DISCUSSION OF THE AMERICAN STATISTICAL ASSOCIATION'S STATEMENT (2014) ON USING VALUE-ADDED MODELS FOR EDUCATIONAL ASSESSMENT

Raj Chetty, Harvard University
John Friedman, Harvard University
Jonah Rockoff, Columbia University

May 2014

In a recent statement, the American Statistical Association (ASA) discusses the use of value-added measurement (VAM) to evaluate teacher quality. The ASA statement is carefully crafted and explicitly states that its intention is not to "promote or condemn specific uses of VAM." However, some commentators have interpreted the statement as expressing a negative view of VAM, while others disagree.

In this note, we present our views on the issues raised by the ASA in light of research we and others have done on this subject over the past five years. While we agree with many of the points made by the ASA, their statement provided no references to the literature. Hence, it is unclear whether the ASA's statement fully incorporates the results of recent research that addresses many of the concerns it raises.

In what follows, we provide a point-by-point discussion of the ASA statement in the context of the recent academic literature and policy debates on value-added measurement. Like the ASA, our objective is not to advocate a specific policy, but rather to clarify which issues have been largely resolved and which issues remain open to debate and should be a priority for future research.

***ASA Point #1**: Data and statistical models should be used wisely and experiments should be designed for improving the quality of education.*

Discussion: This statement is undoubtedly correct: no one involved in the academic or policy debate supports unwise use of data or statistics or the neglect of experimental design in efforts to improve the quality of education. Consistent with the ASA's recommendation, there are now multiple experiments that have been designed specifically to evaluate the validity of VAM estimates, which we return to below.

***ASA Points #2 and #3**: 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.*

Discussion: We agree that some VAMs are complex statistical models. However, the development of a VAM model and the interpretation of its results require very different sets of skills. School administrators, teachers, and other relevant parties can be trained to understand how to interpret a VAM estimate properly, including measures of precision as well as the assumptions and limitations of VAM (as demonstrated in recent research). In practice, measures of precision and a discussion of a VAM model's assumptions and limitations almost always accompany VAM estimates, both in academic work and in policy, in line with the view expressed by the ASA.

***ASA Point #4**: VAMs are generally based on standardized test scores, and do not directly measure potential teacher contributions toward other student outcomes.*

Discussion: Recent research has shown that VAMs do in fact capture teachers' impacts on students' other outcomes. For example, in a paper on teachers' long-term impacts (forthcoming in the *American Economic Review*) we tested the link between teachers' impacts on standardized test scores, as measured by their value-added, and students' long-term outcomes. We find that being assigned to a high value-added teacher makes students more likely to attend college, increases their earnings, and has a variety of other positive long-term outcomes, such as reduced teenage birth rates. Subsequent work by Chamberlain (2013) using different statistical methods reached similar conclusions. These studies indicate that VAMs capture some portion of teachers' contributions to long-term outcomes of interest. However, this does not mean that VAM's capture all aspects of teacher "quality." Indeed, it is likely that test scores cannot capture some important channels through which teachers affect their students' future outcomes. For instance, a recent study uses VAM procedures to measure high school teachers' impacts on students' non-cognitive skill acquisition and finds these effects are an important complement to teachers' effects on cognitive test scores. This illustrates why policy makers should combine VAM with other measures when evaluating teacher performance.

***ASA Point #5**: 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.*

Discussion: We believe the ASA may have overlooked a large body of recent experimental and quasi-experimental evidence showing that VAM estimates provide information about the causal impacts of teachers on their students' test score growth. This includes evidence from four separate studies that have directly tested whether VAMs measure correlation or causation: (1) an experimental study by Kane and Staiger (2008) in the Los Angeles Unified School District, (2) an experimental study funded by the Bill and Melinda Gates Foundation, (3) our own research, which provides a quasi-experimental method for testing whether VAMs measure causation based on teacher turnover; and (4) a replication of our methodology in the Los Angeles Unified School District. All four of these studies reach the same conclusion: VAMs that control for students' lagged test scores primarily capture teachers' causal effects rather than correlations due to other factors

not captured in the model. To our knowledge, there is no experimental or quasi-experimental study to date that reaches the opposite conclusion. While no single experiment is definitive, the academic literature appears to have reached a consensus on the issue of correlation vs. causation.

***ASA Point #6**: 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.*

Discussion: The meaning of "substantial" is unclear, but our reading of the literature is that the various methods used in research and policy to estimate value-added yield very similar estimates (i.e., correlations usually well above 0.90) so long as they control flexibly for students' prior achievement. We demonstrate this in Table 6 of our paper on bias in VA estimates, but this point has been made elsewhere, including two studies published earlier this year in *Statistics and Public Policy*, a journal of the American Statistical Association.

