

Administration & Supervision Program
Department of Education
University of New Hampshire

Policy Brief
#11-01

**STUDENT OUTCOMES, TEACHER EFFECTIVENESS:
RAISING A CAUTIONARY FLAG**

Todd A. DeMitchell
Professor & Chair, Department of Education
Lamberton Professor, Justice Studies Program
University of New Hampshire

Douglas Gagnon
Doctoral Student, Leadership & Policy Studies
Department of Education
University of New Hampshire

DEPARTMENT OF EDUCATION

at the University of New Hampshire

engaged scholarship

Professional Graduate Study at a
Research University with
Recognized Scholars and Practitioners

www.unh.edu/education

STUDENT OUTCOMES, TEACHER EFFECTIVENESS: RAISING A CAUTIONARY FLAG

Todd A. DeMitchell
Professor & Chair, Department of Education
Lamberton Professor, Justice Studies Program
University of New Hampshire

Douglas Gagnon
Doctoral Student, Leadership & Policy Studies
Department of Education
University of New Hampshire

"It is well established that teacher quality makes a difference in student learning."¹

The issue of linking student achievement scores to teacher effectiveness is on the education agenda at both the national and state levels. For example, the New Hampshire Department of Education through its Task Force on Teacher Effectiveness is reviewing the use of student outcome data in teacher evaluations. Some researcher and policymakers assert that the endpoint of accountability is "holding individual teachers (not just schools) accountable for results."² While it is intuitive and rational that the work of effective teachers should lead to positive student outcomes, the reality of using

¹ Patricia H. Hinchey. (December 2010). *Getting Teacher Assessment Right: What Policymakers Can Learn From Research*. Boulder, CO: National Education Policy Center, p.1. Erick Hanushek. (December 2010). *The Economic Value of Higher Teacher Quality*. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research, Calder The Urban Institute. "First, teachers are very important; no other measured aspect of schools is nearly as important in determining student achievement." P. 3.

assessments designed to measure student knowledge and skills as valid and reliable for the assessment of teacher effectiveness may be counterintuitive and unreasonable. This Brief will explore the issue of using student assessment data as a proxy for teacher effectiveness. Value-added modeling (VAM) is the common name for several different iterations that seek to link, or establish causality between a teacher's performance and student outcomes. It has "become the latest lightning rod in the policy and practice of educational accountability."³ If teacher effects on student learning can be isolated through VAM, then important personnel decisions can be based on the findings of VAM. Retention and compensation are two of the important personnel decisions that could be based, at least in part, on VAM scores. If VAM becomes a widespread practice, a teacher's VAM rating in one school district could be part of the hiring process in another school district.

This Brief is intended to raise a cautionary flag for school leaders and policy makers in the race to implement VAM. We recommend that student achievement data be used as one of the multiple data sets when assessing teacher effectiveness, and only when

² Dan Goldhaber & Michael Hansen. (February 2010). *Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions*. Washington, D.C.: The National Center for Analysis of Longitudinal Data in Education Research, Calder The Urban Institute.

³ Derek Briggs & Ben Domingue. (February 2011). *Due Diligence and the Evaluation of Teachers: A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by the Los Angeles Times*. Boulder, CO: National Education Policy Center, p. 1. Stephen Sawchuk (2011). *Wanted: Ways to Measure Most Teachers*. *Education Week* (February 2), 1, 15 ("The debate about 'value added' measures of teaching is the most divisive topic in teacher-quality policy today.") P. 1.

done so by trained evaluators. However, its use in high stakes personnel decisions must be very carefully considered so as to treat teachers fairly.⁴ Given the limitations of VAM, we strongly recommend that caution be used on the weight and extent of student achievement data. We are guided by a conclusion of the Economic Policy Institute convocation of scholars on VAM: “While there are good reasons for concern about the current system of teacher evaluation, there are also good reasons to be concerned about claims that measuring teacher effectiveness largely by student test scores will lead to improved student achievement.”⁵ We need to improve our evaluation system in order to better assist teachers with the complex job of teaching children.

