teachers in their initial years, but that the lateral entrants who remain in teaching (far fewer than those who start) are just as effective as those with a regular license; and that the National Board Certification process at the high school level both identifies more effective teachers and promotes better teaching. Working in the other direction, we find that master's degrees, especially for teachers at the elementary level, are not predictive of higher student achievement. Such findings can help policymakers design policies to improve the quality and distribution of teachers.

At the same time, the credentials themselves clearly do not substitute for careful observation of teachers in the classroom. Such observations—potentially supplemented with value-added measures and student evaluations—will be far more useful than value-added measures alone in providing the type of feedback they need to improve their teaching.

References

- Baker, E., L. Darling-Hammond, E. Haertel, H. Ladd, R. Linn, R. Rothstein, R. Shavelson, & L. Shephard (2008). Problems with the use of Test Scores to Evaluate Teachers. Economic Policy Institute Briefing Paper 278.
- Ballou, D., W. Sanders, & P. Wright (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37-66.
- Campbell, D.T. (1979). Assessing the Impact of planned social change. *Evaluation and Program Planning*, 2, 67-90.
- Clotfelter, C.T., H.F. Ladd, & J.L. Vigdor (2006). Teacher–Student Matching and the Assessment of Teacher Effectiveness. *Journal of Human Resources*, 41(4), 778-820.
- Clotfelter, C.T., H.F. Ladd, & J.L. Vigdor (2007a). Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects. *Economics of Education Review* (December).
- Clotfelter, C.T., H.F. Ladd, & J.L. Vigdor (2007b). How and why teacher credentials matter for student achievement. National Bureau of Economic Research Working Paper 12828. Cambridge MA: NBER. Also available on the CALER website (www.caldercenter.org).
- Clotfelter, C.T., H.F. Ladd, & J.L. Vigdor (2010). Teacher Credentials and Student Achievement in High School: A Cross Subject Analysis with Student Fixed Effects. *Journal of Human Resources*. Earlier versions are also available as NBER and CALDER working papers.
- Clotfelter, C.T., H.F. Ladd, & J.L. Vigdor (2010). Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects. *Journal of Human Resources*, 45(3).
- Clotfelter, C.T., H.F. Ladd, J.L. Vigdor, & J. Wheeler (2007). High Poverty Schools and the Distribution of Teachers and Principals. *North Carolina Law Review*, 85(5), 1345-1379.
- Jacob, B., & L. Lefgren (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, 26(1), 101-136.

- 
- Ladd, H.F. (2008). Teacher Effects: What Do We Know? In G. Duncan & J. Spillane, *Teacher Quality: Broadening and Deepening the Debate*. Northwestern University, Multidisciplinary program in the Education Sciences.
- Lockwood, J.R., & D.F. McCaffrey (2007). Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement. *Electronic Journal of Statistics*, 1, 223-252.

Robert Pianta, University of Virginia

“Signal in a noisy system: Can VAM serve a purpose for evaluation and improvement of teaching?”

The last decade has witnessed a marked increase in research examining the nature of children’s experiences in classrooms and the ways in which these experiences uniquely contribute to children’s social, cognitive, and academic development. Evidence in support of classrooms as a focus and lever for policy is fairly strong, and a substantial fraction of this evidence is derived from studies using value-added models (VAMs) as a means of statistically estimating the impact of a teacher on student learning (at least as defined by state standards tests). In thinking about the use of VAMs in the context of policymaking and decisionmaking (high and low stakes) for evaluating and improving the performance of teachers, a couple of considerations arise.

First, VAM scores based on state standards tests are of very limited utility in a system of workforce evaluation and development. There are a number of reasons for this. For example, although state standards tests are reasonable estimates of achievement in a curriculum, they should be better (deeper, more contextualized); are likely to change with Common Core Standards; currently, and for the foreseeable future, apply to only about half of the teacher workforce; and are only one source of information on teacher performance. Moreover, as a criterion assessment that takes place at the end of a year, state tests are a post-test of an outcome desired as a result of instruction that already took place; they carry little if any information that would proactively or formatively shape teachers’ instruction in ways it might improve over the course of a year. In fact, most districts now use an assortment of other tests throughout the year to shape efforts to improve instruction (resulting in students spending increasing amounts of time taking tests). VAM scores’ retroactive relationship to teacher performance, in my view, greatly limits their utility in a comprehensive system of teacher performance evaluation, whether high or low stakes. In low-stakes (improvement-focused) contexts, the complexity of VAM metrics themselves and their retrospective nature make them a poor choice for providing feedback relevant to performance improvement. In high-stakes (evaluation, such as tenure decisions), they are useful only when embedded in a system of annual performance evaluation and improvement in which they trigger a set of supports that provide opportunities for improvement in subsequent years.

