# Abstract Title Page

**Title:** Latent class models for teacher observation data

**Authors and Affiliations:** Peter F. Halpin, New York University

## Abstract Body

### Background / Context:

Recent research on multiple measures of teaching effectiveness has redefined the role of in-classroom observations in teacher evaluation systems. In particular, most states now mandate that teachers are observed on multiple occasions during the school year, and it is increasingly common that multiple raters are utilized across the different rating occasions (White, 2014). Teacher observations also continue to make up the majority of weight in many districts' evaluation systems. For example, the NYC DOE current allocates 60% of the weight in teacher evaluations to between four and six observations that are made throughout the school year, with the remaining 40% split over value-added measures and other local criteria.

In-classroom observations are typically conducted using a rating rubric, with the Danielson Framework (Danielson, 2013) being one prominent example. Despite the growing evidence that rating rubrics can provide useful information about teaching practices (Kane & Staiger, 2012), it remains less clear how that information should be summarized to support consequential inferences about individual teachers. Researchers in the teacher evaluation literature have summarized the rubrics with a total score, which has led to g-theory studies of the reliability of these scores over multiple rating occasions and raters (Ho & Kane, 2013). However, related research has found that many rubrics measure multiple dimensions of instructional quality (Grossman et al., 2014; Kane & Staiger, 2012; Savitsky & McCaffrey, 2014), suggesting that teachers' practices are not well described in terms of a total score. Halpin & Kieffer (2015) argued for the use of latent class analysis (LCA) as a means of capturing the multidimensional features of rating rubrics, while also providing the standard error of measurement for each teacher, and item-level diagnostic information that can be used as the basis of feedback to educators and for professional development. However, the analysis of Halpin & Kieffer was conducted using teacher-aggregate data and therefore did not allow for a model-based investigation of reliability over multiple rating occasions and raters.

### Purpose / Objective / Research Question / Focus of Study:

The main purpose of the present research is to develop a multilevel extension of the LCA methodology described by Halpin & Kieffer (2015). For a given rating rubric, the multilevel LCA approach is specifically intended to answer the following questions: (a) How reliably (precisely) is a teacher's teaching ability measured during any single observation session? (b) How consistently does a teacher perform over observation sessions? (c) For a given teacher, how many observation sessions are required before his/her teaching ability has been measured with a desired level of precision? The last question in particular has relevance for policy, in that multi-rater systems can place heavy financial demands on school districts in terms of deploying a sufficient number of trained raters to meet required number of observation sessions per teacher. The proposed methodology allows for decisions about the required number of observations to be made on a teacher-by-teacher basis, and to be informed by the data collected from each teacher.

An additional purpose of this research is to provide a satisfactory solution to the problem of rater reliability within the multilevel LCA framework. At the time of this writing, this is a continuing area of research. The specific topics to be addressed are (a) how to control for the effects of raters when making inferences about individual teachers, (b) requirements on how raters should be deployed to teachers in order for the rater effects to be identified, and (c) methods for inferring whether a rater is performing within expectation for a given population of teachers. Initial work is outlined in the description of the model below.

## Significance / Novelty of Study:

Teacher observations are taking an increasingly prominent role in teacher evaluation systems. Previous research on the reliability of teacher observations has employed a true-score model of teaching ability in combination with generalizability theory (Kane & Staiger, 2012; Ho & Kane, 2014). While leading to significant advances in the practice of teacher observations, this work can also be recognized as having a number of shortcomings. In particular, related research has made it apparent that many rating rubrics tend to be multidimensional (Grossman et al., 2014; Kane & Staiger, 2012; Savitsky & McCaffrey, 2014), and hence the utility of a true-score approach to reliability is brought into question. Halpin & Kieffer (2015) proposed an approach based on LCA that is compatible with the evidence that teachers' practices are multidimensional, but this did not address the reliability of scores over observation sessions. The present work fills this void, while working to address the role that raters play in generating the session-level data.

