The Education Policy Center
AT MICHIGAN STATE UNIVERSITY

Findings and Preliminary Recommendations from the Michigan State and Indiana University Research Study of Value-Added Models to Evaluate Teacher Performance

Cassandra M. Guarino, Indiana University

December 20, 2013

Author Information

Cassandra M. Guarino
Associate Professor of Educational Leadership and Policy Studies
Indiana University

This work has been supported in part by the Institute of Educational Sciences (IES) Pre-doctoral Training grant, No. R305B030011 and IES grant No. R305D100028, through the U.S Department of Education.

Suggested citation:

Guarino, C. 2013. Findings and preliminary recommendations from the Michigan State and Indiana University Research Study of value-added models to evaluate teacher performance. East Lansing: Education Policy Center at Michigan State University. Retrieved [Date] from http://education.msu.edu/epc/publications/documents/white-paper-teacher-value-added-project-synthesis-12-20-2013.pdf

White Paper:

Findings and Preliminary Recommendations from the Michigan State and Indiana University Research Study of Value-Added Models to Evaluate Teacher Performance

December 20, 2013

By
Cassandra M. Guarino, PhD
Principal Investigator and Project Director

I. Introduction

The push for accountability in public schooling has extended to the measurement of teacher performance, accelerated by federal efforts through Race to the Top. Currently, a large number of states and districts across the country are computing measures of teacher performance based on the standardized test scores of their students and using them in combination with other indicators to help categorize teachers as effective or ineffective.

With the implementation of the Common Core State Standards in a majority of states and the development of specific assessments that align with them, it is particularly urgent to open up the field right now to discuss and shed light on best practices in computing teacher performance measures. The market for new assessments coupled with derivative products that compute teacher effectiveness measures—each promoting a particular methodology—is becoming increasingly competitive. Policy makers must make informed decisions on which products and procedures they will use.

A number of research studies have recently come out concerning the strengths and limitations of different methodological choices. Despite mounting research evidence, however, states and districts have tended to avoid using the best value-added modeling procedures in favor of weaker methods, such as "growth" models, possibly because they lack information on the relative merits of different procedures. Therefore it is incumbent upon the research community to disseminate its findings to the policy community to aid in effective decision-making.

A research team at Michigan State and Indiana Universities has contributed a number of key insights to the body of research on best practices in computing teacher performance measures and is in the process of disseminating its findings. The next sections describe the project, its findings, and its preliminary recommendations for policy. The work of the research team is highly technical in nature; the descriptions that follow attempt to capture the import of the work in more general language. We encourage readers to look to the papers for a more complete and precise treatment of these important ideas.

II. Description of the Research Project

Principal Investigators Cassandra Guarino, Mark Reckase, and Jeffrey Wooldridge have led a large study entitled "Constructing Value-Added Models of Teacher Effectiveness that We Can Trust" to evaluate and identify how well commonly-used value-added and growth models estimate teacher effectiveness.

The study is supported by a research grant of $1.2 million in funding from the Institute of Education Sciences at the U.S. Department of Education.[1] The study began in May 2010 and is slated to end in May 2014. The three principal investigators have been supported by a team of research assistants composed of doctoral students from Michigan State University.[2]

The project addresses the following main research questions:

- How well do various approaches capture the true effectiveness of teachers in producing student achievement growth?
- How sensitive are these estimates of teacher effectiveness to different approaches and contexts?
- What can statistical tests tell us about how students are assigned to teachers and how these assignments affect the results provided by the different statistical models?

The team uses a helpful combination of analyses based on simulated and actual data. The simulations create student achievement data in which teacher effects are known and investigate the ability of different statistical approaches to recover the true teacher effects under different scenarios for assigning students to teachers. The actual data consist of longitudinal testing and administrative data from a large, diverse Southern state over the course of several years, from 2001 through 2008. The data are at the test item, student, teacher, and school levels, and students are linked to teachers.

The project has produced several products thus far, consisting primarily of research papers and presentations at research conferences. These are described in the next section. In addition, the team recently hosted an informative national conference. The national conference was entitled "Using Student Test Scores to Measure Teacher Performance: The State of the Art in Research and Practice" and took place on October 10-11 at Michigan State University. The conference featured more than 20 prominent researchers and policy makers as speakers and more than 80 active audience participants.

The project website provides information on project activities and provides access to all research papers and products: http://vam.educ.msu.edu/. Videos, slides, and papers associated with all of the national conference panel discussions and presentations can also be found on the project website, as well as a policy brief entitled: *"Highlights of the conference on Using Student Test Scores to Measure Teacher Performance: The State of the Art in Research and Practice."* To disseminate this information further, the Education Policy Center at Michigan State University has featured the policy brief as a "Hot Topic" on its website.

---

[1] IES grant No. R305D100028.