A second consideration is a more general one that applies to understanding and estimating the contribution of a teacher to student learning and development. Fundamentally, assessing a teacher’s contribution to student learning is a signal-to-noise challenge. Classrooms and learning are inherently noisy systems with lots of moving parts. Perhaps no study illustrates this reality as clearly as the Gates Foundation’s Measures of Effective Teaching (MET) study, precisely because it included a number of assessment methods and approaches and directly tackled the task of estimating teacher-related signal within and across all these assessments. Given the noisiness associated with estimating an individual teacher’s contributions to student learning, the challenge is how to extract information (signal) that is reliable and valid for particular purposes. Much of the extant literature on VAM focuses on its signal detection properties and capability. Other briefs in this compilation do a good job of summarizing and highlighting those properties. It is quite clear across many studies that it is not unreasonable to infer that VAM does carry some signal regarding an individual teacher’s contribution to student learning. That conclusion is qualified by a range of other considerations (e.g., school effects) and appears to apply best to the middle 70-80 percent of the distribution. This means that applications of VAMs for high-stakes decisions, in and of itself, might be rather limited with regard to the confidence that it is carrying enough signal to warrant a particular decision (such

as denying tenure). It does not mean that VAMs could not be applied in a more comprehensive, multi-measure system of evaluation and decisionmaking that could include high-stakes decisions. It simply means that the inferences being drawn about any measure or cluster of measures align well with the technical properties of those instruments as they pertain to that decision. That is, we want models of assessment and decisionmaking in which the signal value of a measure maps to the purpose for which that measure is being used.

Third, and related to the point on signal and noise above, we know there are other sources of information on teacher performance that can be standardized and scaled and there are other criterion variables that perhaps should be assessed, such as student engagement and motivation, social skills, or even attendance, because these are also the public's goals for the education of students and for what they view as the aims of teachers' contributions to their children's development. This reality, that there are non-VAM ways to estimate teachers' contributions (including quantitative ratings of classroom behavior and student surveys) and that there is a broader portfolio of outcome assessments suggests that VAMs are only one slice of the signal bandwidth (to extend the metaphor). In a comprehensive system of evaluation and improvement, each assessment—of teacher performance and of student outcomes—might map onto different aims or features of evaluation and improvement (e.g., observations may be more useful for improvement of practice). The research on VAMs has been extraordinarily helpful because of the intense focus on the accurate estimation of effects, isolation of a wide range of correlated factors, and a level of confidence about inference that is not typical of the research literatures on many of the other complementary assessments. Again, MET has helped raise the bar for other assessments, such as observation and student surveys, in this regard and the literature on observation in many ways now reflects some of the technical and inferential considerations (e.g., stability, validity, causal impacts) that have characterized VAM studies. All this seems good and likely to advance estimation and decisionmaking and ultimately lead to systems of evaluation that have fewer unintended consequences because they have a broader base of signal estimates.

All this leads to a focus on how VAMs might contribute to decisionmaking both in high-stakes and low-stakes contexts. The problems and possibilities associated with using VAMs as a sole source of information in high-stakes decisions are well described in other memos in this compilation and in references noted below. Yet states and districts are desperate for actionable information to improve the teaching workforce, and many are driving ahead with using VAMs in ways that expose its flaws, or in using alternative models (such as observation or surveys) that rely on rather poor technical properties (e.g., using locally developed observations with little to no technical information on reliability or validity). These efforts are rolling forth now and many tens of thousands of teachers will be affected soon. The focal challenge for now is not VAMs or no VAMs, it's building capacity to make wise, defensible decisions using a portfolio of data on teacher performance (of which VAMs could be a part) and rolling these systems out in ways that their properties for various decisions (tenure, rewards, assignments) are known to some degree before high stakes are attached.

References

- Allen, J., Pianta, R., Gregory, A., Mikami, A., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, 333, 1034.
- Chetty, R., Friedman, J.N., & Rockoff, J.E. (December, 2011). The Long-term Impacts of Teachers: Teacher Value-added and Student Outcomes in Adulthood. National Bureau of Economic Research.

