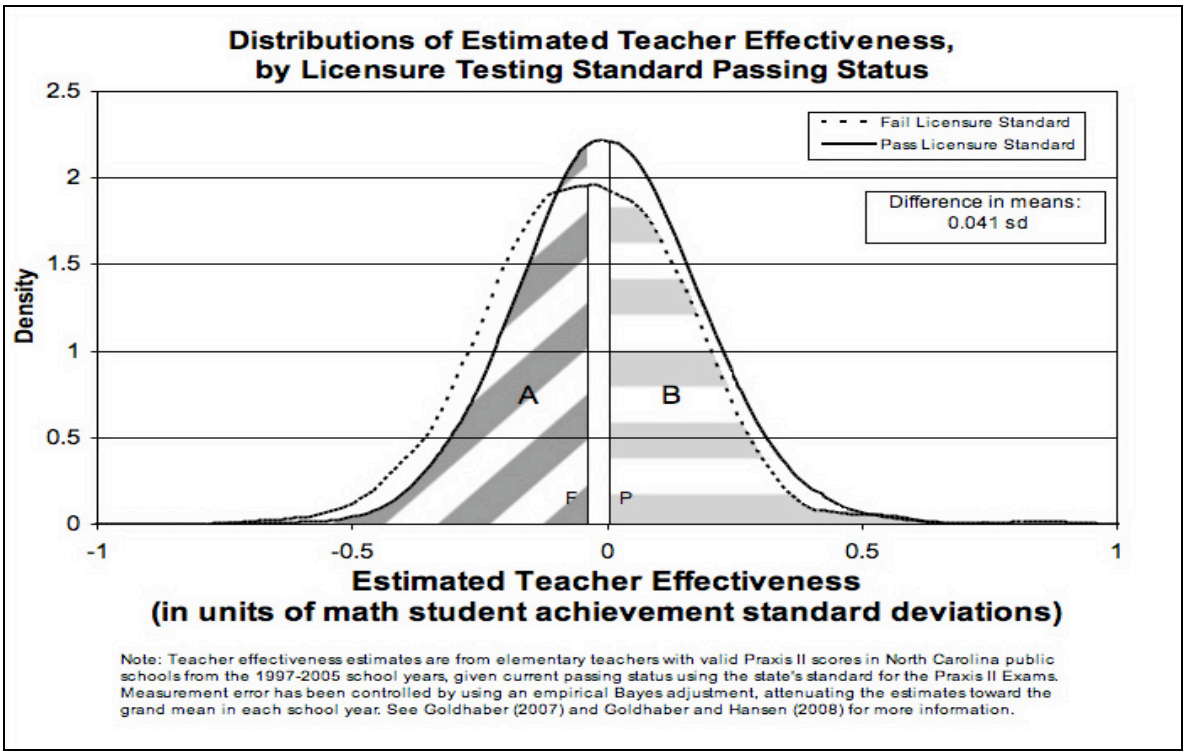
The Promise and Potential Pitfalls of Value-Added Assessment for High-Stakes Personnel Decisions

Dan Goldhaber
Center on Reinventing Public Education
University of Washington

Why VAM?

A growing body of research demonstrates that teacher quality is the most important *schooling* factor influencing student achievement. The difference between having a very effective teacher and a very ineffective teacher may be as great as a full year in learning growth (Hanushek, 1992). In statistical terms, estimates suggest, for example, that a one standard deviation increase in teacher quality raises student achievement in reading and math by 10 to 25 percent of a standard deviation. To put this in perspective, this teacher quality effect is roughly equivalent to lowering class size by 10 to 13 students (Rivkin et al., 2005).

Yet, inputs currently used to measure and evaluate teacher quality - licensure status, degree and experience level - are only weakly linked with teacher effectiveness (Goldhaber, 2002). In fact, the differences between the best and worst teacher who hold a particular credential far outweigh the differences between those with and without the credential. For example, Kane et al. (2006) find that the gap in teacher effectiveness (measured by value-added) within each certification category is about ten times larger than the average gap between certification categories. In work on the relationship between teacher performance on licensure tests and student achievement (Goldhaber, 2007), I find similar results. **Figure 1** below is derived from this work. This figure shows the estimated distribution of effectiveness for teachers who passed the licensure exams required for full certification in North Carolina (the solid line) and those who failed them (the dotted line).