To be clear, the key finding is that VAM predicts how much teachers will raise the test scores of their future students on average, even though the test scores of any individual child can be affected by other factors unaccounted for by VAM , such as illness, stress, or lucky guessing. The role of luck is inescapable in any measure of job performance. While there is no doubt that VA measures are not perfectly reliable, the question is how reliable they are relative to other potential methods of evaluation. For example, standard teacher evaluations are usually based on a single observer who views only one or two of the hundreds of lessons that a teacher delivers over the course of a school year. Research confirms that it matters if an observer shows up on an especially good or bad day, and that this type of standard evaluation procedure may be less reliable than VAM estimation. The fact that classroom observation and VAM are both imperfect measures underscores why educators and policymakers are likely to make better decisions if they are based on multiple measures of job performance rather than any stand-alone metric.

***ASA Point #7**: 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.*

Discussion: The ASA is correct in noting that the majority of variation in student test scores is "attributable to factors outside of the teacher's control," and that this "is not saying that teachers have little effect on students." These statements sum up some of the major findings from VAM research. First, comparing teachers based on raw student tests scores without a VAM approach would be biased against teachers serving students from disadvantaged backgrounds. A VAM approach helps to level the playing field, so that students' knowledge and skills when they enter a classroom are not an impediment to a teacher receiving a positive evaluation of their performance.

Second, while it is true that a single teacher is unlikely to turn a remedial student into an honors student, our paper on teachers' long-term impacts of shows that teachers do have meaningful effects on students. For example, we estimate that being assigned to a high-value added (top 5%) rather than an average teacher for a single grade raises a student's lifetime earnings by more than $50,000. The fact that there is a lot of variance in student achievement due to numerous other factors – such as parents, neighborhoods, or health – does not take away from the important role that teachers can and do play in improving students' outcomes.

The ASA appropriately warns that "ranking teachers by their VAM scores can have unintended consequences that reduce quality." In particular, it is possible that teachers may feel pressured to teach to the test or even cheat if they are evaluated based on VAMs. The empirical magnitude of this problem – and potential solutions if it turns out to be a serious concern – can only be assessed by studying the behavior of teachers in districts that have started to use VAMs.

In summary, our view is that many of the important concerns about VAM raised by the ASA have been addressed in recent experimental and quasi-experimental studies. Nevertheless, we caution that there are still at least two important concerns that remain in using VAM for the purposes of teacher evaluation. First, using VAM for high-stakes evaluation could lead to unproductive responses such as teaching to the test or cheating; to date, there is insufficient evidence to assess the importance of this concern. Second, other measures of teacher performance, such as principal evaluations, student ratings, or classroom observation, may ultimately prove to be better predictors of teachers' long-term impacts on students than VAMs. While we have learned much about VAM through statistical research, further work is needed to understand how VAM estimates should (or should not) be combined with other metrics to identify and retain effective teachers.

References

American Statistical Association. April 8, 2014. "ASA Statement on Using Value-Added Models for Educational Assessment." http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Chamberlain, Gary. 2013. "Predictive Effects of Teachers and Schools on Test Scores, College Attendance, and Earnings." *Proceedings of the National Academy of Sciences* 110(43).

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." Forthcoming, *American Economic Review* (September 2014).

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." Forthcoming, *American Economic Review* (September 2014).

Ehlert, Mark, Cory Koedel, Eric Parsons, and Michael J. Podgursky. 2014. "The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence from School- and Teacher-Level Models in Missouri." *Statistics and Public Policy*, 1(1): pp. 19-27.

Glazerman, Steven, Ali Protik, Bing-ru Teh, Julie Bruch, and Jeffrey Max. 2013. "Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Experiment (NCEE 2014-4003)". Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Goldhaber, Dan, Joe Walch and Brian Gabele. 2014. "Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments." *Statistics and Public Policy*, 1(1): pp. 28-39.

Ho, Andrew D., and Thomas J. Kane. 2013. "The Reliability of Classroom Observations by School Personnel." Bill and Melinda Gates Foundation.

Jackson, C. Kirabo. 2012. "Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina." NBER Working Paper No. 18624.

Kane, Thomas J. and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper No. 14607.

Kane, Thomas J., Douglas O. Staiger, and Andrew Bacher-Hicks. 2014. "Validating Teacher Effect Estimates using Between School Movers: A Replication and Extension of Chetty et al." Harvard University Working Paper.

Neal, Derek. 2011. "New Assessments for Improved Accountability." Hamilton Project Policy Brief 2011-09.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools." *American Economic Review*, December 2012, pp. 3184-3213.