After we lay our foundation we will offer a concise review of value-added modeling and then discuss what we consider the major cautions about the rush to use student outcomes as measures of teacher effectiveness.

⁴ “The bigger issues with value-added estimates of teacher effectiveness concern their use in personnel compensation, employment, promotion, or assignment decisions. . . . Despite the strength of the research findings, concerns about accuracy, fairness, and potential adverse effects of incentives based on limited outcomes raise worries about using value-added estimates in education staffing and policy.” Eric A. Hanushek & Steven G. Rivkin. (May 2010). *Using Value-Added Measures of Teacher Quality*. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research, Calder The Urban Institute p. 4.

⁵ Eva L. Baker, et al. (August 29, 2010). *Problems with the Use of Student Test Scores to Evaluate Teachers*. Washington, D.C.: Economic Policy Institute, p. 1.

FOUNDATION for TEACHER EVALUATION

First, our beginning point is the acknowledgement of the critical role that teachers play in student achievement. Succinctly stated, “Teacher quality matters.”⁶ Classroom educator’s decisions are directed at assisting and guiding their students to meet and exceed the learning outcomes, educational, social, and personal, established by the School Board. The teacher stands at the crossroads of a student’s education. Therefore, excellence in schools is most directly related to the performance of individuals acting in concert and individually. Second, while teacher effectiveness varies, current systems of evaluation do not sufficiently differentiate among teachers.⁷ Third, the evaluation of teachers is an important component in the delivery of a quality education to students. Consequently, teachers have the right to have accurate and fair feedback on their efforts.⁸

The purposes of the supervision/evaluation process include:

- Develop, improve, and maintain teaching skills and behaviors that result in student achievement, and
- Provide a means for the identification and resolution of problems in work performance, up to and including non-retention or dismissal.

⁶ Jesse Rothstein. (May 2008). *Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement*. Princeton University and NBER
<http://www.irs.princeton.edu/pubs/pdfs/25ers.pdf> (last visited April 24, 2011).

⁷ See Steven Glazerman, et al. (November 17, 2010). *Evaluating Teachers: The Important Role of Value-Added*. Center for Education Policy Analysis: Stanford University; Anthony T. Malinowski, Herbert G. Heneman III, & Steven M. Kimball. (August 2009). *Review of Teaching Performance Assessments for Use in Human Capital Management. Working Paper*. Madison, WI: Strategic Management for Human Capital. Center for Policy Research in Education, University of Wisconsin.

⁸ See, *infra*, *Personnel Evaluation Standards*

VALUE-ADDED MODELING

Value-added modeling is an inclusive term for statistical models that calculate the value a teacher adds to the education of a student. It purports to separate out the numerous factors that impact a student's achievement as measured by a standardized test. It uses multiple years of student test scores to determine how much a teacher is contributing to the growth, or lack of growth, of her/his pupils. There are several different VAM models. The original VAM was developed by William Sanders at the University of Tennessee. His model does not control for the variables that research has found impact student learning, such as socio-economic status, English language learner, minority status to name a few. "Because the value-added method measures gain from a student's starting point, it implicitly controls for socio-economic status and other background factors to the extent that their influence is already reflected in the pre-test scores."⁹ In other words, the longer the student takes the test the less the influence these variables affect the student's learning. There are other models, which select some variables for a regression analysis.

Value-added modeling is used to differentiate the effectiveness of teachers. Teacher VAM scores, the difference between the expected student outcomes and their actual outcomes, are hierarchically ordered into quintiles or deciles, thus comparing teachers by student outcomes without relation to normative data. This can be done at the school and the school district level. VAM data will, therefore, identify the most effective from the least effective and identify those in the middle; it distributes the effectiveness of

⁹ Dale Ballou, William Sanders, & Paul Wright. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics*, 29, 37-65.

teachers from least effective to most effective by assigning an effectiveness score. There is an implicit assumption that if VAM measures the effectiveness of a teacher, especially when used in comparison with other teachers, the impact of the effectiveness would persist. In other words, teacher effects are a “fixed construct that is independent of the context of teaching [for example, class composition, class size, English language learners, etc.] and stable across time.”¹⁰

We will next survey some of the cautions about using VAM for high stakes personnel decisions.