[2] Students assisting the project are predoctoral fellows from the Department of Economics and School of Education Measurement and Quantitative Methods program. They are: Andrew Bibler, Steven Dieterle (now Assistant Professor of Economics at the University of Edinburgh), Eun Hye Ham, Michelle Maxfield, Francis Smart, Brian Stacy, Paul Thompson, and Kelly Vosters. All but one student received support from IES Predoctoral Training Grant No. R305B090011 to participate in this project.

III. A Synopsis of Project Findings

The project currently has four web-accessible working papers that can be found on the website of the Education Policy Center at Michigan State University, as well as the project website, and the website of the Institute of Labor Studies (IZA). The following are brief descriptions of the various studies and the relevant findings for policymakers:

- The study entitled **Can Value-Added Measures of Teacher Performance be Trusted?** is forthcoming in the scholarly journal *Education Finance and Policy*. It is the foundational study for the project and uses simulations to show how a relatively simple value-added approach, which we term "Dynamic OLS," can produce better estimates of teacher performance across realistic student-teacher sorting scenarios than some of the more popular methods being used by researchers and policymakers. We point out that no one method accurately captures true teacher effects in all scenarios, and the potential for misclassifying teachers as high- or low-performing can be substantial depending on the context and the model used. We find, however, that the recommended model is more reliable across scenarios and is demonstrably preferable to other approaches. http://education.msu.edu/epc/publications/documents/WP18Guarino-Reckase-Wooldridge-2012-Can-Value-Added-Measures-of-Teacher-Performance-Be-T_000.pdf

- The study entitled **How do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value-Added** directly addresses the effects of assigning students to teachers in different ways. It shows that assigning students based on prior test scores occurs in a nontrivial number of schools, particularly as we approach the middle school grades. This important fact is obscured by earlier methods developed in prior research. Importantly, we find that performance estimates for teachers in schools that engage in sorting are affected by the particular type of value-added model being used. http://education.msu.edu/epc/publications/documents/WP30Dieterle-Guarino-Reckase-Wooldridge-2012-Student-TeacherAssignments-and-Value-added.pdf

- The study entitled **An Evaluation of Empirical Bayes Estimation of Value-Added Teacher Performance Measures** uses both simulation and actual data to show that some commonly used techniques, known in the field of statistics as Empirical Bayes (EB) estimation, do not reduce errors in ranking teachers by their effectiveness and, in many plausible situations, are outperformed by a simpler technique. In this paper we review the theory of EB estimation and use simulated data to study its ability to properly rank teachers. We find that EB estimators generally perform well if teachers are assigned randomly to classrooms. Otherwise, simpler techniques that control for teacher assignment, such as the Dynamic OLS approach, perform the best out of all the estimators we examine. http://education.msu.edu/epc/publications/documents/WP31Guarino-et-al-2012-Empirical_Bayes_Estimation-of-Value-added.pdf

- The study entitled **Does the Precision and Stability of Value-Added Estimates of Teacher Performance Depend on the Types of Students They Serve?** investigates how the precision and stability of a teacher's value-added estimate relates to student characteristics. It finds that the year-to-year stability of a teacher's value-added estimates can depend on the previous achievement level of his or her students. Teachers of lower achieving and more disadvantaged students have less stable value-added estimates and may be more likely to receive sanctions in a high stakes policy environment. This finding has important implications for practitioners implementing high stakes accountability policies, as teachers serving certain groups of students may be unfairly targeted for positive or negative sanctions simply because of the composition of their classroom and the variability this creates in their estimates. http://education.msu.edu/epc/publications/documents/WP35DoesthePrecisionandStabilityofValue-AddedEstimates.pdf

A number of other papers near completion are the following:

- **A Comparison of Growth Percentile and Value-Added Models of Teacher Performance**
  - This paper demonstrates that certain value-added models perform better than growth percentile models when students are not randomly assigned to teachers.

- **Classical Solutions to New Problems: Exploring Measurement Error Corrections in the Context of Teacher Value-Added**
  - This paper demonstrates how measurement error is unevenly distributed across student test scores and how this can affect value added estimates of teachers at the top and bottom of the effectiveness distribution.

- **Evaluating Specification Tests in the Context of Value-Added Models of Teacher Performance**
  - Recently, papers have appeared arguing that certain statistical tests reveal fundamental flaws in the assumptions made by some value-added models. This paper discusses different statistical tests that can be used to evaluate the methods used to estimate teacher effectiveness, including the "falsification" test proposed by Jesse Rothstein. We find that these tests do not accurately tell us whether models will work. Very helpful models can fail these tests for reasons that do not compromise their ability to capture teacher effectiveness.

- **State of the Art in Value-Added Models of Teacher Performance: Taking Stock of What We Know and Still Don't Know**
  - This paper presents a summary of what is currently known about value-added models of teacher performance and provides explanations and data analysis to illustrate differences across modeling approaches. It also highlights remaining questions for exploration.

