Conference Papers

Authors: Douglas Harris, Andrew A. Anderson
Paper Title: *Bias of Public Sector Worker Performance Monitoring: Theory and Empirical Evidence from Middle School Teachers*

Abstract: Monitoring workers in the public sector services presents distinctive challenges because output is multidimensional and preferences for each dimension vary across clients, but for legal and political reasons performance measures often have to be standardized and cannot be based on market prices. We show that this creates three types of bias and identify these biases by applying "value-added" panel data techniques to a statewide sample of middle schools where students and teachers are assigned to explicit tracks. Omitting course tracks, as is typical in actual high-stakes teacher value-added estimates leads to arguably large biases: 30-70 percent of teachers are placed in the wrong performance quartile. In addition, we are able to decompose the sources of bias to better understand the production function. We estimate that 75-95 percent of the bias is due to student sorting, which is difficult to avoid even by incurring additional monitoring costs. The remaining two sources of bias work in opposite directions and thus partly cancel out: first, the test score metric does not equally capture output in all tracks; second, student preferences for outcomes not intended to be measured by the test also vary by track. This has implications for monitoring teachers and other public sector workers.

Authors: Dan Goldhaber and Duncan Chaplin
Paper Title: *Assessing the "Rothstein Test," Does It Really Show Teacher Value-Added Models Are Biased?*

Abstract: In his influential paper, Jesse Rothstein (2010) finds standard value-added models (VAMs) suggest implausible and large future teacher effects on past student achievement. This is the basis of a falsification test that appears to indicate bias in VAM estimates of current teacher contributions to student learning. This paper argues the Rothstein test does show that students are tracked to teachers, but the tracking could be based on lagged achievement. Our results indicate that the Rothstein test does not appear to provide additional guidance regarding the efficacy of VAMs, suggesting a more encouraging picture for using VAMs for policy purposes.

Authors: Jeffrey Wooldridge, Cassandra Guarino, Mark Reckase, and Brian Stacy
Paper Title: *Evaluating Specification Tests in the Context of Value-Added Estimation*

Abstract: We study the properties of two specification tests that have been applied to a variety of estimators in the context of value-added measures (VAMs) of teacher and school quality: the Hausman test for choosing between random and fixed effects and a test for feedback (sometimes called a "falsification test"). We discuss theoretical properties of the tests to serve as background. An extensive simulation study provides important provides further insight to the VAM setting. Unfortunately, while both the Hausman and feedback tests have good power for detecting the kinds of nonrandom assignment that can invalidate VAM estimates, they also reject in situations where estimated VAMs perform very well. Consequently, the tests must be used with caution when student tracking is used to form classrooms.

Authors: John Engberg, Juan Saavedra, Jennifer Steele, Gema Zamarro
Paper Title: *Disentangling Disadvantage: Can We Distinguish Good Teaching from Classroom Composition?*

Abstract: This paper focuses on the use of teacher value-added estimates to assess the distribution of effective teaching across students of varying socioeconomic disadvantage. We use simulation methods to examine the extent to which different commonly used teacher-value added estimators accurately capture the distribution of effective teaching as we vary our model specifications as well as our assumptions about student sorting and classroom composition effects. With the exception of an

aggregated residuals model, most of the models, including teacher fixed and random effects models and a student fixed effects model, are able to recover unbiased estimates of a distribution parameter that measures the relationship between teacher effectiveness and student socioeconomic disadvantage, even under extreme sorting assumptions. The bias in estimates of the distribution parameter from the aggregated residual model increases with the degree of student sorting and with the inclusion of classroom-level controls. The findings may be of interest to researchers or policy leaders who seek to understand how teacher effectiveness is distributed with regard to student demographics in a particular school district, state, or charter management organization.

Authors: Matthew Johnson, Stephen Lipscomb, and Brian Gill
Paper Title: *Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables*

Abstract: The validity of value-added models (VAMs) of teacher effectiveness depends on the ability of the measures to isolate teachers' contributions to their students' achievement growth. Existing VAMs differ in key aspects of their empirical specifications, however, leaving policymakers with little clear guidance on what factors are important to include when constructing a fair model. We examine the sensitivity of teacher value-added estimates obtained under different model specifications. Our models differ based on whether student-level background characteristics, peer-level background characteristics, a double-lagged achievement score, and/or interactions between prior achievement scores and demographic variables are included. Using data from a northern state and a medium-sized, urban district in that state between 2008-09 and 2010-11, we find that teacher estimates are, in general, highly correlated across model specifications. The lowest correlation we observe—which still exceeds 0.91—is between (1) a model that includes one year of lagged scores with student and peer background characteristics, and (2) a model that includes two years of lagged scores and no student or peer background characteristics. Despite high correlations for teachers overall, the specifics of VAM specifications can affect the placement of some teachers across performance categories. The most-affected teachers serve classrooms of students that differ substantially relative to the state in terms of characteristics that are being included or excluded across VAM specifications (e.g., low-income status). To the extent that these factors are common to teachers in a district, the state's choice of VAM specification may systematically affect performance estimates for teachers in a district within the state.

