AERA
OFFICIAL DOCUMENTS

AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs

Approved by AERA Council, June 2015

Summary

The purpose of this statement is to inform those using or considering the use of value-added models (VAM) about their scientific and technical limitations in the evaluation of educators and programs that prepare teachers. The statement briefly reviews the background and current context of using VAM for evaluations, enumerates specific psychometric problems with VAM, and addresses the validity of inferences from VAM, given the challenges of isolating the contributions of teachers and school leaders from the many other factors that shape student learning. The statement also addresses the limitations of using VAM to evaluate educator preparation programs, given the wide variation of experiences and settings in which graduates from those programs work and the lack of comparable and complete information on programs. In addition, the statement goes beyond a consideration of challenges and limitations by specifying eight technical requirements that must be met for the use of VAM to be accurate, reliable, and valid. The statement concludes by stressing the importance of any educator evaluation system meeting the highest standards of practice in statistics and measurement. It calls for substantial investment in research on VAM and alternative methods and models, and cautions against VAM being used to have a high-stakes, dispositive weight in evaluations.

Introduction

The purpose of this statement is to inform those using or considering the use of VAM about the scientific and technical limitations of its inclusion in the implementation of evaluation systems.

The use of VAM to evaluate educators and educator preparation programs continues to be the subject of discussion and debate. There is a shared interest in policy and practice communities in implementing educator evaluation systems that can lead to improvements in instructional practices and that are fair and free of bias. Nevertheless, there is considerable disagreement among education policy makers and decision makers about whether the state of knowledge about VAM, alone or in combination with other indicators, is sufficiently well developed to be incorporated into accountability systems.

The scientific issues at stake, as well as the disagreements surrounding VAM, are documented in an extensive literature[1] and, most recently, in a 2015 special issue of the *Educational Researcher* titled "Value Added Meets the Schools: The Effects of Using Test-Based Teacher Evaluation on the Work of Teachers and Leaders."[2] This statement does not review that literature. Rather, it draws on testing, statistical, and methodological expertise in the field of education research and related sciences and on the standards that guide research and its rigorous applications in policy and practice.[3]

Background and Issues

There is broad consensus about the need for high-quality teachers and principals for all students, especially underserved learners. In an effort to increase teacher and principal quality, many states are devising educator evaluation systems that employ, to varying degrees, statistical indicators related to changes in their students' test-based performance. Some jurisdictions are also extending the use of these systems[4] to evaluate educator preparation programs. Research evidence on the accuracy, reliability, and stability of such indicators, the validity of the underlying measures, and the consequences of the use of such indicators in educator evaluation systems is still accumulating. Thus, the technical foundations for their use in evaluation systems are far from settled.

For purposes of this statement, the phrase *value-added models* is used as an umbrella term to refer to a variety of "true" value-added models, student growth percentiles, and certain growth models that are used for evaluation.[5] In the newly devised educator evaluation systems referenced above, VAM are employed in an attempt to determine teachers' and leaders' contributions to student learning outcomes, as captured by standardized tests,

The *AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs* was prepared under the auspices of and approved by the AERA Council. A subcommittee of Council prepared initial drafts that received blind review from ten experts. Henry Braun, Boston College, served as the independent monitor and presided over the review process. Based on the reviews, further revisions, and subsequent discussion at Council, this statement received final review from Dr. Braun and approval by Council in June 2015.

Educational Researcher, Vol. XX No. X, pp. 1–5
DOI: 10.3102/0013189X15618385

and are typically employed to identify educators who appear, by these measures, to have been particularly effective or ineffective. In teacher evaluation, VAM scores are derived from the aggregate of test-score changes of students in their classrooms. In principal evaluation, there are further aggregations of those changes across multiple grades and classrooms. In program evaluation, scores are also being used retrospectively to draw inferences about the preparation programs in which educators have been trained.

VAM are generally viewed as superior to status models for gauging impacts on student learning outcomes because they are based in some way on changes in test-based performance. Status models simply reflect the proportion of students meeting or exceeding a performance threshold at the end of the school year, without regard to their academic standing at the start of the year. Under a status model, a teacher with a higher-scoring entering class typically will be advantaged in comparison to a teacher with a lower-scoring entering class. In contrast, VAM focus on test-based changes so that teachers or leaders with higher-scoring entering student cohorts are not necessarily advantaged.