As we might expect, on average, teachers who achieved the standard are more effective than those who do not, by about 4 percent of a standard deviation (the horizontal distance between F and P). But, there is also considerable overlap (more than 90%) in the effectiveness distributions. For example, all of the teachers shown in area **A** passed the required licensure tests but were *less* effective than the average teacher who failed to achieve the standard; teachers shown in area **B** failed the standard, but were *more* effective than the average teacher who passed.

What this all means is that the research has painted an elusive picture when it comes to thinking about upgrading the quality of the teacher workforce through changes in teacher *inputs*. In theory one could leave the process of determining teacher effectiveness and acting on those determinations up to local school districts, but there is mounting evidence that teacher evaluations are typically less than rigorous (Toch and Rothman, 2008), that few teachers ever receive anything but a stellar evaluation, the top score on any rating rubric (Tucker, 1997), and there are seldom consequences for being an ineffective teacher (The New Teacher Project, 2009). Note that it is not necessarily the case that local administrators don't know who the more effective and ineffective teachers are (Jacob and Lefgren, 2005), rather that school culture, politics, and cost considerations (the cost of dismissing poor performing teachers) leads to a situation aptly described by the New Teacher Project (2009) where teachers are treated like "widgets," in the sense that their performance is all the same.

Value added models (VAMs) offer policymakers a potential work-around what appears to be the very unsatisfying teacher-as-widget system in the sense that VAMs provide a cost-effective, objective way of fairly evaluating teachers, presumably so the teacher effect estimates can be used for policy purposes. There is, however, a fair amount we don't know about the

properties of value added models or using VAM effects. In what follows I describe some of the concerns that arise and the research that either supports or refutes those concerns. I close by reminding readers that the relevant alternative (today's teacher policy system) appears to be far from a nirvana.

Concerns Over VAMs and the Use of VAM Estimates

In this short piece I will not go into great detail about any of the issues that arise in the context of using value-added models to inform personnel decisions. I urge readers wishing a more comprehensive overview of the issue to see the forthcoming issue of *Education Finance and Policy* (Volume 4, Issue 4), which is devoted to the topic, and to read an exchange between Doug Harris and Heather Hill in the most recent edition (Volume 28, Issue 4) of the *Journal of Policy Analysis and Management*.

Concerns about VAMs and the use of them for policy fall into roughly six categories: 1) the use of student achievement tests as a measure of teacher effectiveness; 2) bias in estimates of teacher effects; 3) measurement error; 4) perverse incentives; 5) defining the counterfactual; and 6) logistical issues.

Using Student Achievement on Standardized Tests as a Metric for Teacher Effectiveness

The idea of using student achievement on tests to make any kind of judgment about schools is controversial. Standardized tests only measure a sample of what students are learning in the classroom (Koretz, 2008). Thus, they can only be used to assess a subset of teaching objectives (and even these objectives will be measured with error, discussed below). In short, we really care about the contributions that teachers are making toward a student's later life outcomes: Do they go to college?; Are they successful there?; Are they likely to be gainfully employed?; What wage can they command? Performance on tests is, at best, a proxy for these types of outcomes, and the question is whether or not it is a good proxy. Of course there is not research connecting performance on all the myriad assessments used by states to all of these outcomes, but the research that does exist connecting this measure of student achievement to later life outcomes suggests that tests are an important predictor of college-going behavior, employment probability, earnings and a host of other non-financial measures (e.g. Grogger and Eide, 1995; Murnane et al., 1995). Moreover, more recent research has shown that aggregate student test achievement is related to national competitiveness (Hanushek et al., 2008).

It is ultimately a political question as to whether student test performance ought to be used as a metric for measuring teacher or schools success, but I would argue there is good evidence supporting this type of outcome measure. Moreover, under No Child Left Behind student performance on state assessments is already being used for this purpose when it comes to judging schools.

