Using Value-Added Measures to Assess Teacher and School Performance:

US Lessons for Developing Countries

The use of student test scores to evaluate school and teacher performance is growing in the United States. One common method, called Value-Added-Models, attempts to adjust for factors outside of the control of educators by controlling for past student performance and other observable characteristics of students (Koedel et al 2015). Value-Added models are not common in developing countries but their use is growing (Muralidharan and Sundararam 2011; Araujo, Carneiro, Cruz-Aguayo, and Schady 2014; Azam and Kingdon 2013; De Ree 2014; Bruns et al 2014). In this presentation I discuss potential lessons learned from the US with a particular focus on potential bias in value-added measures (Goldhaber and Chaplin 2015) and the apparent quality of common alternatives (Chaplin et al, 2014; Lipscomb et al, 2015).

Value-added measures are being used in the United States by states, districts, and schools (Lipscomb et al, 2015; Chaplin et al 2014; Walsh and Eisenberg 2015). Common alternatives are average test scores, especially for schools, and classroom observations, especially for teachers. Compared to value-added-models, average test scores probably produce more biased estimates of school effectiveness because they do not adjust for factors beyond the control of the schools. Classroom observations may also be problematic if they are not well implemented (Lipscomb et al, 2015; Chaplin et al 2014).

Theoretical and simulation results suggest that bias in VAM is not substantial (Goldhaber and Chaplin 2015). Similar work by Chetty et al 2014 provides empirical support for this conclusion. However, single-year VAM can be quite noisy (Schochet and Chiang 2010). This noise can be reduced by using multiple years of VAM which may be an option in some developing countries at the school level. The noise can also be reduced by combining VAM with other measures. This, in turn, can work if the other measures are well implemented (Kane et al, 2010). However, those other measures are sometimes lacking in practice (Lipscomb et al, 2015; Chaplin et al, 2014).

In the rest of this presentation I cover details of my own work in this area focusing on the potential for bias in VAM (Goldhaber and Chaplin 20150 and how VAM measures relate to other measures of student performance (Chaplin et al 2014; Lipscomb et al 2015).

Falsification Tests and Value-Added Models (Goldhaber and Chaplin, 2015)

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 typical VAM estimates of current teacher contributions to student learning. Rothstein’s critique of value-added methods used to estimate teacher effectiveness has been cited by both researchers and policymakers as a reason to doubt the wisdom of using VAMs for high-stakes purposes.

In this paper the Rothstein test is investigated using both theoretical considerations and simulation methods. The findings call into question whether the Rothstein falsification test (RFT) provides accurate guidance regarding the magnitude or existence of bias of teacher effect estimates. Ideally, the RFT could be used to identify VAMs that produce biased estimates of current teacher effects. It is shown, however, that the RFT identifies important control variables left out of a VAM only under conditions that are not plausible. More precisely, we find that one cannot use the RFT to reject the hypothesis that students were effectively randomly assigned conditional on lagged achievement. In addition, we find that when data are generated that appear similar to the data analyzed in Rothstein (2010), estimated future teacher effects, from his tests, are similar in magnitude to the true teacher effects, but the bias for current teacher effects is extremely small, suggesting that the magnitude of the future teacher effects does not provide useful information about the magnitude of the bias. In a nutshell, the RFT can be used to identify the existence of tracking, but the tracking could well be a function of lagged achievement, the variable that is included in most VAMs. It does not appear that the RFT can be used to tell us much more.

The authors think that Rothstein's 2010 paper raised important concerns about the ability of VAMs to produce unbiased estimates of teacher effectiveness, but the RFT itself does not provide useful guidance regarding VAMs. Given this, more work needs to be done to understand the potential reasons why VAMs might produce biased teacher effect estimates. This will likely involve a closer look at the various factors affecting student sorting into classrooms so that one can better account for student sorting when estimating teacher effects.

From a policy perspective, the important question may not be whether there is any bias, but the potential magnitude of any bias. It is quite likely that teacher effectiveness estimates generated from VAMs are biased, at least to some small degree but, as shown in Rothstein (2009), Kinsler (2012), and our simulations, the magnitude of bias may be inconsequential. Decisions about using VAMs should consider how this bias compares to potential information that value-added models can provide about teacher effectiveness over, or in addition to, other means of assessment.

City of Pittsburgh Teacher Performance Measures (Chaplin et al, 2014)

Responding to federal and state prompting, school districts across the country are implementing new teacher evaluation systems that aim to increase the rigor of evaluation ratings, better differentiate effective teaching, and support personnel and staff development initiatives that promote teacher effectiveness and ultimately improve student achievement. States and districts are implementing richer measures of professional practice alongside "value-added" measures of student achievement growth and in some cases are incorporating additional measures, such as student surveys. Pittsburgh is a leader in the nationwide movement to evaluate, enhance, and reward effective teaching. The analyses presented in this report were conducted to assist Pittsburgh Public Schools in refining its multiple measures of teacher effectiveness, to create a rich, valid, and comprehensive combined measure. In addition, Pittsburgh's work is based on an approach that is being used or considered elsewhere, the findings have important implications for districts and states across the country.

The Pittsburgh Public Schools teacher evaluation system includes three types of measures. The first—the Research-based Inclusive System of Evaluation (RISE), based on Charlotte Danielson's Framework for Teaching (Danielson, 2013)—is an observation-based professional practice measure that relies on principals' assessments. The second measure is based on a student survey called the 7Cs, which incorporates students' perceptions of teachers' practices and was developed by Ronald Ferguson of Harvard University as part of the Tripod Project and administered by Cambridge Education. The third measure is a value- added measure that uses changes in student test scores to estimate each teacher's contribution to student achievement over up to three years of teaching. This study used 2011/12 data to describe how the ratings on the three measures are distributed across teachers and how the ratings are correlated.

We find that all three measures have the potential to differentiate among teachers. While all three composite measures suggest a wide range of teacher effectiveness, only the district's value-added measures have been shown to reliably differentiate among teachers (Johnson et al., 2012); the reliability of the RISE and 7Cs composites cannot be determined without multiple ratings per teacher (ideally by multiple raters). However, the components of each of these measures are highly correlated, indicating that the composites have acceptable levels of internal consistency.

We find that teachers with high RISE ratings tend to have high 7Cs ratings and high value-added measure estimates as well. The correlations are moderate but statistically significant—consistent with other research on similar measures of teacher effectiveness. These results suggest that the measures capture teaching skills that overlap but are not identical—as the district intended in creating multiple measures.

We find that systematic differences in RISE ratings remain between school even after accounting for differences in value- added and 7Cs measures, suggesting that some principals are tougher or more lenient than others in applying RISE. Using an additional rater for each teacher could help principals better calibrate their RISE ratings, thus enhancing the consistency, fairness, and validity of the ratings (particularly if the additional raters work in more than one school).

State of Pennsylvania Teacher Performance Measures (Lipscomb et al, 2015)

Pennsylvania bases annual teacher evaluations on several measures, including supervisor observations using the Framework for Teaching (FFT) and, for many teachers, their contributions to student achievement growth from a value-added model (VAM). In the past, there was concern that supervisor observations did not differentiate performance well or relate to true teacher performance. We investigate how well the current system addresses these issues using evaluation data covering 6,676 teachers from 269 districts. We find that, although FFT scores are overwhelmingly concentrated in the top few performance categories, the positive correlations with VAM suggest that the FFT provides some meaningful differentiation and captures aspects of teacher skills related to student achievement growth.