Title:
Is Seeing Believing? Exploring the Empirical Evidence for the Validity of Teacher Observations

Session Chair:
Mark Chin, Harvard Graduate School of Education (mark_chin@gse.harvard.edu)

Authors and Affiliations:
Andrew Bacher-Hicks, Harvard Kennedy School of Government (abacherhicks@g.harvard.edu)
David Blazar, Harvard Graduate School of Education (db239@mail.harvard.edu)
David Braslow, Harvard Graduate School of Education (dab736@mail.harvard.edu)
Charalambos Y. Charalambous, University of Cyprus (charalambous.y.charalambous1@ucy.ac.cy)
Mark Chin, Harvard Graduate School of Education (mark_chin@gse.harvard.edu)
Dan Goldhaber, CALDER, American Institutes for Research (dgoldhab@u.washington.edu)
Thomas J. Kane, Harvard Graduate School of Education (tom_kane@gse.harvard.edu)
Douglas O. Staiger, Dartmouth College (douglas.o.staiger@dartmouth.edu)

Conference Section:
First Choice – Instruction and Student Achievement
Second Choice – Understanding the Effects of Education Policies

Paper Presentation Order:
1) Dimensionality of Upper Elementary Mathematics Instruction: Exploring Factors Across Two Observational Instruments
2) Validating Components of Teacher Effectiveness: A Random Assignment Study of Value-Added, Observation, and Survey Scores
3) Impacts of Multidimensionality and Error: Simulating Explanations for Weak Correlations between Measures of Teacher Quality

Discussant:
Heather C. Hill (heather_hill@gse.harvard.edu)
For decades, academics have considered observations of teacher instruction as indicators of teacher quality (Brophy & Good, 1986), and administrators have used observations to assess teachers in schools. More recently, observation instruments have been designed to capture an array of practices considered important to instruction. For example, content-independent instruments, such as the Classroom Assessment Scoring System, assess teacher skills in classroom organization and in the creation of supportive learning environments. Instruments like the Mathematical Quality of Instruction capture a teacher’s ability to use reform-based practices to deliver subject-specific content. Both instruments are used for evaluation purposes and development efforts.

To date, much research has explored the validity of newer observation instruments in representing teacher quality. For example, empirical findings suggest that teacher scores on these rubrics can, under certain scoring conditions (i.e., multiple evaluators assessing multiple instances of instruction), demonstrate acceptable levels of reliability (Bell et al., 2012; Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013). Other studies have found evidence for the convergent validity of such measures, with teachers’ content knowledge and knowledge for teaching demonstrating strong relationships to teachers’ content-specific instructional practices (Hill, Kapitula, & Umland, 2010; Hill, Blazar, & Lynch, submitted).

Some issues remain, however, in the validation of observation scores as metrics of teacher quality. First, few empirical studies have investigated how multiple observation instruments relate to one another, and how they jointly relate to the underlying trait of instructional quality. Specifically, little research provides insight as to whether teacher scores across general and subject-specific instruments derive from single or multiple constructs. Second, studies have largely found teacher observation scores to weakly relate to teachers’ impacts on student achievement (most commonly known as “value-added”; Kane & Staiger, 2012). This finding challenges either the convergent validity of observation scores or value-added, as both measures supposedly represent the same latent trait. Intuition would suggest that classroom instruction should correlate with student achievement; it is possible, however, that both measure different aspects of teacher ability.

Our symposium attempts to address these issues concerning the validity of observation scores as metrics of teacher quality. Analyzing 1300 videos of upper-elementary math lessons, the first paper investigates teacher performance on both a general and a subject-specific observation instrument to better understand the dimensionality of teacher instruction. The second paper uses a dataset in which classrooms are randomly assigned to teachers to explore whether non-experimental estimates of quality, including observation scores, predict student outcomes without bias. Finally, using simulated data, our third paper explores explanations for weak correlations between value-added and observation scores in the literature, such as error in scores and multidimensionality in teacher quality.

Our symposium provides further evidence regarding the relationship of different instructional practices and student achievement, and suggests to policymakers areas of focus for teacher evaluation and professional development. Such knowledge should subsequently provide tools to
improve the long-term outcomes of students, such as college attendance and career earnings (Chetty, Friedman, & Rockoff, 2011).

References


