On October 10-11, researchers Cassandra Guarino, Mark Reckase, and Jeffrey Wooldridge hosted an IES sponsored conference at Michigan State University on the timely topic of "Using Student Test Scores to Measure Teacher Performance: The State of the Art in Research and Practice."

The conference brought together more than 80 researchers, policy-makers, and practitioners from across the U.S. to discuss how to promote best practices in constructing and implementing value-added and growth measures of teacher performance.

The primary goal of the conference was to make a positive impact on policy by (1) raising awareness regarding the strengths and weaknesses of different methods of computing teacher performance measures and (2) discussing how to disseminate this information in a way that will enable school systems to make informed choices as they implement teacher evaluation systems.

The conference occurred at a critical point in time. 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 to help categorize teachers as effective or ineffective.

The market for assessments coupled with derivative products that compute teacher effectiveness measures—each promoting a particular methodology—is becoming increasingly competitive. Simultaneously, a large number of research studies have come to light regarding strengths and limitations of different methodological choices.

With the implementation of the Common Core State Standards in the majority of states and the development of specific sets of assessments that align with them, it is particularly urgent to open up the field right now to talk about the best way to compute teacher performance measures based on the evidence that has been compiled.

The conference specifically addressed pressing issues related to choosing and implementing particular methodologies. Panelists from the National Center for Assessment, AIR, the North Carolina Department of Public Instruction, the Value-Added Research Center, the American Federation of Teachers, and the Council of Chief State School Officers highlighted the following issues:

- The focus on school accountability has shifted the theory of action over to classroom instruction, with the idea that great teachers are the primary instruments to close the achievement gap.
- Credibility and communication are key. Stakeholders are looking for transparency and a language for talking about teacher performance models that they can understand. Moreover, there are many methodological and logistical choices to be made in constructing teacher performance measures, thus it is important to be able to explain the nuances in the system.
- Value-added models are viewed as highly controversial, and it is preferable not to label methods using that terminology.

- Decisions about which students and teachers to be included in the calculations are important and not simple. It is difficult to include highly mobile students, and, currently, fewer than half of teachers can be assessed with value-added or growth models due to the subjects and grades tested.
- Preparation is an important feature of successful implementation, given the fast timeline states and districts have to implement teacher evaluation systems. It is advisable to set up the infrastructure in advance and do dry runs with simulated data.

Presentations by prominent researchers from several universities and research institutes focused on six main topics: (1) model specification tests, (2) comparisons of the growth and value-added models, (3) the influence of tracks and course timing on teacher performance measures, (4) whether or not specifications should include peer variables, (5) how different methods of linking students to teachers affect value-added, and (6) the impact of measurement error on value-added. The research presented suggested the following:

- Model specification tests can be misleading. Although the results of such tests, such as the "falsification" tests discussed in Rothstein (2010), have been cited as evidence to discredit value-added models, they in fact tell us very little about whether or not particular models produce good estimates of teacher effectiveness. A value-added model can fail these tests and yet do a good job of identifying teacher effects.
- Comparisons of the Colorado Growth Model (CGM) with various value-added models show that the CGM does as well as any value-added specification under conditions in which students are randomly assigned to teachers. However, value-added models that include both prior test scores and teacher variables as fixed regressors are better at measuring a teacher's effectiveness in the many cases in which students are tracked into classrooms on the basis of prior performance and assigned to teachers in a nonrandom fashion.
- Assessing teacher value-added is more difficult in middle and high school than in elementary school due to the prevalence of tracking and course-sequencing. Accounting for current and prior tracks is important in estimating teacher performance to avoid bias. Using End of Course (EOC) exams is an alternative way to incorporate student performance measures into teacher evaluation but there are issues to be dealt with regarding course timing and the fact that not all students take EOC exams for all courses.
- The inclusion of classroom covariates in value-added models can substantially affect how teachers are classified. Moreover, it can be difficult to disentangle the teacher effect from the peer effect when students are nonrandomly assigned to teachers. Measures based on multiple classrooms and years of data can help alleviate this issue somewhat.
- Difficulties in properly linking teachers to students can stem from inaccuracies in class rosters and team teaching, especially at the elementary level. Although a few districts have undertaken efforts to confirm class rosters, many districts link large proportions of students to the wrong teachers. Roster-confirmed data and estimators that assign dosages to teachers can help account for these issues.
- Measurement error in test scores is not properly addressed by commonly applied correction techniques that rely on classical error assumptions.

For more information about the conference, please go to: http://vam.educ.msu.edu/. The website contains links to videos of the conference, conference slide presentations, and papers to provide information to policy makers, researchers, and educators.

Cassandra Guarino