Bias in Estimates of Teacher Effects

Efforts to identify the causal impacts of teachers, i.e. their contribution toward student achievement, are complicated by the complexity of the learning environment and process. VAMs rely on strong assumptions about the nature of student learning over time (Todd and Wolpin, 2003). In a perfect world we would observe all differences in student and school quality and control for these differences when measuring teacher effects. But VAMs can only control for *observed* differences in student background, school quality, and peer effects. We don't observe other factors that may be related to the nonrandom distribution of students and teachers.

And failure to account for these factors can lead to biased VAM estimates. There is some evidence, for instance, that the nonrandom assignment of students to teachers and its relationship to other factors that influence student achievement might lead typical VAMs to misestimate the contributions that teachers make (Rothstein, 2009). For example, imagine that parents respond to the quality of the teacher their students have. In years in which their students are assigned to a lower quality teacher they might help their children more with homework or get them supplemental tutoring, whereas when assigned to a really effective teacher they might ease off the pedal assuming their children are in good hands at schools. Further assume that none of this is directly observable in the sense that it is not accounted for in statistical models.

If students were randomly assigned to teachers this type of situation might not be a problem, but if the assignment is based on student achievement, say principals assign lower performing students to their best teachers, then this type of parental "compensating behavior" would lead to biased VAM estimates. In this scenario it would lead to underestimates of the true contributions of effective teachers and overestimates of the true contributions of ineffective teachers, but note that a different type of sorting or parental behavior could result in bias in the oppose direction.[1]

Measuring and partitioning credit for teacher contributions is more complicated in some contexts than in others. At the high school level, for instance, students typically switch between teachers as they move from one class to another and it's entirely possible that the performance of students in one subject might be influenced by the quality of teachers in one or more other subjects (Koedel, 2009). Research is really only at the leading stages of understanding these issues, but there is some reason for optimism since a recent study by Kane and Staiger (2008) shows that VAM teacher effect estimates in an experimental period, when teachers are randomly matched to their classrooms, are similar to estimates derived under nonexperimental conditions.

Measurement Error

Even if VAM teacher effects are unbiased, they are certainly measured with error, a situation often referred to as being "noisy". Tests themselves are inherently noisy measures of student knowledge and thus when used in models to assess teacher effects will provide imprecise measures of true teacher contributions. The precision of estimates of teacher value added will depend on the number of students taught (the class size and/or number of years of data in the VAM). And imprecision is clearly an issue because imprecise estimates mean that any "cut-point" policies *will always* result in errors (false positives and false negatives). For example, were teachers falling into the top quintile of performance based on VAM estimates to be awarded a performance bonus, the teacher at the 89th percentile could make a very good statistical argument that there is really no difference between the estimate of her effect and the estimate for the teacher at the 90th percentile who received a bonus.

Measurement error in the context of teacher effect estimates has received a good deal of research attention of late. Studies, for instance, show that teachers, when grouped into performance quartiles or quintiles tend to jump from one quintile to the next from year-to-year, and some percentage from the top to bottom categories or bottom to top (e.g. Aaronson et al., 2007; Ballou, 2005; Koedel and Betts, 2007). One cannot definitively know whether these movements represent true changes in performance from one year to the next or instability that results from measurement error.

[1] VAMs have been shown to fail various falsification tests (Rothstein, 2009), heightening concerns about bias.

In general, various estimates suggest the intertemporal stability of teacher effect estimates in the 0.3 to 0.5 range, and while these correlations have been categorized as small, Goldhaber and Hansen (2008) note that they are not very different from estimates of job performance in sectors of the economy that consider them for high-stakes purposes (such as job retention and pay determination). There is no real solution to the problem of measurement error since it is inherent in any statistical model, but policymakers do face tradeoffs when using VAMs. Multiple years of job performance data improve the reliability of our estimates because they provide more information and enable us to use more sophisticated statistical approaches, but multiple year estimates also dampen incentives and mean that there is no VAM information on first-year teachers.