Authors: Elias Walsh and Eric Isenberg
Paper Title: *How Does a Value-Added Model Compare to the Colorado Growth Model?*

Abstract: We compare teacher effectiveness results from a typical value-added model with results from the Colorado Growth Model (CGM), a student-level model that compares the percentile of each student's achievement to other students with similar past test scores. The CGM assigns a growth percentile to each student, and the median relative percentile of a teacher's students provides the measure of teacher effectiveness. The CGM does not account for other student background characteristics. This may lend the CGM more transparency for educators than a value-added model, but also could result in teachers with many reduced-price or free lunch, English language learner, or special education students being unfairly disadvantaged. The Colorado Growth Model (CGM) is approved by the U.S. Department of Education for use in AYP determinations, and it is being used in 15 states. Using data from the DC Public Schools, including teacher-student links that have undergone a roster confirmation process by teachers and principals, we look at the stability of results across the two methods for all teachers and for teacher subgroups. In particular, we compare differences between the benchmark and CGM value-added for teachers with few disadvantaged students relative to teachers with many disadvantaged students. If the results are similar, the perceived face validity benefit of the CGM may make it an appealing alternative to the benchmark value-added model.

Authors: Cassandra Guarino, Mark Reckase, Brian Stacy, and Jeffrey Wooldridge
Paper Title: *A Comparison of Growth Percentile and Value-Added Models of Teacher Performance*

Abstract: Currently researchers and policymakers can choose among a number of statistical approaches to measuring teacher effectiveness based on student test scores. Given a relative lack of easily

accessible information on the pros and cons of different methodological choices, the choice of a method is often based on replicating what others in similar contexts or disciplines have done rather than carefully weighing the relative merits of each approach. The distinction between growth modeling procedures and OLS-based value-added models in the context of teacher performance evaluation---and the relative merits of each approach---has not been fully explored. This paper addresses this task. To explore this research question, we evaluate the merits of growth models versus VAMs with regard to the goal of ranking teachers, since both approaches can accomplish this task. Both types of approaches face a common set of challenges when applied to the task of determining teacher effectiveness rankings. Perhaps the most important of these is the issue of bias under conditions of nonrandom assignment of students to teachers. To compare how well the two approaches deal with these challenges, we use them to rank teachers using simulated data in which the true underlying effects are known. The simulated datasets are created to represent varying degrees of challenge to the estimation process: some of our data generating processes (DGPs) randomly assign students to teachers, others do so in nonrandom ways. In addition to the simulation study, we compare growth percentile models to VAMs using administrative data from a large diverse southern state.

Authors: Bing-ru Teh, Elias Walsh, and Eric Isenberg
Paper Title: *Is It Better To Estimate Value Added Using Elementary School Teachers?*

Abstract: It is often presumed that upper elementary school grades 4 and 5 provide better data for estimating teacher value added for research purposes than middle school grades 6 to 8 for two reasons: a) elementary school teachers teach two subjects in individual, self-contained classrooms, and b) there is little tracking of students to teachers in elementary school grades compared to middle school grades. We examine both assumptions. First, to document whether the first assumption holds in practice, we use data on teacher-student links from DC Public Schools that have undergone a roster confirmation process, whereby teachers and principals verify which subjects and students a particular teacher taught and for what length of time. We also compare the teacher-student links that result from this process to data that approximate the typical quality of teacher-student links in unconfirmed administrative data. Second, we compare variation in baseline student achievement within and between classes at the same school at upper elementary and middle school grades to examine whether tracking of students by ability grouping is actually more common at the middle school level compared to the upper elementary school level. Preliminary results show that departmentalization of teaching instruction across math and English/Language Arts is actually quite common in grades 4-5, at least in DCPS, where about one in six teachers of these subjects is linked to a subject in the administrative data that the teacher does not teach. Put another way, between one third and one half of DCPS upper elementary school students are taught by a departmentalized teacher. As an example of how using unconfirmed administrative data can affect results, we examine how calculations of the year-to-year and cross-subject stability of value-added estimates depend on the quality of the data used.