THE CAUTIONARY FLAGS

Does VAM appropriately Control for Student Variables?

Students bring their lives into the classroom; they cannot check the influences in their lives at the schoolhouse gate. Student learning is influenced by home, peer, neighborhood factors, previous learning, attendance, and the fixed school factors (e.g., how are students assigned, what is the class size, what is the makeup of the classroom, what remedial programs are available, what gifted programs are available, how good is the principal?); influences that the classroom teacher takes as a given when the student enters the classroom. “Student characteristics such as poverty, non-English language status, and minority status are

¹⁰ Xiaoxia A. Newton, et al. (September 30, 2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Education Policy Analysis Archives*, 18 923) <http://epaa.asu/ojs/article/view/810> (last visited April 25, 2011) p. 18.

negatively correlated with student outcomes, and usually significantly so.”¹¹ Student motivation at any point in time is another factor that all educators know influences student achievement.

Newton et al. designed a study using student scores for math and English Language Arts on the California Standards Test to investigate whether teacher rankings were consistent across different VAM models. The alternative models controlled for student characteristics (see discussion above on the impact of student variables on VAM scores). The data found significant negative correlations ($*p < .10$ $**p < .05$ $***p < .01$) for English language learners, free lunch recipients, or Hispanic students. Positive correlations were found with proportions of students the teacher had in class who were Asian or whose parents were more highly educated.¹² “This suggests either that teachers who were teaching greater proportions of more advantaged students may have been advantaged in their effectiveness rankings, or that more effective teachers were generally teaching more advantaged students.”¹³

In controlling for the non-random assignment of students into classrooms, one cannot be certain that the ultimate variable of interest—teacher quality—is not in fact being controlled away with it. However, failing to control for some of the

¹¹ Linda Darling-Hammond. (December 1999). *Teacher Quality and Student Achievement: A Review of State Policy Evidence*. Seattle, WA: Center for the Study of Teaching and Policy, University of Washington, p. 29.

¹² Newton, *supra* note 10, p. 11. In addition, they found that English teachers who had more girls in class were more highly ranked (p. 17).

¹³ *Ibid.*

aforementioned variables would systematically benefit the ratings of some teachers. This remains an inescapable reality of such models.

Another perplexing issue in the use of VAM is which students should be counted as part of the analysis, thus subject to the effectiveness of the teacher. How much time a student spends with a teacher in order to be part of the analysis appears to be a policy question with real consequences for teachers. Teachers know the challenge of teaching a mobile student who has moved around or even a student who is new to the classroom when the testing takes place. Who owns the student's learning? Should the student be in the teacher's classroom for 50 percent of the year, 75 percent of the year in order to be counted? In the Ballou, Sanders, and Wright study¹⁴, the policy decision of only using students with 75 percent attendance is shrugged off by the authors as out of their control, but it reveals an unresolved difficulty which is what to do about mobile students (who are typically lower SES, who may have different learning projections and therefore differ systemically from other students). In the sample studied, the percentage of 7th and 8th grade students claimed in estimates of reading is only 56.7 percent and 56.9 percent, respectively. Averaging all subject-grade percentages yields 76.6 percent of students claimed. Although the VAM uses all available data to estimate within-student variation and district means, almost a quarter of students do not enter a teacher's estimated effectiveness. Besides creating bias, this could also have the perverse effect of

¹⁴ Ballou, Sanders, & Wright, *supra* note 9.

teachers caring less about students that do not count towards their rating is VAM is used in decisions of tenure and pay.

Should the Use of Different VAM Models using the same Data Yield the Same Results?