Although VAM may be superior to status models, it does not mean that they are ready for use in educator or program evaluation. There are potentially serious negative consequences in the context of evaluation that can result from the use of VAM based on incomplete or flawed data, as well as from the misinterpretation or misuse of the VAM results. Teachers and leaders, for example, with low VAM scores can experience loss of advancement, lost compensation, and even termination. Also, when large numbers of teachers and leaders are misidentified, then resources may be misdirected, and the educational system as a whole can be degraded. Only if such indicators are based on high-quality, audited test data and supported by sound validation evidence for the specific purposes proposed, can they be appropriately used, along with other relevant indicators, for professional development purposes or for educator evaluation.

Limitations of the Use of VAM for Evaluation

There are fundamental issues that need to be addressed to use VAM to evaluate teachers and other educators. Currently, longitudinal records of students' scores on standardized tests serve as the input to VAM. Standardized tests, however, vary in the degree to which they fully capture the target constructs, as well as in their levels of precision across the range of reported scores. In addition, current state tests, by federal requirement, measure only grade-level standards without including items needed to measure growth for students who perform well below or well above grade level. Therefore, caution about the psychometric quality of the assessment should be exercised if VAM are being considered for purposes of teacher evaluations.

Moreover, the use of VAM presents additional substantial challenges in the evaluation of principals and nonteaching staff. Existing VAM estimates have not been shown to isolate sufficiently the effectiveness of teachers, principals, or other nonteaching professional staff. Extant research, for example, suggests that principal effectiveness can only be separated from school effectiveness when applied to relatively new principals or in evaluations of school improvement over multiple years of a principal's tenure in a school.[6]

The limitations of using VAM to isolate the relative effectiveness of teachers and leaders are further compounded when used to compare the effectiveness of educator preparation programs. There is very little evidence that value-added models can be used to evaluate the effectiveness of educator preparation programs based on the aggregation of graduates' performance as teachers or leaders.[7] At first blush, it might seem commonsensical that VAM scores of novice teachers or leaders aggregated back to their preparation programs could serve as a basis for comparison. However, such use presents further challenges since those teachers and leaders are working in a wide range of schools, grades, and districts. Important differences in those settings, including variations in student populations, curricula, class sizes, and resources, as well as in the quality of induction and mentoring, contribute to differences in educators' performances and, therefore, are confounded with differences in the efficacy of their training programs. The difficulties are only compounded if the teachers and leaders included in such evaluations are not representative of all graduates across programs, as is the case, for example, when programs are small or graduates do not work in public school systems.[8] Finally, it is logistically and methodologically problematic to take account of differences among programs in both the prior preparation of matriculants and the type of experiences that they have had since program completion.

Technical Requirements for the Use of VAM

Those engaged in or contemplating the use of VAM in a system of evaluation must weigh the potential benefits against the limitations and complexities described above. Moreover, they must consider whether the consequences of such use will likely lead to improvements in instructional practices and meaningful gains in student learning.

Because of the adverse consequences of faulty evaluations for educators and the students they serve, use of VAM in any evaluation system must meet a very high technical bar. This section sets forth the technical requirements, all of which must be met, for VAM uses to be scientifically rigorous and fair.[9] Any material departure from these requirements should preclude use.

Even if all of the technical requirements listed below are met, the validity of inferences from VAM scores depends on the ability to isolate the contributions of teachers and leaders to student learning from the contributions of other factors not under their control. This is very difficult, not only because of data limitations but also because of the highly nonrandom sorting of students and teachers into schools and classes within schools. Consequently, such disentangling can be accomplished only imperfectly and with an unknown degree of success. The resulting bias will not be distributed evenly among schools, given wide variation in critical factors like student mobility, and could in itself make some students, schools, and teachers appear to be underperforming. This residual bias in the VAM scores can be exacerbated by measurement error in the predictors employed in the model. Therefore, due caution should be exercised in the

interpretations of VAM scores, since we generally do not know how to properly adjust for the impact of these other factors.

(1) *VAM scores must only be derived from students' scores on assessments that meet professional standards of reliability and validity for the purpose to be served.*

For assessment scores to be used in VAM for any purpose, it is essential that the assessments meet professional standards for assessments as described in the *Standards for Educational and Psychological Testing* issued in 2014 by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). Relevant evidence should be reported in the documentation supporting the claims and proposed uses of VAM results, including evidence that the tests used are a valid measure of growth by measuring the actual subject matter being taught and the full range of student achievement represented in teachers' classrooms.