Perverse Incentives

One of the major concerns about using VAMs for policy purposes (the above sub-sections just deal with the estimates themselves) is that their use might create perverse incentives for teachers. For example, teachers worried about improving student scores in reading and math may forgo non-tested subjects such as history, music and art or narrow their instruction to unimportant test skills not associated with greater comprehension and understanding. And when teachers start to view their peers as competitors, collaboration is less likely (Murnane and Cohen, 1986). Some empirical work even suggests that incentives linked to student achievement on tests leads to corrupt teacher behavior such as cheating or the intentional exclusion of potentially low-scoring test takers (Jacob and Levitt, 2003; Cullen and Reback, 2006; Figlio and Getzler, 2002). While these are legitimate concerns, there little evidence to suggest that high-stakes performance incentives have lead to *widespread* problems and there are certainly ways that one could design incentive programs to avoid these types of problems (e.g. have strong sanctions against cheating or a schoolwide component of a teacher incentive).

Defining the Counterfactual

VAMs allow for comparisons of teachers within schools or districts or the state as a whole. However, between school or district teacher comparisons will necessarily conflate school or district effects and teacher effects. For example, a comparison of teachers at the district level will combine teacher effects with school effects so teachers in schools with very effective principals or a working environment conducive to student achievement will, for instance, benefit by having somewhat larger teacher effects. This argues for more localized comparisons between teachers. However, comparisons that are more localized, at the school-level, for instance, are also potentially problematic. The most effective teacher in school A might, for instance, be less effective than the least effective teacher in school B. Moreover, in small districts or schools there may not be enough teachers teaching in the same context to enable models to estimate reasonable apples to apples teacher effects (e.g. there may be only one teacher in a school teaching an advanced section of math).

Defining the counterfactual involves inescapable tradeoffs, but research on teachers does offer some guidance. Most of the variation in teacher effectiveness is estimated to be *within* school variation (Gorden et al., 2006), implying that school or district effects are relatively unimportant. So, while it's a value judgment, I would argue for comparisons at the highest level in which students are subject to the same assessments.[2]

[2] This standard would therefore preclude cross-state comparisons since information on student achievement is typically based on state assessments.

Logistical Issues

I'll now touch upon some of the logistical challenges that arise for those wishing to use student test scores as a metric to judge teacher effectiveness. While certainly not limited to these, they include: the timing of student achievement tests so that "gains" in achievement correspond to particular teachers (e.g. tests administered mid-year are problematic); at what point student achievement ought to be attributed to particular teachers (e.g. if a student enters a teacher's class in October); how to handle cases (primarily at the elementary level) where teachers appear to be differentially effective in instructing students in different subjects; the long-lag that often occurs between the time that students are assessed and VAM estimates are available; and how to judge teachers in non-tested areas where VAMs do not apply.

Of the above, I will only expand on the last issue, what to do about non-tested grades and subject areas as this issue has not, at least to my knowledge, received much attention. There are of course myriad ways to evaluate teachers - supervisory or peer assessments, review of teacher portfolios or observations of teachers - to name just a few. A few of these have been validated against student achievement (see Toch and Rothman, 2008, for a review). What I want to suggest here, and it is mere speculation, is that the use of VAM assessments would have a spillover impact in non-tested areas. Specifically, I would argue that the use of VAM would encourage more rigorous evaluations across the board and make it less likely that teachers would receive uniformly top ratings, as they typically do today, either in the grades and subjects covered by VAMs or in non-tested areas. The reason is that it would seem incongruous if either VAM showed a spread of teacher effectiveness and a non-VAM means of assessing the same teachers presented a very different picture, or if teacher ratings only varied in grades and subjects covered by VAM assessments.

In the Eye of the Beholder: What is the Relevant Alternative?

The discussion in the prior section may lead readers to the conclusion that I would not favor using VAM assessments for policy purposes. There are certainly a lot of open questions when it comes to how VAM teacher effect estimates might or ought to be used. But we also know that the status quo is not very appealing. As I noted in the introduction, there exists substantial evidence that we are currently making some bad investments by relying on input-based teacher policies. The master's pay premium is a particularly stark example. Teachers typically get an annual salary increase of about $5000 for receiving their master's degree, but we know that under current policies (where any master's degree is rewarded) that the credential is no better than a flip of a coin when it comes to predicting teacher effectiveness.