Authors: Eric Isenberg and Elias Walsh
Paper Title: *Accounting for Co-Teaching with the Full Roster and Full Roster Plus Methods*

Abstract: Building on earlier work, we propose an improvement to a method of accounting for co-teaching that treats co-teachers as teams, with each teacher receiving equal credit for co-taught students. Hock and Isenberg (2012) described a method known as the Full Roster Method (FRM) that links students to each of their teachers. It involves creating unique records for each teacher-student combination in cases where students are co-taught by two or more teachers. In the regression analysis, each teacher-student combination is weighted according to the fraction of the year the student spent with the teacher. While providing a feasible and simple solution to estimating effects for teacher teams, the FRM effectively double counts co-taught students—these students receive a full weight with each of their teachers, but this causes these students to receive extra weight when calculating the relationship between student characteristics and achievement. The improvement, known as the Full Roster Plus Method, continues to allow co-taught students to receive full weight with their teachers, but also makes all students contribute equally to the calculation of the relationship between student characteristics and achievement. This method may be particularly appealing when estimating value added across school districts, each of which has its own set of practices and data systems for instituting and measuring co-

teaching. To investigate how the application of this method empirically changes value-added estimates, we use data from DC Public Schools, which uses a roster confirmation process that allows teachers to verify which of the students listed on their administrative rosters they actually taught. The roster confirmation process has uncovered a substantial level of co-teaching. For example, in the 2010-2011 school year, 29 percent of math teachers and 40 percent of reading teachers shared students with another teacher receiving a value-added score, and 9 percent of math teachers and 13 percent of reading teachers shared all of their students with another teacher.

Authors: Mark Ehlert, Cory Koedel, Eric Parsons, Michael Podgursky, and Peng Xiang
Paper Title: *Incorporating End-of-Course Exam Timing into Educational Performance Evaluations*

Abstract: There is increased policy interest in extending the accountability framework in K-12 education to include student achievement in high school. High-school achievement is typically measured by performance on end-of-course exams (EOCs), which test course-specific standards in subjects including algebra, biology, English, geometry, and history, among others. However, unlike standardized testing regimes in the early grades, students take EOCs at different points in their schooling careers. Recent research indicates that school and district policies that determine when students take particular courses can have important implications for student achievement and longer-term outcomes, like advanced course taking. In this paper we propose an approach for modeling EOC test performance that disentangles the influence of school and district policies regarding the timing of course taking from other factors. After separating out the timing issue, better measures of the quality of instruction provided by districts, schools and teachers can be obtained. Our approach also offers diagnostic value because it explicitly separates out the influence of school and district course-taking policies from other factors that determine the total performance evaluation.

Authors: Cassandra Guarino, Eun Hye Ham, Mark Reckase, Brian Stacy, and Jeffrey Wooldridge
Paper Title: *Sending Value-Added into Tailspin: Measurement Error in Models of Teacher Performance*

Abstract: As a result of federal efforts to hold individual teachers accountable for student learning based on the standardized test results of their students (Race to the Top, U.S. Department of Education, 2009), the use of value-added models of teacher performance as a component of teacher evaluation has been spreading rapidly. Value-added models have come under strong criticism for their imprecision and potential for bias (e.g., Rothstein, 2008; Baker et al., 2010). However, a working paper by Kane and Staiger (2008) and a recent one by Chetty, Friedman, and Rockoff (2011) have largely assuaged fears about bias in these models. Given these findings and the pressures to incorporate test-based performance measures into teacher evaluation, policymakers have forged ahead in their efforts to implement and attach stakes to these measures. Often researchers and policymakers alike feel comfortable with imprecision in the models because they are usually employed primarily to identify teachers in the *high* and *low* ends of the distribution in effectiveness. Using these models to reward teachers at the top and sanction those at the bottom seems a relatively attainable goal. However, the question of whether we can "get the tails of the distribution right" has not been adequately resolved in the research literature. Related to this is the issue of measurement error in student test scores. Most models either do not address the possibility that measurement error could be affecting value-added estimates in important ways or employ measurement error correction techniques derived from other applications to deal with the problem. In this paper, we address the issue of measurement error and its potential to introduce bias in the tails of the effectiveness distribution and sound a warning note that identifying teachers at the top and bottom of the scale may be less straightforward than previously thought. We produce simulation evidence on the performance of various value-added estimators in the presence and absence of measurement error and show that under certain conditions measurement error can induce a visible bias in estimation. Included in the set of estimators that we consider are estimators that explicitly "correct" for this problem. Moreover, we find that estimators found in the literature that attempt to correct for measurement error are shown to be largely unbiased when assignment is based on true scores but biased when assignment is based on observed scores. These correction techniques may exacerbate rather than mitigate the problem of diagnosing teachers in the tails of the distribution.