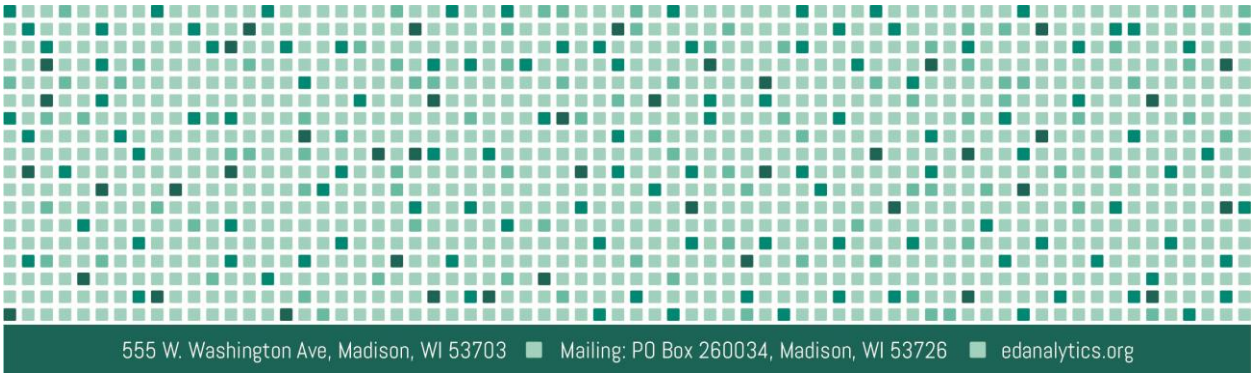
Education Analytics INC.

Using Value-Added Models for Educational Assessment

Robert H. Meyer and Stephen M. Ponisciak

December 3rd, 2014



USING VALUE-ADDED MODELS FOR EDUCATIONAL ASSESSMENT

Robert H. Meyer and Stephen M. Ponisciak[1]

On April 8, 2014, the American Statistical Association released the "ASA Statement on Using Value-Added Models for Educational Assessment." Below, we provide comments on the ASA statement and provide examples from our work at the Value-Added Research Center and Educational Analytics Inc. with districts and their stakeholders on the development and implementation of value-added models. We also provide a list of resources that provide additional information on these issues.

We agree with many of the points in the ASA's statement on value-added models, and our value-added models are already aligned with many of them. In particular, we concur that there are strengths and limitations associated with all prospective measures of school and educator effectiveness, including both measures of educators' professional practice as well as various student outcome measures. It is for precisely this reason that states and districts adhering to best practices have moved to incorporate a "multiple measures" approach that avoids placing too much weight on any single measure, although states are clearly adopting different definitions of how much weight constitutes "too much."

We also agree with the claim that "VAMs are only as good as the data fed into them." This is one reason why it is important to accurately capture and store student-teacher links and use a roster verification process to audit this data (either the full data set or a sub-sample). Many of our partner districts conduct an audit of student-teacher links for all teachers and thus can use this corrected data to estimate value-added models. In addition, all test records used in the value-added model are validated by district central office staff, and the value-added results themselves are subject to a rigorous series of quality control (QC) checks by our organizations.

ASA is correct in stating that "VAM scores are calculated from classroom-level heterogeneity [differences in student achievement] that is not explained by the background variables in the regression model." ASA further notes that "[t]he 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." It is for this reason that our partner districts have engaged with local stakeholders, Technical Advisory Committees, and national experts to identify and include in the model a comprehensive set of student-level

[1] Meyer is Research Professor, La Follette School of Public Affairs and Wisconsin Center for Education Research, and Director, Value-Added Research Center, at the University of Wisconsin-Madison; and President and CEO of Education Analytics Inc. Ponisciak is Researcher at the Wisconsin Center for Education Research and Education Analytics Inc.

predictors of student achievement growth that are typically beyond the control of schools and educators, including for example:

- prior mathematics and reading achievement (corrected for test measurement error)
- grade level
- gender
- race/ethnicity
- free/reduced-price lunch status
- homelessness
- English learner status
- special education status
- an indicator of student mobility (to account for the potentially disruptive effect of mobility)
- student/classroom (and student/school) "dose" variables that represent the fraction of the school year in which a given student was taught by a given teacher (in a given subject area) – equal to one (1) if a student was taught by a single teacher.

