American Statistical Association

Promoting the Practice and Profession of Statistics

ASA Statement on Using Value-Added Models for Educational Assessment

April 8, 2014

Executive Summary

Many states and school districts have adopted Value-Added Models (VAMs) as part of educational accountability systems. The goal of these models, which are also referred to as Value-Added Assessment (VAA) Models, is to estimate effects of individual teachers or schools on student achievement while accounting for differences in student background. VAMs are increasingly promoted or mandated as a component in high-stakes decisions such as determining compensation, evaluating and ranking teachers, hiring or dismissing teachers, awarding tenure, and closing schools.

The American Statistical Association (ASA) makes the following recommendations regarding the use of VAMs:

- The ASA endorses wise use of data, statistical models, and designed experiments for improving the quality of education.
- VAMs are complex statistical models, and high-level statistical expertise is needed to develop the models and interpret their results.
- Estimates from VAMs should always be accompanied by measures of precision and a discussion of the assumptions and possible limitations of the model. These limitations are particularly relevant if VAMs are used for high-stakes purposes.

- VAMs are generally based on standardized test scores, and do not directly measure potential teacher contributions toward other student outcomes.
- VAMs typically measure correlation, not causation: Effects – positive or negative – attributed to a teacher may actually be caused by other factors that are not captured in the model.
- Under some conditions, VAM scores and rankings can change substantially when a different model or test is used, and a thorough analysis should be undertaken to evaluate the sensitivity of estimates to different models.

- VAMs should be viewed within the context of quality improvement, which distinguishes aspects of quality that can be attributed to the system from those that can be attributed to individual teachers, teacher preparation programs, or schools. Most VAM studies find that teachers account for about 1% to 14% of the variability in test scores, and that the majority of opportunities for quality improvement are found in the system-level conditions. Ranking teachers by their VAM scores can have unintended consequences that reduce quality.

As the largest organization in the United States representing statisticians and related professionals, the American Statistical Association (ASA) is making this statement to provide guidance, given current knowledge and experience, as to what can and cannot reasonably be expected from the use of VAMs. This statement focuses on the use of VAMs for assessing teachers' performance but the issues discussed here also apply to their use for school or principal accountability. The statement is not intended to be prescriptive. Rather, it is intended to enhance general understanding of the strengths and limitations of the results generated by VAMs and thereby encourage the informed use of these results.

Value-Added Models and Their Interpretation

In recent years test-based accountability for schools and educators has become a prominent feature of the education landscape. In particular, the use of sophisticated statistical methods to create performance measures from student achievement data, often through VAMs, has become more prevalent. VAMs attempt to predict the "value" a teacher would add to student achievement growth, as measured by standardized test scores, if each teacher taught comparable students under the same conditions. VAM results are often regarded as more objective or authoritative than other types of information because they are based on student outcomes, use quantitative complex models, and rely on standardized test scores and common procedures for all teachers or schools.

This statement by the American Statistical Association provides guidance as to what can and cannot be reasonably expected, given current knowledge and experience, from use of VAMs. It is intended to enhance general understanding of the strengths and limitations of the results generated by VAMs and thereby encourage the informed use of these results. It is not meant to be prescriptive or advocate any particular VAM specification or promote or condemn specific uses of VAM.

Value-added models typically use a form of regression model predicting student scores or growth on standardized tests from background variables (including prior test scores), with terms in the model for the teachers who have taught the student. The model coefficients for the teachers are used to calculate their VAM scores. In related models known as "growth models" a regression model is fit to predict students' current test scores from previous test scores. A percentile is calculated for each student from the model, relating his or her growth to the growth of other students with similar previous test scores. The median or average of the percentiles of a teacher's students is then used to calculate the teacher's VAM score. The statistical issues underlying the use of these various types of models are similar, and in this statement, the term "VAM" is used to describe both traditional value-added models and growth models. In both types of models, if a teacher's students have high achievement growth relative to other students with similar prior achievement, then the teacher will have a high VAM score. Some VAMs also include other background variables for the students.

There are a number of key questions states and districts should address regarding the use of any type of VAM. VAMs are being used for the evaluation of individual teachers on the basis of claims that they can measure those teachers' effects on student achievement growth. These questions are concerned with how well VAMs measure these effects and how the results should be interpreted.

- The measure of student achievement is typically a score on a standardized test, and VAMs are only as good as the data fed into them. Ideally, tests should fully measure student achievement with respect to the curriculum objectives and content standards adopted by the state, in both breadth and depth. In practice, no test meets this stringent standard, and it needs to be recognized that, at best, most VAMs predict only performance on the test and not necessarily long-range learning outcomes. Other student outcomes are predicted only to the extent that they are correlated with test scores. A teacher's efforts to encourage students' creativity or help colleagues improve their instruction, for example, are not explicitly recognized in VAMs.
- VAM scores are calculated from classroom-level heterogeneity that is not explained by the background variables in the regression model. Those classroom-level differences may be due in part to other factors that are not included in the model (for example, class size, teaching "high-need" students, or having students who receive extracurricular tutoring). The validity of the VAM scores as a measure of teacher contributions depends on how well the particular regression model adopted adjusts for other factors that might systematically affect, or bias, a teacher's VAM score.