- Gates Foundation. (2012). *Gathering Feedback to Improve Teaching*. Bill and Melinda Gates Foundation, Seattle.
- Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review*, 100(2), 250-255.
- Gordon, R., Kane, T.J., & Staiger, D.O. (2008). *Identifying effective teachers using performance on the job*. Washington, DC: The Hamilton Project, The Brookings Institution.
- Hanushek, E., & Rivkin, S. (2010). Generalizations about using value-added measures of teacher quality." *American Economic Review*, 100(2), 267-271.
- Ladd, H.F. (2008). *Teacher effects: What do we know?* Unpublished manuscript, Teacher Quality Conference, Northwestern University.
- Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26, 237-257.
- Pianta, R.C., & Hamre, B.K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109-119. doi:10.3102/0013189X09332374
- Rivkin, S.G., Hanushek, E.A., & Kain, J.F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417-458.
- Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2), 247-252.
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement," *Quarterly Journal of Economics*, 125(1).

Jonah Rockoff, Columbia Business School

The two properties of any estimator in which researchers are generally most interested are bias and reliability (or precision). Value-added—an estimate of a teacher’s performances in raising student achievement test scores—is no exception. I begin by focusing on the recent literature on bias and reliability in VA measures. At the end of the brief, I highlight a number of other issues and recent studies that I think have presented significant findings.

The key issue for bias is whether variation in test scores (after controlling for student and classroom characteristics) is driven at least in part due to persistent differences in the students assigned to different teachers.¹¹ An important study by Rothstein (2010) raised a flag that such bias might be present and caused a stir among researchers, essentially by showing that students’ future teacher assignments were correlated with their prior test score gains, even after applying controls that researchers had used in prior studies.

Since Rothstein, research has been generally supportive of unbiased VA estimates. Koedel and Betts (2011) and Goldhaber and Chaplin (2012) focus directly on Rothstein’s econometric test for bias, presenting evidence that his results need not indicate bias but may be due to idiosyncratic noise or even subtler issues regarding test non-linearity. Kane and Staiger (2008) present the results of an experiment where VA measured in previous years accurately predicted student test scores under random assignment.¹² This provides powerful evidence that their VA estimates were unbiased, though the external validity of their experiment is uncertain. Finally, Chetty, Friedman, and Rockoff (hereafter CFR, 2012), create quasi-experimental tests for bias in VA measures and find little evidence of significant bias using a large dataset from an urban district. My overall interpretation of the literature is that VA typically provides unbiased estimates of teachers’ skill in raising test scores and, while there is no guarantee that VA measures will be unbiased in every setting, the magnitude of such bias is likely to be small.

Reliability is just as important as bias, though it receives less attention from researchers. Findings have been pretty consistent on the extent to which variation in VA reflects real differences in teachers that persist over time. Year-to-year correlations in VA estimates range from 0.3 to 0.5, which is similar to performance statistics in other settings; for example, year-to-year correlation in batting average for professional baseball players is roughly 0.4.

The only controversy on this topic is whether VA reliability is enough to be useful in evaluation. In a 2010 paper, Staiger and Rockoff use a straightforward framework to show that, at these levels of reliability, there are large potential gains from the use of VA for teacher personnel decisions. Baker et al. (2010) and Corcoran (2010) are less sanguine about the usefulness of VA given its reliability, but offer no analytic framework to support their conclusions.

Thus, researchers agree on the reliability of VA—positive, significant, but far less than perfect—and only disagree on whether the glass is half-empty or half-full. The main question, in my opinion, is whether there exist other valid measures of teacher effectiveness that are so reliable so as to make VA redundant. Here, all the evidence suggests that this is not the case. For example, the Gates MET project (2012) makes it clear that class observation reliability is only moderate, and Rockoff et al. (forthcoming) show that principals who learn about teachers’ VA

¹¹ If VA is unbiased, then this variation must be driven only by teacher performance and idiosyncratic noise (e.g., test measurement error, idiosyncratic class chemistry, etc.).

¹² The 2008 date on the paper by Kane and Staiger is based on their working paper, whereas 2010 reflects publication of Rothstein’s paper in a peer reviewed journal. Rothstein’s original findings were available as a working paper as early as 2007.

estimates incorporate them into their holistic evaluations. Thus, VA has a role to play, and the size of its role should be proportional to its accuracy relative to other performance measures.¹³

Teaching to the Test

In addition to their work on bias, CFR present compelling evidence that the test score gains attributable to high VA teachers also improve students' academic and labor market outcomes later in life. This proves that their VA measures capture something that goes beyond whether some teachers "teach to the test," while others focus on non-tested (but just as valuable) material. The value of a high VA teacher could diminish once VA is used for high-stakes evaluations (e.g., by inducing more test preparation or cheating that improve scores in a way that provides little of real value), but this will have to be examined in future research.

Portability of VA

An important question is the extent to which the skills/abilities captured by VA are very specific to the teaching context or whether teachers measured to be high VA would perform well with different student populations or in a different subject or grade level. The quasi-experimental estimates shown by CFR suggest that, to some extent, VA is portable. However, the best evidence on this question is likely to come from the Talent Transfer Initiative study, being conducted by Mathematica, which I imagine could be available within the year.