Briggs and Domingue replicated Buddin’s VAM analysis of elementary school student scores for math and reading on the California Standardized Test from 2003 to 2009.¹⁵ Their replication was directed at evaluating the effectiveness ratings of the teachers found by Buddin and later published in the *Los Angeles Times*. Briggs and Domingue added some variables¹⁶ and focused on grade 5 from 2005-2009. Their key question was whether they find a significant shift in the classification of teachers as “effective” or “ineffective” when applying the alternative VAM model. They found that less than half (46.4%) of the teachers’ scores remained in the same quintile for reading, and 61 percent remained in the same quintile in math. When the researchers focused on those teachers in the “more” and “most effective quintiles, 8.1 percent shifted to the “less” or “least” effective quintiles.¹⁷ The shift from ineffective to effective was larger, 12.6 percent. For math the shifts were much smaller. Briggs and Domingue assert, “Given these results it would be hard to argue that teachers or stakeholders evaluating

¹⁵ Briggs & Domingue, *supra* note 3.

¹⁶ “We focus on these particular variables because they have been widely discussed as plausible confounders in the research literature.” *Ibid.*, p. 12.

¹⁷ *Ibid.*, p. 14.

them would be indifferent to the choice of model used to produce these ratings . . .”¹⁸

Consequently, the choice of VAM may have differing effects for teachers.

If a Teacher is Truly Effective Shouldn't Her or His Effectiveness be Stable?

When a new program or instructional strategy is implemented we evaluate the effectiveness of the new treatment to see if we will adopt it. While a short term gain may be encouraging, if the gain does not persist over time we question the efficacy of the treatment. Similarly, if VAM provides a statistical statement of effectiveness at a point in time for a teacher but that effectiveness is not stable over time can we rely on the use of the statistical treatment?

For example, Corcoran found in analysis of Houston and New York City VAM programs that teachers who were placed in one quintile one year moved quintiles including to moving from the lowest to the highest in a single year. This raises the serious issue of stability of scores. In Houston, only 36 percent of the teachers in the lowest quintile one year remained in the lowest quintile the following year. This is amazing growth. Similarly, only 38 percent remained in the top quintile one year later. This is an amazing loss of effectiveness. To further confound the use of VAM to assess effectiveness, 23 percent who were in the lowest quintile one year then moved to the highest quintile the following year; a truly amazing turnaround. These numbers were similar for New York City.¹⁹ There was no indication that the teachers received intensive

¹⁸ *Ibid.*

¹⁹ Sean P. Corcoran. (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 at Brown

professional development during the year in which the data were gathered. This might suggest that some of this change is likely due to the uncertainty in the models rather than growth in the effectiveness of particular teachers.

Sass (2008), as reported by Newton and colleagues, found in a study of five urban school districts “considerable instability in VAM rankings.”²⁰ For example, among those teachers ranked in the lowest quintile of effectiveness one year, only 25 percent to 35 percent were similarly placed the following year. Similarly, a review of those in the highest quintile one year only 20 percent were in the same quintile while a comparable proportion dropped to the bottom two quintiles. What happened? Can we ascribe the movement to changed effectiveness of the teacher? Is it more likely that the factors that brought about the change were beyond the control of the teacher’s instructional and curriculum decisions?

The use of a single year’s VAM rankings for personnel decisions is highly problematic. Some researchers recommend the use of multi-year VAM data. Multi-year data “improves [the] statistical power in estimating teacher effectiveness.”²¹ The use of multiple assessments also reduces the standard error of measurement associated with the use of statistical data on teacher effectiveness. (See below for a discussion of the standard error of measurement.)

University, p. 6. *See, also* Newton, et al, *supra* note 9. For example, for those teachers ranked in the lowest quintile of effectiveness of effectiveness, only 25% to 35% were similarly ranked a year later. A similar finding surfaced for those in the upper quintile only 20% to 30% were ranked there a year later.

²⁰ Newton, et al., *supra* note 10, p. 5.

²¹ See Goldhaber & Hansen, *supra* note 2, p. 18.

Is There Enough Data to Assess Beginning Teachers?

