ASA Statement on Value-Added Models

David Morganstein and Ron Wasserstein

The ASA issued a statement to inform the current national discussion on the use of value-added models (VAMs) in making high-stakes decisions regarding teacher performance appraisals and compensation. This article outlines a number of issues behind the decision to make this public statement.

1. INTRODUCTION

Sallie Keller’s address as 2006 ASA President challenged the statistical community to engage in improving science for policy (Keller-McNulty 2007). Keller said that statisticians are well poised to assume roles as integrators of science but need to be positioned closer to where policy is made.

Keller led with words and with actions. Under her leadership, the ASA created a Director of Science Policy and has in the ensuing years made a mark advocating for the statistics profession and for science in general.

An important part of that advocacy is issuing public statements about topics related to using data and statistical tools (http://www.amstat.org/policy/boardstatements.cfm). These statements are member-driven and Board-approved. That is, members identify an issue to address and provide the expertise and energy to develop a well-reasoned statement. The Board reviews the statement, modifies it (or asks that it be modified) as necessary, and then approves and publicizes it.

Part of the ASA’s mission is “promoting sound statistical practice to improve public policy and improve human welfare” (http://www.amstat.org/about/index.cfm). In that spirit, the ASA decided to make a statement to inform the discussions of the use of value-added models (VAMs) for educational assessment because a statistician’s perspective is important, especially where states and local governments use VAMs to make high stakes decisions regarding teacher performance appraisals and compensation.

2. ASSUMPTIONS AND LIMITATIONS

As with all statistical inference, VAMs are based on assumptions. The assumptions may be about the form of an assumed model. They may relate to the distribution of errors and have an impact on the inferences that can be drawn about future estimates made from the model results. It is critical, however, that individuals using these models understand the assumptions on which they are based as well as their limitations.

For example, some procedures assume that the underlying error is distributed normally (the “bell-shaped curve”) or that cases are assigned randomly to various treatment conditions. Statistical models often include an assumed form with unknown coefficients, such as, $y = mx + b + \text{error}$, where the association between $x$ and $y$ is measured through the coefficients $m$ and $b$. Often the coefficients in the model are estimated from data collected in a random way via an experiment. In such cases, the standard errors of the coefficients can be used to determine the statistical significance of the model coefficients and to create confidence intervals around a prediction of $y$ from a specific $x$.

Most VAMs are linear models albeit complex ones. These and other statistical models can be valuable in many areas of decision-making. The value statisticians add is to appreciate the risks associated with the decision by incorporating variability into an understanding of the likelihood of gains or losses estimated by these models. Two errors are commonly discussed when using data in decision-making such as high-stakes decisions about teacher performance: saying something is effective (or ineffective) when it isn’t (over-reacting) and saying it isn’t when it is (missed opportunity). These risks must be assessed if the user is to understand the limitations of the results.

3. IDENTIFYING SPECIAL CAUSES

To determine whether VAMs can identify very strong or very weak teachers, at least as they relate to improvement in test scores, requires a next step. Not only is an individual VAM score calculated for each teacher but the standard error of that score is also needed to determine if the individual teacher’s score came from a distribution of “strong” teachers or a distribution of “weak” teachers. Alternatively there may be only one overall teacher distribution indicating they are all part of the same system variation. Deming and other quality experts referred to this as looking for special causes in the presence of common cause (or system) variation. In the linear VAM models mentioned earlier, this would appear as a statistically significant, nonzero teacher effect.

Two key benefits are derived from using statistical methods: an appreciation for the importance of including error estimates and an understanding of how generalizable any particular survey or data experiment might be to other situations.

The error estimates are necessary to distinguish special causes, for example, an exceptional teacher, from system variation. Ranking teachers by a score and identifying the “top five” of them as somehow preferred is not statistically sound. After all, a “top five” and a “bottom five” always exist, even among the 100 greatest teachers of all time. Ranks are only useful when the error estimates for individual teacher scores indicate they (or any subset of teachers) appear to come from a different distribution.

David Morganstein, Department of Statistics, Westat Inc., Rockville, MD 20850 (E-mail: davidmorganstein@westat.com). Ron Wasserstein, American Statistical Association, Alexandria, VA (E-mail: ron@amstat.org).
4. RANDOMNESS

Students are generally not assigned randomly to classrooms. However, the foundations of the statistical theory behind inferences, where extrapolations are made from a sample to a population, generally assume random assignments have been made. If a comparison is to be made, say between two teachers (or teaching methods, or two schools), students would be randomly assigned to the alternatives. Observed differences in student growth scores would be compared to the differences that could have arisen from the random assignment. Statistically significant differences are noted as variation that was unlikely to occur had it arisen solely from the underlying system variation in the process.

When there is no random assignment, for example, students generally are not randomly assigned to teachers, some other mechanism or assumption is needed for inferences. The challenge with VAMs is to include all the important factors that might contribute to the observed differences in test scores. Many potential explanatory variables are not available for inclusion or have many missing values in estimating teacher VAM scores.