(2) *VAM scores must be accompanied by separate lines of evidence of reliability and validity that support each claim and interpretative argument.*

Each type of evaluation (teachers, leaders, preparation programs) requires evidence to support the validity argument for that particular application. That evidence must take into account the potential impact of contextual factors and selection bias on the appropriateness of the inferences made.[10] The validity arguments used to support the use of students' value-added scores for program evaluation (which are less error-prone) are insufficient to support the aggregation and use of such scores for individual personnel evaluation.

(3) *VAM scores must be based on multiple years of data from sufficient numbers of students.*

The precision of VAM scores depends on the amount and quality of the data available, as well as on the features of the model. Therefore, VAM scores should not be used unless they are derived from data obtained from sufficient numbers of students over multiple years. VAM scores should always be accompanied by estimates of uncertainty to guard against overinterpretation of differences. Further, care must be taken to address estimate instability that results from teacher mobility across schools, grades, and subjects.

(4) *VAM scores must only be calculated from scores on tests that are comparable over time.*

Many states are currently transitioning to new assessment systems and adopting new or revised performance standards. Major transitions typically affect student performance both directly and indirectly, as teachers and leaders adapt to the new standards, assessments, and expectations. Although such changes are to be expected, they pose a threat to the validity of the interpretations of VAM scores, especially when these scores are compared before, across, and after the transition. Transitions in student assessment not only pose difficulties for VAM, but also interrupt longitudinal trends of status measures. Such situations can lead to misinterpretations of student progress. In these instances, assessments across years may no longer be equated and the statistical links between scores are not sufficiently strong to support the validity arguments and interpretations required for VAM. Although consistent categories can be established across assessments for some models, the interpretations of growth from before or after the transition to the current assessment may not be comparable. Consequently, VAM scores should generally not be employed across transitions.

(5) *VAM scores must not be calculated in grades or for subjects where there are not standardized assessments that are accompanied by evidence of their reliability and validity.*

When standardized assessment data are not available across all grades (K–12) and subjects (e.g., health, social studies) in a state or district, alternative measures (e.g., locally developed assessments, proxy measures, observational ratings) are often employed in those grades and subjects to implement VAM.[11] Such alternative assessments should not be used unless they are accompanied by evidence of reliability and validity as required by the AERA, APA, and NCME *Standards for Educational and Psychological Testing*. Because the validity of VAM scores is so dependent on the quality of the underlying assessment, they should not be implemented in grades or subjects where there is a lack of evidence of reliability and validity.

(6) *VAM scores must never be used alone or in isolation in educator or program evaluation systems.*

If VAM scores are used, they should only be one component in a more comprehensive educator or program evaluation. Also, their meaning should be interpreted in the context of an individual teacher's curriculum and teaching assignments, with cautions issued regarding common interpretation problems, such as ceiling and floor effects of the tests for estimating growth for high- and low-achieving students. Other measures of practice and student outcomes should always be integrated into judgments about overall teacher effectiveness.

(7) *Evaluation systems using VAM must include ongoing monitoring for technical quality and validity of use.*

Ongoing monitoring is essential to any educator evaluation program and especially important for those incorporating indicators based on VAM that have only recently been employed widely. If authorizing bodies mandate the use of VAM, they, together with the organizations that implement and report results, are responsible for conducting the ongoing evaluation of both intended and unintended consequences. The monitoring should be of sufficient scope and extent to provide evidence to document the technical quality of the VAM application and the validity of its use within a given evaluation system. When there is credible evidence that there are negative consequences, every effort should be made to mitigate them. Although when multiple indicators are used, it may be difficult to determine which technical challenges

are attributable to VAM, the presence of negative consequences alone should trigger a "red flag."

(8) *Evaluation reports and determinations based on VAM must include statistical estimates of error associated with student growth measures and any ratings or measures derived from them.*

There should be transparency with respect to VAM uses and the overall evaluation systems in which they are embedded. Reporting should include the rationale and methods used to estimate error and the precision associated with different VAM scores. Also, their reliability from year to year and course to course should be reported. Additionally, when cut scores or performance levels are established for the purpose of evaluative decisions, the methods used, as well as estimates of classification accuracy, should be documented and reported. Justification should be provided for the inclusion of each indicator and the weight accorded to it in the evaluation process.