We know that teachers need time on teaching to gain the necessary knowledge and skills. Most VAM models need at least three years to build enough data to have any meaning. When the issue of the standard error of measurement is factored in the cause for concern for beginning teachers is magnified because there is greater error with fewer years of data. This seriously limits its usefulness in making tenure decisions.

Are VAM Scores Precise Enough to Make High Stakes Decision?

All measures have what is called the standard error of measurement. This means that the score may fluctuate; it is not precise. Some scores have larger standard errors of measurement (SEM) than others. The NECAP scores have a standard error of measurement. This impacts the inferences that are drawn from the scores. A large SEM means that the reported score may have a large range of where the actual score may be located. A small SEM means that the actual score is closer to the reported score. Corcoran reports that VAM scores for math over multiple years from New York had a SEM of 34 percent. In other words the true score, for example, for teacher's VAM could be anywhere from the 46th percentile to the 80th percentile. However, related to the problem of a beginning teacher, a one-year SEM was 61 percent (e.g., from the 30th percentile to the 91st percentile).²²

The SEM does decrease with the number of years of data. This helps the validity of the scores for more experienced teachers but does little for beginning teachers. Hanushek and Rivkin assert that “measurement of error issues largely go away if teachers

²² *Ibid.*

are observed over multiple years and with large numbers of children.”²³ The need for multiple years of data and large number of students underscores the caution we raise above regarding beginning teachers. They have few years of experience and few students, therefore their error of measurement is high. These teachers are precisely the ones who need the greatest supervision and the value of VAM to their supervision is highly questionable.

Can VAM be Applied Fairly to Teachers Who do Not Teach Subjects that Are Tested?

This issue is one of fundamental fairness. If student test scores are used to evaluate the effectiveness of a teacher and the teacher does not teach a subject that has a standardized student test score how is that teacher evaluated. For a number of States and researchers the answer is to attribute a certain percentage of the overall student scores for the school to all teachers. The argument is that all teachers impact the school’s learning environment.

We believe that all educators influence the culture of the school. However, the assignment of a percentage of student learning in reading and math and possibly other subjects strikes us as highly arbitrary. While all teachers have influence, the calculus of the level of influence is not currently accessible to statistical analysis. The influence likely changes and is thus liable for the issue of stability discussed above.

Fundamentally, is it fair to hold a person responsible for an outcome in which the individual’s influence over the outcome is unknown? What will the principal tell the

²³ Hanushek & Rivkin, *supra* note 4, p. 4.

teacher whose subject is not tested how to improve her or his contribution to the specific math or reading score of the students?

Can a teacher's contribution to a student's reading score, for example, be accurately assessed when the student has not received any instruction from the teacher?

An assessment of teacher effectiveness with an unknown, unspecified association of impact is more of a drive-by assessment attached to high-stakes personnel and personal consequences than a valid and reliable use of data. It is an assessment "too far."²⁴

Business Uses Data to Gauge Effectiveness: Why Can't Education?

A common argument is that business uses statistics why can't statistics on student achievement be used to gauge the effectiveness of teachers. Examples include batting average as the measure for baseball player effectiveness.²⁵ Several problems are associated with this analogy. First, the baseball player may be traded or dropped for reasons other than his/her batting average (a disruptive influence in the dugout, failing to hit for power, poor fielding skills in the National League with no designated hitter). Seldom is effectiveness reduced to a single statistical number. Multiple measures such as runs, stolen bases, and runs batted in are needed to get a truer picture of the player. Second, the batting average is by in large a product of the batter. The batter must not rely

²⁴ Cornelius Ryan wrote a classic World War II history, *A Bridge Too Far* (1974) of a failed attempt to attempt to break through the German Lines at Arnhem, Netherlands in 1944. The title has come to mean overreaching.