IV. Preliminary Recommendations

Several recommendations have emerged from the above studies. The overall message for policymakers is that certain simpler, more robust value-added methods, such as the Dynamic OLS approach, often perform as well as or better than many of the models currently being proposed or adopted. More specifically:

1) **Value-added models should include so-called "teacher fixed effects" to take account of the non-random assignment of students to teachers.**
    - Efforts to evaluate the effectiveness of doctors typically take into account the ease or difficulty of the individual cases they handle. Similarly, the essence of this recommendation is that value-added models should avail themselves of the basic strength of statistical analysis—i.e., statistical control—and take into account the ease or difficulty of the individual cases teacher handle. Many models in current use, including those used by the Measures of Effective Teaching (MET) project funded by the Gates Foundation, do not account for student assignment and thus lose some of the model's capability to isolate teacher contributions from other factors that may influence a student's achievement. From a practical standpoint, incorporating teacher fixed effects is relatively easy to do.

2) **Models should control for prior student achievement and not use a gain score as a dependent variable**
    - This recommendation again stems from basic statistical principles—gain-score models miss the opportunity to account for student assignment-related factors and therefore produced biased results. As we show in the actual data in our second paper described above, gain score models can produce quite different estimates of teacher performance than those produced by stronger models such as the Dynamic OLS approach in cases where teachers teach in schools that "track" students by grouping them in classrooms with other students of similar achievement levels.

3) **Value-added models are preferable to growth models**
    - Growth models, such as the Colorado Growth Model, describe the collective amount of growth that occurs in a group of students—in this case, in a teacher's classroom. As such, they merely describe this growth but do not make use of basic statistical techniques that allow us to isolate the teacher's contribution to that growth from that of other possible sources of learning.

4) **"Shrinkage" procedures are of limited usefulness**
    - The term "shrinkage" is often used in connection with a variety of different types of value-added models and appears as a component of many widely used approaches. However, as more students contribute to a teacher's effect estimate (for example, by using more than one year of data for teachers), the results of some methods that use shrinkage become indistinguishable from those obtained using a more robust value-added approach.

The project will formulate additional recommendations in the coming months. For example, we are currently working on the issues of measurement error in test scores and whether or not the typical measurement error corrections currently in use (such as those employed in the Wisconsin Value-Added Research Center models) are helpful in this regard.

To be sure, even the best value-added models have limitations, and a degree of imprecision exists with every approach. Good teaching is an art that can never fully be captured by a number. Teacher performance measures computed from student test scores should always be combined with other types of performance measures, such as classroom observations, in evaluating teachers. An outcome-based value-added measure, however, can form an integral part of an overall picture of a teacher's performance and, if computed thoughtfully, provide an important piece of information. Given that states and districts are computing these measures, it is our hope that they will move in the direction of adopting best practices in choosing their methods.

Appendix: Bios of Principal Investigators

Dr. Cassandra Guarino, Principal Investigator and Project Director, is an associate professor of Educational Leadership and Policy Studies at Indiana University Bloomington and an economist and experienced policy analyst with a background in teacher quality, teacher labor markets, and program evaluation. She has successfully led the current IES study of value-added modeling and is an author on several current studies of value-added methods. In the past, she has led and published several studies of teacher labor markets, school choice, and studies of early elementary teacher effectiveness, such as that conducted for the National Center for Education Statistics, and has served on the editorial board for *Educational Evaluation and Policy Analysis*. Formerly a faculty member at Michigan State University and before that an economist at the Rand Corporation, she has extensive experience leading and participating in research projects of all sizes.

Dr. Mark Reckase, Principal Investigator, a Michigan State University Distinguished Professor, is a past president of the National Council on Measurement in Education and the leader in the field on measuring the types of construct shift that cause bias in VAMs. He specializes in the development of educational and psychological tests, educational policy related to testing, and the psychometric theory that supports the assessment of cognitive skills and content knowledge. In particular, he conducts research on applications of unidimensional and multidimensional item response theory (IRT) models, computerized adaptive testing (CAT), assessment using performance tasks, and standard setting on educational tests. He has authored a book on multidimensional item response theory published by Springer in 2009. He has also served as a member of the Technical Advisory Committee for the assessment system in Florida.

Dr. Jeffrey Wooldridge, **Principal Investigator**, a Michigan State University Distinguished Professor, is a world-renowned expert on longitudinal data with nested structures. He is the author of *Econometric Analysis of Cross Section and Panel Data* (MIT Press, 2010), the leading authoritative text on the subject. He is a past president of the Midwest Economics Association and has served as editor for the *Journal of Econometric Methods, Journal of Economics and Business Statistics, Econometric Theory,* and *Economics Letters* and on the editorial board for the *Journal of Economic Literature*. He has developed tests of assumptions—such as geometric decay and feedback from changes in outcomes to future assignments—that are centrally applicable to VAM settings.