Elements of the report should include: (a) a description of the data and the data quality checks employed; (b) the methodology, statistical models, and computational methods employed; (c) a rationale and explanation of how each indicator has been incorporated into the evaluation system; and (d) validity evidence to support the use of the system. When reporting identifies material problems in the use of VAM, procedures should be established that trigger a review of the evaluation system and possible system modifications necessary for continued use. Reporting can be accomplished through the preparation of a technical manual, an implementation manual, or a set of research reports. Dissemination should include accessible formats that are widely available to the public, as well as to professionals.

Conclusion

Many states and districts have incorporated VAM in a comprehensive system to evaluate teachers, principals, and educator preparation programs. There are considerable risks of misclassification and misinterpretation in the use of VAM to inform these evaluations. As detailed above, the education research community emphasizes that the use of VAM in any evaluations must satisfy technical requirements of accuracy, reliability, and validity. This includes attention not only to the construct validity and reliability of student assessments, but also to the reliability of the results of educator and program evaluation models, as well as their consequential validity. In sum, states and districts should apply relevant research and professional standards that relate to testing, personnel, and program evaluation before embarking on the implementation of VAM.

The standards of practice in statistics and testing set a high technical bar for properly aggregating student assessment results for any purpose, especially those related to drawing inferences about teacher, school leader, or educator preparation program effectiveness. Accordingly, the AERA recommends that VAM (which include student gain score models, transition models, student growth percentile models, and value measures models) not be used without sufficient evidence that this technical bar has been met in ways that support all claims, interpretative arguments, and uses (e.g., rankings, classification decisions). Although there may be differences in views about the desirability of using VAM for evaluation purposes, there is wide agreement that unreliable or poor-quality data, incorrect attributions, lack of reliability or validity evidence associated with value-added scores, and unsupported claims lead to misuses that harm students and educators.

Finally, the AERA recommends substantial investment in research on VAM, as well as on alternative methods and models for educator and program evaluation. There are promising alternatives currently in use in the United States that merit attention.[12] These include use of teacher observation data[13] and peer assistance and review models that provide formative and summative assessments of teaching[14] and honor teachers' due process rights.[15] There is also research that considers the relationship between educator practice and student outcomes, and the relationship between features of preparation programs and their graduates' performance outcomes.[16]

The value of high-quality, research-based evidence cannot be overemphasized. Ultimately, only rigorously supported inferences about the quality and effectiveness of teachers, educational leaders, and preparation programs can contribute to improved student learning.

NOTES

[1]See Lockwood & McCaffrey, 2007; Rothstein, 2009.

[2]See Harris & Herrington, 2015.

[3]This statement on the use of VAM and other similar models for educator and program evaluation builds on and complements professional standards and recommendations. The first is the *Standards for Educational and Psychological Testing* (2014), which outlines sound and appropriate test use in education and psychology. These standards and recommendations are sponsored and endorsed by the American Educational Research Association, the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). Second, the National Research Council and National Academy of Education published a workshop report, *Getting the Value Out of Value-Added* (Braun et al., 2010, which raised several key issues, particularly that (1) the "PBI model" (PBI) has been applied to a range of approaches, varying in their data requirements, statistical complexity, and evaluation use; and (2) there are many concerns over the tests used and the technical aspects (particularly sources of bias and imprecision) and issues of transparency and public understanding (Braun et al., 2010). The third is AERA's *Position Statement on High-Stakes Testing in Pre-K–12 Education* (American Educational Research Association, 2000).

[4]There are a variety of models which aggregate students' growth or value-added indicators (VAM) to provide some measure of change that is incorporated in the evaluation of educators or educator preparation programs. This statement refers to all models as VAM and does not address the distinctions of different models throughout. These models include (a) Gain Score–based Models (e.g., Growth) or Mean Gain, which simply aggregate difference scores derived from subtracting previous scores from current scores on tests; (b) Transition-based Models (or Categorical Models), which compute aggregate changes in performance categories over a period of 2 or more years; (c) Student Growth Percentiles–based models (SGPs), which answer the question "What is the percentile rank of a student's current test score compared to students with similar score histories?" and then evaluate teachers based on the median or mean percentiles aggregated across their students; and (d) Value-Added Measures–based Models (VAM), which establish an expected current test score for students based on test scores from previous years, along with (possibly) other demographic characteristics of the student, classroom, and the school in attempting to account for the impact of factors beyond student achievement to isolate the teacher's

impact. Each of these models has different strengths and drawbacks that need to be considered when interpreting their results, but such discussion is beyond the scope of this statement (see Braun et al., 2010).

[5]See American Statistical Association, 2014.