Some of the issues that arise with VAMs exist in the context of today's input-based teacher policies. Licensure tests, which are commonly used to determine teacher employment eligibility, have the same cut-score problem I described above in the context of VAMs: prospective teachers immediately above and below the cut-scores established by states are unlikely to be truly different from each other, but those below the cut-score are deemed ineligible to teach. In the context of teacher evaluations, VAM estimates of teacher effectiveness need not be perfectly accurate, just substantially more accurate than what is used today. That is not hard to imagine. There is clearly a Lake Wobegon effect when it comes to teacher evaluations - so much so that research shows that a substantially higher portion of teachers receive better ratings than their supervisors or peers feel they merit (Tucker, 1997). And there are high-profile stories of the tens of millions of dollars that large districts are spending on teachers housed in "rubber rooms"

because no principals want them in their schools, but the system cannot easily dismiss them (often because there is no established track record of poor performance).

When focusing only on the potential warts of VAMs, it is easy to forget that the current input-based teacher policy system also suffers from many of the issues described in the "Concern Over VAMs" section, and, as a consequence policymakers may hold alternatives like VAM up to a higher standard of evidence than the status quo. Moreover, there is substantial evidence that the current system used to determine employment eligibility and compensation is not working well to attract and keep the most talented people in teaching. There has been a long-term downward trend in the academic skills of the U.S. teacher workforce (Corcoran et al., 2004). And today, on average, teachers graduate from less-selective undergraduate institutions, have lower standardized test scores (Ballou, 1996; Goldhaber and Liu, 2003; Hanushek and Pace, 1995), and require more remediation in college (U.S. Department of Education, 1996). As Richard Murnane and colleagues (1991) summarize, "college graduates with high test scores are less likely to take [teaching] jobs, employed teachers are less likely to stay, and former teachers with high test scores are less likely to return."

We do not yet know whether alternatives to the status quo that utilize VAMs would have long-lasting positive impacts on the quality of the teacher workforce, but there is reason for optimism given the well-established labor economics principal that occupations that reward individuals for key outcomes will draw in those individuals whose skill sets give them an advantage in producing those outcomes. And using VAM teacher effect estimates for policy purposes – e.g. to help determine eligibility to remain in the teacher workforce, or to allocate performance awards, or to direct professional development offerings – offers the promise of a much tighter connection between teacher policies and investments and student achievement.[3] This is clearly a value judgment, but the position that I would take is that the input-based system is pretty clearly unappealing in terms of its connection to student outcomes so it presents a low bar for considering alternative VAM-based teacher policies.

[3] For a more thorough discussion of the potential virtues of such a system, see Goldhaber (2006).

References

Aaronson, D., L. Barrow, and W. Sanders (2007), "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics*, Vol. 25, No. 1, pp. 95-135.

Ballou, D. (2005a), "Value-Added Assessment: Controlling for Context with Misspecified Models." Paper presented at the Urban Institute Longitudinal Data Conference, March 2005.

Ballou, D. (1996), "Do Public Schools Hire the Best Applicants?" *The Quarterly Journal of Economics*, Vol. 111, No. 1 pp 97-133.

Corcoran, S.P. W.N. Evans, and R.M. Schwab (2004), "Changing Labor-Market Opportunities for Women and the Quality of Teachers, 1957–2000." *American Economic Review*, Vol. 94, no. 2, pp. 230-235.

Cullen, J.B., and R. Reback (2006), "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System." NBER Working Paper No. W12286, National Bureau of Economic Research, Cambridge, MA.

Figlio, D. N., and L. S. Getzler (2002), "Accountability, Ability, and Disability: Gaming the System." Working Paper 9307. National Bureau of Economic Research, Cambridge, MA.

Goldhaber, Dan. 2002, "The Mystery of Good Teaching." *Education Next*, Spring 2002.