²⁵ Jim Hull. (n.d.) *Building a Better Evaluation System*. Alexandria, VA: Center for Public Education, National School Boards Association
<http://www.centerforpubliceducation.org/Main-Menu/Staffingstudents/Building-A-Better-Evaluation-System/Building-A-Better-Evaluation-System.html> (last visited April 23, 2011).

upon the actions of another motivated person in order to produce hits. Student achievement is reliant on a host of individuals—administrators, aides, other teachers, and the students themselves—beside the classroom teacher. Third, the batter is not dependent upon the hitting ability of the prior batter, as a teacher may be dependent upon the prior learning (the previous teachers) that a student brings to the class. Learning gains (or lack of gains) produced by a teacher persists for many years after a student leaves the classroom. Fourth, some batters can only hit left-handed pitchers so they are only used in certain situations. This option is usually not available to teachers, who must teach the students they are assigned. Pinch hitting is not a part of teaching.

Some of the Consequences of the Use of VAM

There are consequences to using VAM, some are unintended. The following are a few of the consequences we have identified. There may be more that thoughtful discussion may reveal.

- VAM may indicate problem but it cannot identify the problem or point to a preferred solution. It cannot be used for formative assessments. VAM cannot identify deficits, what needs to be improved, and what steps need to be taken to improve. The decision making based on VAM is dichotomous – retain or dismiss.²⁶ What and how to

²⁶ For example, Goldhaber & Hansen, *supra* note 2, conclude that the “deselection”, the dismissal of teachers, to the bottom quarter of teachers is a more cost-effective alternative to increasing student achievement than reducing class-size. P. 31. In other words, fire the 25 percent lowest performing teachers on a VAM. The way VAM is constructed there will always be a lowest 25 percent of teachers. Is firing 25 percent of the teachers every year the preferred solution?

improve is not available under VAM. Only a highly trained educator can take these important steps working one on one with the teacher.

- Will VAM distort the educational process? Will teachers who are being publically designated as ineffective with the result of less compensation or dismissal alter their behavior in response to increasing their score as opposed to building their teaching skills?²⁷ Will the curriculum shrink further to that which is tested? Will our notion of effective teaching, which is far broader than the ability to raise standardized test scores, be changed for the worse?

- “If teacher evaluations are sensitive to student assignments, teachers may attempt to manage their assignments so as to maximize their personal estimated value added, perhaps by lobbying their principals for preferred students”²⁸ Some students may be easier to help, while the achievement of some students is not even taken into account in calculating teacher ratings (due to attendance requirements, for example). In one model, only about three-quarters of students are used to calculate a teacher’s effectiveness score.²⁹ This could create a perverse situation, whereby teachers feel

²⁷ Brian Stecher of RAND asks how teacher evaluation systems such as VAM affect teacher behavior. “Are teachers still making decisions in the best interests of good instruction, or are they trying to game the system to do what they think will artificially raise their value-added scores? A lot of these things can’t be known until after the fact.” Stephen Sawchuck. (2011). Building Systems for Evaluation of Teachers Poses Challenges. *Education Week* (April 27) 1, 18. P. 18.

²⁸ Rothstein, *supra* note 6, p. 5.

²⁹ Ballou, Sanders, & Wright, *supra* note 9.

pressured to give attention to some students more than others. Who can raise my VAM score?

- Are teachers of the neediest students disadvantaged by VAM? Given the findings of researcher like Newton et al. – “a substantial portion of the variation in teacher rankings is attributable to selected characteristics”³⁰ – will teachers in systems that use a VAM that does not take into account student characteristics “create disincentives for teachers to want to work with those students with the greatest needs.”³¹ Even more pernicious is the inevitability that outright cheating would occur.

- Will the use of VAM hurt morale and collaboration among teachers? Given the relative nature of individual VAM ratings, a loss in one teacher’s effectiveness is another’s gain. For teachers who do not teach subjects that are assessed will a dip in student scores be an opportunity for collaboration or recrimination because of the high-stakes attached to the dip. Furthermore, VAM ratings may systematically favor certain grades or subjects. How will VAM effect staffing issues?

CONCLUSION

The need to gauge teacher effectiveness is as great as ever. With more sophisticated statistical tools comes the hope that a more objective measure of effective teaching can be implemented. VAM has great intuitive appeal, as it purports to separate out the impact that a particular teacher has on student achievement. In low-stakes

³⁰ Newton, *supra* note 10, p. 18.

