Education reform is increasingly focused on teachers and teaching with current policy initiatives (RTTT, NCLB waivers, etc) promoting evaluating teacher in terms of their relative effectiveness. This emphasis on quantitative approaches to rating teachers grew out of a deep frustration with teacher evaluations in which almost all teachers receive the same evaluation, and unsatisfactory evaluations are rare. In response, states and districts across the country are designing and implementing new teaching evaluation systems which have at their heart two classes of measures—one based on student achievement and the other based on classroom observation.

The revised evaluation systems rely on a new generation of standardized classroom observation protocols that can generally be described as follows: Each protocol is structured around a set of domains that describe the core constructs of teaching valued by the protocol (e.g., instructional support, classroom organization, emotional support). Each domain is defined by a set of dimensions (e.g., engagement, productivity). Each dimension is assigned scores by a trained observer on a numerical scale that typically has descriptive anchors explaining most, if not all, of the scale points. Observations of the teacher are made for some specified number of lessons over the school year. For each observation, observers assign scores using the method mandated by the protocol system. Dimension scores are subsequently aggregated to obtain an overall observation score, or set of domain scores, for the teacher for the year.

There is a pressing need to determine if the inferences made from this new wave of protocols will be accurate and lead to the appropriate conclusions about teachers and guidance on how they can improve. A few studies have calculated at the reliability of these measures (cite MET, Mashburn) but no studies have addressed issues such as what constructs the protocol scores measures, whether observing video will yield the same inferences as live observations, or what contributes to the large variability among ratings observed in the reliability studies. In addition, to date, studies have not used measurement models that fully account for the complex hierarchical multivariate ordinal nature of the protocol scores which are collected from one or more independent raters observing multiple lessons taught to multiple classrooms by each teacher.

This symposium includes four papers that address the methodological challenges of collecting and modeling classroom observation scores. The first two papers identify the constructs measured by the Classroom Assessment Scoring System Secondary, CLASS-S. The first paper develops a Bayesian hierarchical multivariate ordinal model for the observed CLASS-S scores. The second paper applies the model to CLASS-S scores from two studies of secondary teaching – Understanding Teaching Quality (UTQ) and Toward and Understanding Classroom Context (TUCC). The third paper addresses how best to conduct observations by studying difference in scores from raters observing lessons live in the classroom with raters observing videos of the lessons. The final paper studies changes in scores as raters become more experienced with four different observation protocol used in the UTQ study.
Papers

Bayesian Hierarchical Multivariate Formulation with Factor Analysis for Nested Ordinal Data
Terrance D. Savitsky, RAND Corporation

Analyzing the Factorial Structure of the Classroom Assessment Scoring System-Secondary Using a Bayesian Hierarchical Multivariate Ordinal Model
Kun Yuan, Daniel F. McCaffrey, and Terrance D. Savitsky, RAND Corporation

Effect of Observation Mode on Measures of Teaching
Daniel F. McCaffrey and Jodi M. Casabianca, RAND Corporation and Carnegie Mellon University

Rater Drift and Time Trends in Classroom Observations
Jodi M. Casabianca and J. R. Lockwood, Carnegie Mellon University and RAND Corporation

Discussant
J. Patrick Meyer, University of Virginia