We have also explored including classroom-level measures to capture peer effects.

ASA states that "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." This comment is routinely invoked for evaluation models in which assignment to alternative "treatments" (classrooms and schools in our context) is not random, as in a classical randomized control trial (RCT). It is apparent that random assignment is generally infeasible in this context (assignment of teachers to students), and may even be undesirable from the standpoint of limiting the ability of principals and other school leaders to make optimal "matches" between the individual needs of students and the strengths of teachers.

Instead, VAMs and related growth models rely on quasi-experimental methods to control as much as possible for differences in student growth that are not due to teachers or schools. In particular (as mentioned above), the strategy is to include measures of prior achievement (in one or more subject areas) and an extensive list of student (and possibly classroom) variables, to control for differences in student growth. In our value-added models, these predictors typically capture 60-80% of the variability in post student achievement. The recent Measures of Effective Teaching (MET) Study (Kane et al.,2013; Kane, 2014) addressed the issue of whether, after controlling for prior achievement and other predictors of student achievement, there were unobserved differences in growth between students assigned to different classrooms within the same school. The MET Study tested this hypothesis by working with participating schools in the study to randomly assign students to classrooms. Then, value-added estimates were computed using the data based on random assignment and data from the previous year in which students were assigned in the usual manner. The study found no evidence that the non-experimental estimates differed from the experimental estimate. This suggests that estimated value-added effects may be highly correlated with causal effects. In order to protect against the possibility that value-added effects may be biased

to some degree, our value-added models and estimates are subject to a battery of quality control (QC) processes designed to flag possibly biased or non-neutral results. The QC process continues to be refined over time as new and better methods are developed to assess validity and robustness of the results produced by quasi-experimental models.

Value-added or growth models that use limited (or nonexistent) sets of control variables may yield estimates of educator effectiveness that are significantly biased (thus, the need to properly specify these models and use well-measured data). Student growth percentile models (which are often defined as descriptive, but used as if they are causal) are subject to this criticism since they do not include student or classroom variables (e.g., free lunch and EL status) that are predictive of student growth and do not control for test measurement error (Akram and Meyer, 2014). In particular, schools and teachers that serve a larger share of certain subgroups, such as students with disabilities, will typically fare worse when no attempt is made to control for these factors in growth models. This raises important policy questions around consequential validity, in the sense that creating disincentives to avoid serving these types of students seems unwise.

ASA states that "under some conditions, VAM scores and rankings can change substantially when a different model or test is used." It is for this reason, as discussed above, that it is our practice to assess the robustness of the adopted model to alternative assumptions and to alternative sets of predictor variables and to share the results of this analysis with district central office staff and Technical Advisory Committee (TAC) members. This process has resulted in models that are quite robust.

With respect to the issue of the choice of assessments used to measure student achievement, growth in student achievement, and educator effectiveness, the most useful and valid information is obtained from assessments that are "curriculum sensitive;" that is, aligned with established learning standards and with what is taught in the classroom (Pellegrino et al, 2001). Test reliability – overall and by student – is also important since it affects the precision of estimated educator effects.

Of course, all test scores are imperfect proxies for student knowledge and measure student achievement with some error; this is one reason why our value-added models include an adjustment for test measurement error. But it is also true that other components of teacher evaluation, including commonly-used rubrics for evaluating teachers' professional practice, are subject to statistical error (primarily via lack of consistency across raters) and possible bias, which is frequently not acknowledged or reported (Ho and Kane, 2013; Ponisciak et al, 2014; Whitehurst et al, 2014). We agree that as the stakes associated with value-added models increase, so does the potential for increased instructional time spent on narrow test preparation. On the other hand, it is arguably a very good thing for teachers to teach to high quality learning standards and the content domains that underlie assessments aligned to standards. Most evaluation systems address the two

above concerns by using multiple measures to evaluate teachers. From a measurement perspective, the use of multiple indicators enhances the validity and precision of overall, composite measures of educator effectiveness,

The claim that "most VAMs predict only performance on the test and not necessarily long-range learning outcomes" is not necessarily true, as one can see in the work of Chetty et al (2014a, 2014b). The opposite statement is made later in the ASA document: "Various studies have demonstrated positive correlations between teachers' VAM scores and their students' future academic performance and other long term outcomes."