[6]See Chiang, Lipscomb, & Gill, 2012; Grissom, Kalogrides, & Loeb, 2012.

[7]See Gansle, Noell, & Burns, 2012.

[8]See Henry, Kershaw, Zulli, & Smith, 2012; Knight et al., 2012.

[9]This statement of conditions parallels and is consistent with the American Educational Research Association's *Position Statement on High-Stakes Testing in Pre-K–12 Education*, adopted in 2000.

[10]The development and use of value-added (or growth) results for teachers, leaders, and preparation programs often requires different measures and methods of data aggregation and attention to measurement error. Different VAM may be based on different assumptions, and the degree to which measurement error is accounted for needs to be explicit (Lockwood et al., 2007; Newton, Darling-Hammond, Haertel, & Ewart, 2010; Braun, Chudowsky, & Koenig, 2010).

[11]See Fuller & Hollingworth, 2014; Goe & Holdheide, 2011.

[12]Outside the United States, there are alternative approaches to ensuring high levels of teacher and leader quality that do not use standardized testing (e.g., Finland, Singapore).

[13]See Goldring et al., 2015.

[14]See Goldstein, 2010; Papey & Moore Johnson, 2012.

[15]See Baker, Oluwole, & Green, 2013.

[16]Darling-Hammond, Meyerson, LaPointe, & Orr, 2010; Goldhaber, 2013.

REFERENCES

American Educational Research Association. (2000). *Position statement on high-stakes testing in pre-K–12 education*. Retrieved from http://www.aera.net/AboutAERA/AERARulesPolicies/AERAPolicyStatements/PositionStatementonHighStakesTesting/tabid/11083/Default.aspx

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Retrieved from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Baker, B. D., Oluwole, J., & Green, P. C., III (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archives*, *21*(5).

Braun, H., Chudowsky, N., & Koenig, J. (2010). *Getting the value out of value-added: Report of a workshop*. Washington, DC: National Research Council and National Academy of Education.

Chiang, H., Lipscomb, S., & Gill, B. (2012). *Is school value-added indicative of principal quality*? Cambridge, MA: Mathematica Policy Research.

Darling-Hammond, L., Meyerson, D., LaPointe, M., & Orr, M. (2010). *Preparing principals for a changing world*. San Francisco: Jossey-Bass.

Fuller, E. J., & Hollingworth, L. (2014). A bridge too far? Challenges in evaluating principal effectiveness. *Educational Administration Quarterly*, *50*(3), 466–499.

Gansle, K. A., Noell, G. H., & Burns, J. M. (2012). Do student achievement outcomes differ across teacher preparation programs? An analysis of teacher education in Louisiana. *Journal of Teacher Education*, *63*(5), 304–317.

Goe, L., & Holdheide, L. (2011). *Measuring teachers' contribution to student learning growth for nontested grades and subjects* (Research & Policy Brief). Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D. (2013). *What do value-added measures of teacher preparation programs tell us?* Palo Alto, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/teacher_prep/

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*(2), 96–104.

Goldstein, J. (2010). *Peer review and teacher leadership: Linking professionalism and accountability*. New York: Teachers College Press.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2012). *Using student test scores to measure principal performance*. Nashville, TN: Vanderbilt University.

Harris, H. N., & Herrington, C. D. (Eds.). (2015). Value added meets the schools: The effects of using test-based teacher evaluation on the work of teachers and leaders [Special issue]. *Educational Researcher*, *44*(2).

Henry, G. T., Kershaw, D., Zulli, R., & Smith, A. (2012). Incorporating teacher effectiveness into teacher preparation program evaluation. *Journal of Teacher Education*, *63*(5), 335–355.

Knight, S. L., Edmonson, J., Lloyd, G., Arbaugh, F., Nolan, J., Whitney, E., & McDonald, P. (2012). Examining the complexity of assessment and accountability in teacher education. *Journal of Teacher Education*, *63*(5), 301–303.

Lockwood, J., & McCaffrey, D. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics*, *1*, 223–252. Retrieved from http://www.rand.org/content/dam/rand/pubs/reprints/2007/RAND_RP1266.pdf

Lockwood, J. R., McCaffrey, D., Hamilton, L., Stecher, B., Le, V., & Martinez, J. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives*, *18*(23).

Papey, J. P., & Moore Johnson, S. (2012). Is PAR a good investment? Understanding the costs and benefits of teacher Peer Assistance and Review programs. *Educational Policy*, *26*(5), 696–729.

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