High School Teachers

Most research on VA is from elementary and middle schools, and research is mixed on the usefulness/validity of VA in high school (Clotfelter et al. 2007; Jackson, 2012). The key seems to be whether tracking and team teaching are so extensive in high school that VA methods cannot adequately separate the contributions of individual teachers.

Test Construction Issues

Papers by Neal (2011) and Lang (2010) present strong arguments for why test scaling can present problems for VA measures and suggest ordinal measures (e.g., student growth percentiles) might be more appropriate in evaluations. Neal also argues that problems of gaming and teaching to the test motivate the use of a different test for teacher evaluation than for measuring student performance levels. Koedel and Betts (2011) present evidence that too many students scoring at a test ceiling creates problems for VA estimates.

VA in Pay for Performance

Research on the use of VA as an input into performance pay plans is mixed (Springer et al., 2011; Imberman and Lovenheim, 2012). My take on the literature is that the structure of teacher performance pay is crucial, and one concern with the use of VA is that teachers might not easily understand the link between their actions and VA measures. This concern might diminish over time as VA becomes more common (and commonly understood).

¹³ Another consideration is whether VA is useful as a tool for helping teachers improve their craft. We have no good evidence on this issue, but my sense is that VA—which is typically only available after the school year ends and is calculated using complicated statistical formulae—is unlikely to be useful as formative feedback, particularly compared with real-time feedback from classroom observation.

References

- Baker, E.L. et al. (2010). Problems with the use of student test scores to evaluate teachers. Economic Policy Institute Briefing Paper 278. Retrieved from <http://www.epi.org/publication/bp278/>
- Chetty, R., Friedman, J.N., & Rockoff, J.E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. National Bureau of Economic Research Working Paper #17699. Retrieved from <http://www.nber.org/papers/w17699.pdf>
- Clotfelter, C.T. et al. (2007). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. National Bureau of Economic Research Working Paper #13617. Retrieved from <http://www.nber.org/papers/w13617.pdf>
- Corcoran, S. (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. Providence, RI: Annenberg Institute for School Reform.
- Gates Foundation (2012). Gathering Feedback to Improve Teaching. Seattle, WA: Bill and Melinda Gates Foundation.
- Goldhaber, D., & Chaplin, D. (January, 2012). Assessing the "Rothstein Test": Does it really show teacher value-added models are biased? CALDER Working Paper #71.
- Imberman, S.A., & Lovenheim, M.F. (2012.) Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. National Bureau of Economic Research Working Paper #18439. Retrieved from <http://www.nber.org/papers/w18439>
- Jackson, C.K. (2012). Teacher quality at the high-school level: The importance of accounting for tracks. National Bureau of Economic Research Working Paper #17722. Retrieved from <http://www.nber.org/papers/w17722>
- Kane, T., & Staiger, D.O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Cambridge, MA, NBER.
- Koedel, C., & Betts, J.R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*, 6(1), 18-42. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/EDFP_a_00027
- Lang, K. (2010). Measurement matters: Perspectives on education policy from an economist and school board member. *The Journal of Economic Perspectives*, 24(3), 167-181.
- Neal, D. (2011). The design of performance pay in education. National Bureau of Economic Research Working Paper #16710. Retrieved from <http://www.nber.org/papers/w16710>
- Rockoff, J.E., Staiger, D.O., Kane, T.J., & Taylor, E.S. Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review*, 102(7).
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1), 175–214. doi:10.1162/qjec.2010.125.1.175
- Springer, M.G. et al. (2011). Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching. National Center on Performance Incentives. Retrieved from https://my.vanderbilt.edu/performanceincentives/files/2012/09/POINT_REPORT_9.21.102.pdf

Jesse Rothstein, University of California, Berkeley

There has been an explosion of research on statistical measures of teacher effectiveness, but this research has yielded remarkably little insight into the design of better approaches to teacher evaluation. It has focused almost exclusively on the statistical properties of value-added models (which I use as shorthand for a broader class of methods, including student growth models and others) in settings in which individual teachers have little or no stake in the outcome. We have learned a great deal about the limitations of VAMs in these settings, but the questions that Rubin, Stuart, and Zanutto (2004) urged us all to focus on remain the most important ones: What reward structures and other policies should we implement based on value-added models? And what will be the effects of these policies? These questions remain under-studied. We know little about whether VAMs can be used to improve real world teacher evaluation, and if so, how.