5. NO UNIVERSALLY ACCEPTED SOLUTION

Many VAM models have been proposed and are being used, each with their own set of assumptions. The same set of data has been observed to yield different conclusions from different VAM models (Newton et al. 2010; Baker et al. 2010; Goldhaber, Walsh, and Gabele 2014). Likewise, many variables may be included in any given model, and the data are often dependent on information available at a specific school system. Various entities, states, and school districts use different VAM models and different variables. The use of various models on the same set of data has resulted in different conclusions about teacher performance. Ignoring for the moment serious concerns about confusing correlation with causality and the relevance of changes in test scores to predicting future student success, the lack of consistency among various methods using the same data alone raises concerns about the application of VAMs for high-stakes decisions.

To date, no universally accepted approach to such models has been carefully tested. Our profession values transparency and whenever a VAM approach is used the statistical methods and software implementing them should be described, tested, and documented thoroughly. Not having a well-accepted, well-tested method increases the pressure to solve this important national concern, potentially resulting in making critical errors in finding solutions.

As one example, the Washington Post reported (Anderson 2013) that scores in Washington, DC, Public Schools (DCPS) “IMPACT” system, which evaluated teachers, had been erroneously calculated for 44 teachers, about 10% of those who had been evaluated through the DCPS VAM approach. One of those teachers was fired as a result of the error. (That teacher was rehired when the error was discovered.) Such unfortunate errors can reduce confidence in the entire process, particularly when the details of the model are not well documented or widely accepted. The higher the stakes, the more important teacher and public confidence in the procedures become.

6. UNINTENDED CONSEQUENCES

Actions always have effects that are unanticipated or unintended. Even if a highly reliable, clearly defined and accepted process were available that could identify superior or inferior teachers based on changes in their students’ test scores, care must be taken in the decision to use it for rewards and punishments. The “law of unintended consequences” must be considered.

Some unintended consequences may ensue if teachers shift emphasis away from activities that inspire their students and encourage them to learn to, instead, applying that effort toward test preparation. The media has reported numerous instances of teachers and principals providing improper help to their students or even falsifying test results (Aviv 2014). If teachers believe that their job, pay, or bonuses derive from improvements in their students test scores, very talented and capable teachers may choose another profession. Students’ educational experience could suffer if the deployment of VAM scores for high-stakes decisions is not done carefully, particularly with regard to how various types of evidence contribute to an overall evaluation and to consequences for teachers.

7. VALUABLE ROLE

VAMs can play a valuable role in improving the quality of education. Their value is clearer in assessing more aggregate level effects, such as curricula and methods for teaching, but less clear in evaluating individual instructor performance or other potentially punitive uses that give undue weight to changes in test scores.

The ASA and the statistical community do much more than issue statements. Members of the community are involved in research to better understand and improve statistical methods that may be used in education. Statisticians continue to have an important role in collaborating with educational organizations that wish to use data to improve education. That collaboration, in many respects, is the most important message in the ASA’s VAM statement: “VAMs are complex statistical models, and high-level statistical expertise is needed to develop the models and interpret their results.”

8. CONCLUSION

VAMs play a role in educational assessment. They can help evaluate teaching programs and, possibly, compare school districts. However, if data and statistical models are used without regard for their assumptions and limitations, there may be unintended consequences that do not result in improving teacher performance or our educational system.

[Received July 2014. Revised August 2014.]

REFERENCES

Anderson, N. (2013, December 23), “D.C. School Officials: 44 Teachers Were Given Mistaken Performance Evaluations,” Washington Post. Available at http://www.washingtonpost.com/local/education/dc-school-officials-44-teachers-were-given-mistaken-performance-evaluations/2013/12/23/5cb9f26-6c0c-11e3-a523-je730f6b8d_story.html. [109]
Aviv, R. (2014, July 21), “Rachel Aviv: A Middle-School Cheating Scandal Raises Questions About No Child Left Behind,” The New Yorker. Available at http://www.newyorker.com/reporting/2014/07/21/140721fa_fact_aviv?currentPage=all. [109]

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., and Shepard, L. A. (2010), Problems With the Use of Student Test Scores to Evaluate Teachers, Washington, D.C.: Economic Policy Institute. [109]

Goldhaber, D., Walch, J., and Gable, B. (2014), “Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments,” Statistics and Public Policy, 1, 28–39. [109]

Keller, S. (2007), “From Data to Policy: Scientific Excellence is Our Future,” Journal of the American Statistical Association, 102, 395–399. Available at https://www.ma.utexas.edu/seminar/sallie.pdf. [108]

Newton, X. A., Darling-Hammond, L., Haertel, E., and Thomas, E. (2010), “Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability Across Models and Contexts,” Educational Policy Analysis Archives, 18. Available at http://dx.doi.org/10.14507/epaa.v18n23.2010. [109]