³¹ *Ibid.*

contexts such as program evaluation, VAM currently offers great promise. However, in using VAM to inform high-stakes decisions such as teacher hiring and tenure, much more caution must be used. As a recent National Research Council report urges³³, VAM should not be the primary indicator used in teacher evaluation systems.

Many measurement challenges persist in the use of VAM to assess teachers. Model choice, instability across years, and a lack of data for beginning teachers remain thorny issues to be resolved before VAM should be used with conviction. Even if these issues are ameliorated, broader questions persist. Is the construct of effective teaching captured by student growth on standardized tests? Isn't the effectiveness of a teacher also determined by their ability to inspire and motivate, serve as a role model, and promote higher orders of thinking? It seems unlikely that standardized testing will ever capture these more abstract abilities. Moreover, what negative unintended consequences may come from instituting VAM in teacher evaluation systems? There exists many ways that VAM could cause more harm than good. Finally, how may VAM results be used in a formative manner? Without addressing how ineffective teachers can improve, VAM is only an indicator. Comprehensive evaluation systems must also address teacher development.

Is VAM a simple solution for measuring a complex activity: how to direct, organize, teach, and inspire someone else to learn? But is it the right solution or just the easy solution? There is very little heavy lifting necessary if we just categorize teachers

³³ Chudowsky, N., Koenig, J. A., Braun, H. I., National Research Council (U.S.), & National Academy of Education. (2010). *Getting value out of value-added: Report of a workshop*. Washington: National Academies Press.

into effective and ineffective and dismiss the ineffective, low score; you're fired in the words of Donald Trump. Or, should test scores of students be an important data point for a more searching professional dialogue about effectiveness and a possible trigger for gathering more data with in-depth analysis. This path may lead to dismissal or it may lead to improvement.

In the interest of students, those who devise and enact education policy must find better ways to promote good teaching. This will entail a reorganization of teacher evaluation systems, which will include a component of student achievement results. Tantamount to enacting new evaluation systems is that administrators are aware of the limitations and potential consequences of their implementation. This paper aims to illuminate these concerns when using VAM to evaluate teachers.

Personnel Evaluation Standards

Summary of the Standards

Propriety Standards

The Propriety Standards are intended to ensure that a personnel evaluation will be conducted legally, ethically, and with due regard for the welfare of the evaluatee and those involved in the evaluation.

- **P1 Service Orientation** Personnel evaluations should promote sound education, fulfillment of institutional missions, and effective performance of job responsibilities, so that the educational needs of students, community, and society are met.
- **P2 Appropriate Policies and Procedures** Guidelines for personnel evaluations should be recorded and provided to the evaluatee in policy statements, negotiated agreements, and/or personnel evaluation manuals, so that evaluations are consistent, equitable, and fair.
- **P3 Access to Evaluation Information** Access to evaluation information should be limited to the persons with established legitimate permission to review and use the information, so that confidentiality is maintained and privacy protected.
- **P4 Interactions with Evaluatees** The evaluator should respect human dignity and act in a professional, considerate, and courteous manner, so that the evaluatee's self-esteem, motivation, professional reputations, performance, and attitude toward personnel evaluation are enhanced or, at least, not needlessly damaged.
- **P5 Balanced Evaluation** Personnel evaluations should provide information that identifies both strengths and weaknesses, so that strengths can be built upon and weaknesses addressed.
- **P6 Conflict of Interest** Existing and potential conflicts of interest should be identified and dealt with openly and honestly, so that they do not compromise the evaluation process and results.
- **P7 Legal Viability** Personnel evaluations should meet the requirements of all federal, state, and local laws, as well as case law, contracts, collective bargaining agreements, affirmative action policies, and local board policies and regulations or institutional statutes or bylaws, so that evaluators can successfully conduct fair, efficient, and responsible personnel evaluations.

Utility Standards

The Utility Standards are intended to guide evaluations so that they will be informative, timely, and influential.