  The form of the model may lead to biased VAM scores for some teachers. For example, "gifted" students or those with disabilities may exhibit smaller gains in test scores if the model does not accurately account for their status.
- VAM scores are calculated using a statistical model, and all estimates have standard errors. VAM scores should always be reported with associated measures of their precision, as well as discussion of possible sources of biases.

VAMs are complicated statistical models, and they require high levels of statistical expertise. Sound statistical practices need to be used when developing and interpreting them, especially when they are part of a high-stakes accountability system. These practices include evaluating

model assumptions, checking how well the model fits the data, investigating sensitivity of estimates to aspects of the model, reporting measures of estimated precision such as confidence intervals or standard errors, and assessing the usefulness of the models for answering the desired questions about teacher effectiveness and how to improve the educational system.

Quality Improvement and Value-Added Models

Statistical science has a rich and continuing history of successful contributions to quality improvement undertakings. While the methods and approaches vary, consensus exists that:

1. The quality improvement process should be monitored and informed using relevant quantitative information;
2. Almost all systems of measurement contain random variation;
3. Attaching too much importance to a single item of quantitative information is counter-productive—in fact, it can be detrimental to the goal of improving quality. In particular, making changes in response to aspects of quantitative information that are actually random variation can increase the overall variability of the system.

When used appropriately, VAMs may provide quantitative information that is relevant for improving education processes. For example, the models can provide information on important sources of variability, and they can allow teachers and schools to see how their students have performed on the assessment instruments relative to students with similar prior test scores. Teachers and schools can then explore targeted new teaching techniques or professional development activities, while building on their strengths.

Using VAM scores to improve education requires that they provide meaningful information about a teacher's ability to promote student learning. For instance, VAM scores should predict how teachers' students will progress in later grades and how their future students will fare under their tutelage. Various studies have demonstrated positive correlations between teachers' VAM scores and their students' future academic performance and other long term outcomes. In a limited number of studies, teachers have been randomly assigned to classes within schools, thus reducing systematic effects that might arise because of assignment of students to teachers. These studies indicate that the VAM score of a teacher in the year before randomization is positively

correlated with the test score gains of the teacher's students in the year after randomization, but the correlations are generally less than 0.5. Also, studies have shown that teachers' VAM scores in one year predict their scores in later years.

These studies, however, have taken place in districts in which VAMs are used for low-stakes purposes. The models fit under these circumstances do not necessarily predict the relationship between VAM scores and student test score gains that would result if VAMs were implemented for high-stakes purposes such as awarding tenure, making salary decisions, or dismissing teachers.

The quality of education is not one event but a system of many interacting components. The impact of high-stakes uses of VAMs on the education system depends not only on the statistical properties of the VAM results but on their deployment in the system, especially with regard to how various types of evidence contribute to an overall evaluation and to consequences for teachers.

It is unknown how full implementation of an accountability system incorporating test-based indicators, such as those derived from VAMs, will affect the actions and dispositions of teachers, principals and other educators. Perceptions of transparency, fairness and credibility will be crucial in determining the degree of success of the system as a whole in achieving its goals of improving the quality of teaching. Given the unpredictability of such complex interacting forces, it is difficult to anticipate how the education system as a whole will be affected and how the educator labor market will respond. We know from experience with other quality improvement undertakings that changes in evaluation strategy have unintended consequences. A decision to use VAMs for teacher evaluations might change the way the tests are viewed and lead to changes in the school environment. For example, more classroom time might be spent on test preparation and on specific content from the test at the exclusion of content that may lead to better long-term learning gains or motivation for students. Certain schools may be hard to staff if there is a perception that it is harder for teachers to achieve good VAM scores when working in them. Overreliance on VAM scores may foster a competitive environment, discouraging collaboration and efforts to improve the educational system as a whole.

Research on VAMs has been fairly consistent that aspects of educational effectiveness that are measurable and within teacher control represent a small part of the total variation in student test scores or growth; most estimates in the literature attribute between 1% and 14% of the total variability to teachers. This is not saying that *teachers* have little effect on students, but that *variation* among teachers accounts for a small part of the variation in scores. The majority of the variation in test scores is attributable to factors outside of the teacher's control such as student and family background, poverty, curriculum, and unmeasured influences.

The VAM scores themselves have large standard errors, even when calculated using several years of data. These large standard errors make rankings unstable, even under the best scenarios for modeling. Combining VAMs across multiple years decreases the standard error of VAM scores. Multiple years of data, however, do not help problems caused when a model systematically undervalues teachers who work in specific contexts or with specific types of students, since that systematic undervaluation would be present in every year of data.

A VAM score may provide teachers and administrators with information on their students' performance and identify areas where improvement is needed, but it does not provide information on how to improve the teaching. The models, however, may be used to evaluate effects of policies or teacher training programs by comparing the average VAM scores of teachers from different programs. In these uses, the VAM scores partially adjust for the differing backgrounds of the students, and averaging the results over different teachers improves the stability of the estimates.

Statistical science has an important role to play in raising the quality of education, through developing and refining statistical models for use in education, providing guidance on designing experiments and interpreting statistical results, and applying quality and process improvement expertise to help guide judgments in the presence of uncertainty. The ASA promotes sound use of statistical methodology for improving education.