Goldhaber, D. and A. Liu (2003), "Occupational Choices and the Academic Proficiency of the Teacher Workforce." In *Developments in School Finance 2001–02*, edited by William Fowler. Washington, DC: NCES, pp. 53-75.

Goldhaber, Dan. (2006), "Teacher Pay Reforms: The Political Implications of Recent Research." Prepared for theCenter for American Progress, December 2006.

Goldhaber, D. (2007), "Everyone's Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness?" *Journal of Human Resources*, Vol. 42, No. 4, pp. 765-794.

Goldhaber, D. and M. Hansen (2008), "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions." CRPE Research Brief. Center on Reinventing Public Education, Seattle, WA.

Gordon, R., T. J. Kane, and D. O. Staiger, "Identifying Effective Teachers Using Performance on the Job." Discussion Paper 2006-01, The Hamilton Project. April 2006.

Grogger, Jeff, and Eric Eide. "Changes in College Skills and the Rise in the College Wage Premium." *Journal of Human Resources* 30, no. 2 (1995): 280-310.

Hanushek, Eric A., 1992, "The trade-off between child quantity and quality." *Journal of Political*

Economy Vol. 100, pp. 84-117.

Hanushek, E.A., and R.R. Pace (1995), "Who chooses to teach (and why)?" *Economics of Education Review*, Vol. 14, pp. 101-17.

Hanushek, E., D.T. Jamison, E.A. Jamison, and L. Woessmann (2008), "Education and Economic Growth." *Education Next,* Vol. 8, No. 2, pp. 62-70.

Jacob, B. and L. Lefgren. (2005). "Principals as Agents: Subjective Performance Measurement in Education." National Bureau of Economic Research Working Papers: 11463.

Jacob, B. A., and S. D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics*118: 843–78.

Kane, Thomas J. , Jonah E. Rockoff, and Douglas O. Staiger. "What Does Teacher Certification Tell Us About Teacher Effectiveness? Evidence from New York City." In *Working Paper Series*. Stanford, CA: National Bureau of Economic Research, 2006.

Kane, Thomas J., and Douglas O. Staiger, 2008, Are teacher-level value-added estimates biased? An experimental validation of non-experimental estimates, National Conference on Value-Added Modeling (University of Wisconsin at Madison).

Koedel, C. (2009). "An Empirical Analysis of Teacher Spillover Effects in Secondary School." *Economics of Education Review*.

Koedel, C. and J.R. Betts (2007), "Re-examining the role of teacher quality in the educational production function." University of Missouri, Columbia, MO.

Koretz, D.M. (2008), *Measuring Up: What Educational Testing Really Tells Us.* Harvard University Press: Cambridge, Massachusetts.

Murnane, R. J., J. B. Willett, and F. Levy. "The Growing Importance of Cognitive Skills in Wage Determination." *Review f Economics and Statistics* 77, no. 2 (1995): 251 - 66.

Murnane, Richard J., and David K. Cohen. "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive." *Harvard Educational Review* 56, no. 1 (1986): 1-17.

Murname, R., J.D. Singer, J.B. Willett, J. Kemple, and R. Olsen (1991), *Who Will Teach? Policies That Matter,* Harvard University Press, Cambridge, MA.

Rivkin, S.G., E.A. Hanushek, and J.F. Kain (2005), "Teachers, schools, and academic achievement." *Econometrica,* Vol. 73, pp. 417-458.

Rothstein, J. 2009 "Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables." NBER Working Paper Series.

The New Teacher Project, 2009. "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness."

Toch, Thomas, and Robert Rothman. 2008. "Rush to Judgment: Teacher Evaluation in Public Education." Washington, Education Sector (January).

Todd, P.E., and K.I. Wolpin. (2003), "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* Vol. 113, pp. F3-F33.

Tucker, Pamela, D. (1997), "Lake Wobegone: Where All Teachers are Competent (Or Have We Come to Terms with the Problem of Incompetent Teachers?)" *Journal of Personnel Evaluation in Education* Vol. 11 pp. 103-126.