The ASA statement makes three points with respect to the magnitude of estimated teacher effects:

1. "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."
2. "Most VAM studies find that teachers account for about 1% to 14% of the variability in test scores."
3. "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."

It is important when considering these points to distinguish between: (1) the contribution of teachers in a single school year and (2) the cumulative contribution of teachers from grades K (or earlier) through 12. The ASA statement focuses on the first point and thus underestimates the important contributions that high-performing teachers make over time to student learning and measured student achievement. Even so, the contributions of high-performing (versus low-performing) teachers over a single school year are practically important. Our review of value-added results from multiple districts indicate that the amount of variation explained by teachers is typically around 9% in mathematics and 5% in reading or English Language Arts (ELA). These numbers correspond to standard deviations of 0.30 and 0.23, respectively, which represent effects that are comparable to the size of achievement gaps between schools with low versus average achievement as well as to the size of achievement effects in randomized studies (Lipsey et al., 2012). The cumulative contribution of teachers over time substantially exceeds the contributions from a single year, although the total effect depends on the degree to which growth over a single year decays. Allowing for decay, our simulations suggest that the cumulative effect over 3 years approximately doubles the effect for a single year, and the effect of having a better teacher in all elementary school grades (K-5) more than triples the effect for a single year. The bottom line is that differences in teachers' contributions to student achievement are practically important, particularly when it is acknowledged that students attend school for multiple years.

We strongly agree with the statement that all results should be reported with standard errors, and that any model assumptions should be explained. Our value-added results are reported with standard errors, and color-coded to indicate statistical significance. In addition, we provide technical documentation outlining the features and assumptions of the model. We would extend

ASA's position on acknowledging statistical precision and model assumptions even further, in fact, by arguing that *all* components of teacher and principal evaluation – and not just value-added - should be subject to these requirements, as should the overall composite measure. Continuing to assume that only value-added measures are subject to imprecision, in other words, is more than a little misleading, and could lead to undermined confidence in educator evaluation systems down the road.

In conclusion, many districts and states use multiple indicators, including value-added measures, to measure the effectiveness of educators and schools. The use of multiple indicators, as discussed above, enhances both the validity and reliability of composite measures of educator effectiveness. One further benefit of this approach is that it signals to students, stakeholders, and the public that it is important for educators to focus on the complementary goals of using the best possible instructional practices and improving the learning outcomes of all students. We would argue that a balanced system that focuses on instructional practices *and* student outcomes is likely, in a "checks and balances" sense, to strengthen the integrity of both dimensions of the system.

RESOURCES

VALUE-ADDED AND GROWTH MODELS

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25 (1), 95–135.

Akram, K. and Meyer, R.H. (2014). *Comparison of Student Growth Percentile and Value-Added Models for Estimating Educator and School Effectiveness*, paper presented at the Annual Meeting of the Association of Education Finance and Policy, San Antonio, Texas.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice* 28 (4), 42-51.

Center for Educator Compensation Reform (undated). Research Synthesis A. Retrieved June 3, 2014 from http://www.cecr.ed.gov/researchSyntheses/Research%20Synthesis_Q%20A1.pdf

Chetty, R., Friedman, J., and Rockoff, J. (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates, forthcoming *American Economic Review*. Retrieved June 12, 2014 from http://obs.rc.fas.harvard.edu/chetty/w19423.pdf

Chetty, R., Friedman, J., and Rockoff, J. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood, forthcoming *American Economic Review*. Retrieved June 3, 2014 from http://obs.rc.fas.harvard.edu/chetty/w19424.pdf

Coleman, J. (1966). Equality of educational opportunity (Report No.OE-38000). Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge, MA: Harvard Education Press.