- **U1 Constructive Orientation** Personnel evaluations should be constructive, so that they not only help institutions develop human resources but encourage and assist those evaluated to provide excellent services in accordance with the institution's mission statements and goals.
- **U2 Defined Uses** Both the users and intended uses of a personnel evaluation should be identified at the beginning of the evaluation so that the evaluation can address appropriate questions and issues.
- **U3 Evaluator Qualifications** The evaluation system should be developed, implemented, and managed by persons with the necessary qualifications, skills, training, and authority, so that evaluation reports are properly conducted, respected and used.
- **U4 Explicit Criteria** Evaluators should identify and justify the criteria used to interpret and judge evaluatee performance, so that the basis for interpretation and judgment provide a clear and defensible rationale for results.
- **U5 Functional Reporting** Reports should be clear, timely, accurate, and germane, so that they are of practical value to the evaluatee and other appropriate audiences.
- **U6 Professional Development** Personnel evaluations should inform users and evaluatees of areas in need of professional development, so that all educational personnel can better address the institution's missions and goals, fulfill their roles and responsibilities, and meet the needs of students.

Feasibility Standards

The Feasibility Standards are intended to guide personnel evaluation systems so that they are as easy to implement as possible, efficient in their use of time and resources, adequately funded, and viable from a political standpoint.

- **F1 Practical Procedures** Personnel evaluation procedures should be practical, so that they produce the needed information in efficient, non-disruptive ways.
- **F2 Political Viability** Personnel evaluations should be planned and conducted with the anticipation of questions from evaluatees and others with a legitimate right to know, so that their questions can be addressed and their cooperation obtained.
- **F3 Fiscal Viability** Adequate time and resources should be provided for personnel evaluation activities, so that evaluation can be effectively implemented, the results fully communicated, and appropriate follow-up activities identified.

Accuracy Standards

The accuracy standards determine whether an evaluation has produced sound information. Personnel evaluations must be technically adequate and as complete as possible to allow sound judgments and decisions to be made. The evaluation methodology should be appropriate for the purpose of the evaluation and the evaluatees being evaluated and the context in which they work.

- **A1 Validity Orientation** The selection, development, and implementation of personnel evaluations should ensure that the interpretations made about the performance of the evaluatee are valid and not open to misinterpretation.
- **A2 Defined Expectations** The qualifications, role, and performance expectations of the evaluatee should be clearly defined, so that the evaluator can determine the evaluation data and information needed to ensure validity.
- **A3 Analysis of Context** Contextual variables that influence performance should be identified, described, and recorded, so that they can be considered when interpreting an evaluatee's performance.
- **A4 Documented Purposes and Procedures** The evaluation purposes and procedures, both planned and actual, should be documented, so that they can be clearly explained and justified.
- **A5 Defensible Information** The information collected for personnel evaluations should be defensible, so that the information can be reliably and validly interpreted.
- **A6 Reliable Information** Personnel evaluation procedures should be chosen or developed and implemented to assure reliability, so that the information obtained will provide consistent indications of the evaluatee's performance.
- **A7 Systematic Data Control** The information collected, processed, and reported about evaluatees should be systematically reviewed, corrected as appropriate, and kept secure, so that accurate judgments about the evaluatee's performance can be made and appropriate levels of confidentiality maintained.
- **A8 Bias Identification and Management** Personnel evaluations should be free of bias, so that interpretations of the evaluatee's qualifications or performance are valid.
- **A9 Analysis of Information** The information collected for personnel evaluations should be systematically and accurately analyzed, so that the purposes of the evaluation are effectively achieved.
- **A10 Justified Conclusions** The evaluative conclusions about the evaluatee's performance should be explicitly justified, so that evaluatees and others with a legitimate right to know can have confidence in them.
- **A11 Metaevaluation** Personnel evaluation systems should be examined periodically using these and other appropriate standards, so that mistakes are prevented or detected and promptly corrected, and sound personnel evaluation practices are developed and maintained over time.