Because so much research has focused on the properties of VAMs in low-stakes settings, I will review the stylized results, then return to discuss policy design and policy effects.

Properties of VAMs in Low-Stakes Settings

Relying primarily on data from districts that either were not computing VAM scores for their teachers or were not using those scores in any meaningful way, researchers have learned a great deal about the statistical properties of VAM scores in these settings.

- Annual VAM scores for individual teachers are quite noisy. Even multi-year averages remain noisy, and annual fluctuations undercut the scores' face validity.
- VAMs do not achieve their original goal of separating the component of student achievement that is due to the teacher from other influences on end-of-grade scores. Rather, most classroom assignments are non-random in ways that violate the strict formal requirements of the available VAMs—both the simpler ones that are widely used and the more complex models explored by some researchers. We do not know whether this creates substantial biases in most teachers' VAM scores, but all of the available evidence is consistent with this.
- Teachers' VAM scores are positively correlated with principals' assessments, with structured classroom observations, and with student evaluations, but the correlations are quite weak—too weak to be consistent with the often-espoused view that we needn't worry about the limitations of VAMs because any measure at all will successfully identify the worst teachers.
- When students are administered two tests with different emphases or styles (e.g., open response vs. multiple choice), the resulting teacher VAM scores are positively correlated but only weakly—disattenuated correlations are around 0.4. Similarly, VAM scores computed from different subtests of the same test are only weakly correlated.
- There is some evidence that teachers' VAM scores are importantly unstable over time—that even after adjusting for annual noise some teachers improve over the years while others get worse. We do not have good statistical models for examining this phenomenon or for incorporating it into teacher evaluations; current approaches are predicated on the assumption that VA is stable but for idiosyncratic noise.
- Whatever impacts are measured by VAMs fade out remarkably quickly—students who have a high-value-added teacher in one grade get higher scores that year but see much smaller improvements to their subsequent scores. Moreover, some teachers have low initial effects but larger longer run effects. Neither the average fadeout nor the heterogeneity in long-run effects is well understood. One study found a suggestive correlation between the initial effects and students' later wages, but the reliability of the short-term VAM score as a proxy for teachers' effects on students' long-run earnings has not been measured.

Two high priority questions that we have learned basically nothing about are:

- How can teachers' VA be compared across heterogeneous contexts (i.e., schools)? Policy uses of VAM scores require making comparisons between high- and low-poverty schools. But research VAMs typically say nothing about these comparisons.
- How do teachers in different grades interact? All extant VAMs assume that one can ignore cross-grade interactions, merely adding up the value-added of the 3rd grade teacher, the 4th grade teacher, and so on. This assumption is totally unfounded. It is not clear whether a richer model that allows for interactions will yield similar or wildly different estimates.

Policy Applications of VAM

Nearly all of the above results derive from settings in which schools may face high stakes but teachers are not accountable for their own students' scores. VAM measures will deteriorate—will become less reliable and less closely tied to true effectiveness—if they are used for high-stakes individual decisions. It has been shown repeatedly that school accountability leads to deterioration, but we do not know how much worse it will get when teachers face individual stakes. How much will teachers change their content coverage (in desired or in undesired ways), neglect non-tested subjects and topics, lobby for the right students, teach test-taking strategies, and cheat outright? And to what extent will these responses prevent us from distinguishing effective from ineffective teachers? We simply don't know.

We also don't know much about how to design a teacher evaluation system predicated on VAMs. In other comparable occupations evaluation systems are more often formative than summative; they keep stakes low and they rely intensively on highly skilled and highly compensated managers. There has been little exploration of such programs in education, and little research into potential formative uses (if any) of VAM scores.

Education policymakers have focused more on high-stakes incentive pay or non-retention policies. There have been a few experiments with individual- or group-level incentive pay, but they have yielded overwhelmingly disappointing results. We don't know how to design a retention policy that will use VAM scores in a wise way or how changes in contracts will change recruitment to the profession. A recent paper of mine suggests that pay-for-performance programs are unlikely to lead to large changes in the quality of teacher recruits, and that non-retention policies can do so only if accompanied by large increases in teacher salaries. But even this derives from simulations rather than from empirical evidence. We need to put a lot more effort into designing new human resource policies. We also need to study their effects when implemented. High-stakes teacher evaluations have been implemented in only a few districts; the impacts of those few programs have not been carefully studied; and the programs have not been designed to permit rigorous evaluation. We need to ensure that, as we embark on a massive national experiment with alternative teacher evaluations, we will learn something from the experience.

References

Rubin, D.B., Stuart, E.A., & Zanutto, E.L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, 29(1), 103-116. doi: 10.3102/10769986029001103