Kane, T.J. (2014). "Do Value-Added Estimates Identify Causal Effects of Teachers and Schools?" *The Brown Center Chalkboard*, Brookings Institution, October 30, 2014.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. MET Project Research Paper, Bill & Melinda Gates Foundation.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1): 67–101

McCaffrey, D., T. Sass, J.R. Lockwood, and K. Mihaly (2009). The Intertemporal Stability of Teacher Effects. *Education Finance and Policy* 4(4), 572-606.

Meyer, Robert H. and Dokumaci, Emin (2015). "Value-Added Models and the Next Generation Of Assessments," in Lissitz, Robert W. (ed.), *Value added modeling and growth modeling with particular application to teacher and school effectiveness*, Charlotte, NC: Information Age Publishing.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics* 125(1), 175-214.

Sanders, W.L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center. Retrieved June 3, 2014 from http://www.cgp.upenn.edu/pdf/Sanders_Rivers-TVASS_teacher%20effects.pdf

Sanders, W. L., Saxton, A. M., and Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A Quantitative, Outcome-Based Approach to Educational Assessment. In J. Millman (ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, CA: Corwin Press, 137-162.

Schochet, P. Z., & Hanley S. C. (2010). Error rates in measuring teacher and school performance based on student test score gains (NCEE 2010-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Value-Added Research Center (2013). *Technical Report on the Chicago School-Level Value-Added Model, Academic Year 2012-2013.* Madison, WI: Value-Added Research Center, University of Wisconsin-Madison.

Measures of Educator Effectiveness

Ho, A. and Kane, T.J. (2013). *The reliability of classroom observations by school personnel*, MET Project Research Paper, Bill & Melinda Gates Foundation.

Ponisciak, S., Gawade, N., Wang, C, & Meyer, R.H. (2014). *Use of Value-Added and Observational Ratings to Measure Educator Effectiveness: Evidence from the Hillsborough County School District*, Technical Report, Education Analytics Inc.

Whitehurst, G.J., Chingos, M.M, and Lindquist, K.M. (2014). *Evaluating Teachers with Classroom Observations Lessons Learned in Four Districts*, Brookings Institution.

Quasi-Experimental Methods

Heckman, J.J. and Vytlacil, E.J. (2007). Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation (Chapter 70), in James J. Heckman, J.J. and Leamer, E.E (ed.), *Handbook of Econometrics*, Volume 6, Part B, 4779–4874, Elsevier.

William R. Shadish, Thomas D. Cook, Donald T. Campbell (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton Mifflin.

Effect Size

Cohen, Jacob (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd edition). (Hillsdale, NJ: Lawrence Erlbaum).

Hill, C. J., Bloom, H.S., Black, A. R. and Lipsey, M. W. (2008), Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives, 2: 172-177.*

Lipsey, Mark W. (1990) *Design Sensitivity: Statistical Power for Experimental Research* (Newbury Park, CA: Sage Publications).

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the IES website at http://ies.ed.gov/ncser/

Student Assessment

Cordray, D., Pion, G., Brandt, C., Molefe, A, & Toby, M. (2012). The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement. (NCEE 2013–4000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academic Press.

U.S. Department of Education, Office of the Deputy Secretary, Implementation and Support Unit, Race to the Top Assessment: Partnership for Assessment of Readiness for College and Careers Year Two Report, Washington, DC, 2013.

U.S. Department of Education, Office of the Deputy Secretary, Implementation and Support Unit, Race to the Top Assessment: Smarter Balanced Assessment Consortium Year Two Report, Washington, DC, 2013.

Data Quality

Battelle for Kids. (2009). *The Importance of Accurately Linking Instruction to Students to Determine Teacher Effectiveness.*

Ponisciak, S., Akram, K., McCants, M., Erickson, F., and Meyer, R. (2012). The Effect of Student – Teacher Linkage Data Errors on Value-Added Results. Paper presented at the Annual Meeting of the Association of Education Finance and Policy, Boston, Mass.

Reporting Standards/Standard Errors

American Educational Research Association (2006). Standards for Reporting on Empirical Social Science Research in AERA Publications. *Educational Researcher*, 35 (6), 33-40.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, 63, 839-851.

International Committee of Medical Journal Editors (2013). *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals.* ICMJE. Retrieved June 3, 2014 at http://www.icmje.org/icmje-recommendations.pdf