## Statistical, Measurement, or Econometric Model:

This section sketches the basic details of how multilevel LCA can be applied as a measurement model for teacher observations. Estimation of multilevel LCA via a multinomial regression specification using the EM algorithm is described by Vermunt (2003, 2004) and this can be applied to the present application. However, the proposed approach to rater effects requires embedding integration over raters on the E-step, which is accomplished using the method described by Hedeker (2003).

**The model.** Let $X_{ijk}$ denote the rating assigned to item $k$ of a rating rubric, observed during session $j$ of teacher $i$. It is assumed that each $X_{ijk}$ is random variable with support $x = \{x_r \mid r = 1, \ldots, R\}$. Let $X_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijK})$ represents the $K$-vector of ratings for the $ij$-th session and let $X_i = \{X_{i1}, X_{i2}, \ldots, X_{in_i}\}$ denote the collection of $n_i$ sessions for teacher $i = 1, \ldots, N$.

At the teacher level, introduce a discrete latent variable $Z_i$ with support $z = \{z_t \mid t = 1, \ldots, T\}$ and assume that the joint probability mass function (pmf) $P(X_i, Z_i)$ is well-defined. Then the pmf the observations of teacher $i$ is

$$P(X_i) = \sum_{z} P(X_i \mid Z_i = z_t)\, P(Z_i = z_t). \tag{1}$$

The basic idea behind the teacher-level model is to select the minimum value of $T$ such that

$$P(X_i \mid Z_i = z_t) = \prod_{j=1}^{n_i} P(X_{ij} \mid Z_i = z_t) \tag{2}$$

for all $i$. The interpretation of $Z_i$ is discussed. Any additional clustering of observation sessions within teachers (due to groups of students, subject taught, specific lessons, time of day, etc.) is considered uninteresting and addressed using established methods for model misspecification (e.g., cluster-robust standard errors).

Similar to the teacher level, introduce a discrete latent variable $Y_{ij}$ with support $y = \{y_s \mid s = 1, \ldots, S\}$ at the session level and assume that $P(X_{ij}, Y_{ij}, Z_i)$ is well-defined. Then

$$P(X_{ij} \mid Z_i = z_t) = \sum_y P(X_{ij} \mid Y_{ij} = y_s, Z_i = z_t)\, P(Y_{ij} = y_s \mid Z_i = z_t). \tag{3}$$

It is assumed that the measurement model is invariant with respect to $Z_i$,

$$P(X_{ij} \mid Y_{ij}, Z_i) = P(X_{ij} \mid Y_{ij}), \tag{4}$$

and that the item responses are independent given $Y_{ij}$,

$$P(X_{ij} \mid Y_{ij}) = \prod_k P(X_{ijk} \mid Y_{ij}). \tag{5}$$

These standard psychometric assumptions and their alternatives are discussed.

Substituting (2) through (5) into (1) gives the basic model.

$$P(X_i) = \sum_z \prod_j \sum_y \prod_k P(X_{ijk} \mid Y_{ij} = y_s)\, P(Y_{ij} = y_s \mid Z_i = z_t)\, P(Z_i = z_t). \tag{6}$$

**Classification of teachers.** In standard applications, posterior analysis is via $P(Z_i \mid X_i)$ and it is assumed that all observations $X_i$ arrive simultaneously. In professional settings, however, teachers are observed on different occasions sequentially throughout the school year. Letting smaller values of $j$ correspond to earlier observations and $X_{i\{j\}} = \{X_{i1}, X_{i2}, \ldots, X_{ij}\}$, this leads to the following Bayesian updating scheme:

$$P(Z_i \mid X_{i\{j\}}) \propto P(X_{ij} \mid Z_i)\, P(Z_i \mid X_{i\{j-1\}}) \tag{7}$$

In the example it is illustrated how this approach leads to reliable classification of some teachers using fewer observation sessions than for other teachers.

**Rater effects.** The choice of rater can influence the observed values recorded during an observation session. Consequently, raters can be conceptualized in terms of violations of measurement invariance:

$$P(X_{ijk} \mid Y_{ij}, W_{ij}) \neq P(X_{ijk} \mid Y_{ij}) \tag{8}$$

where $W_{ij}$ denotes the rater of session $j$ of teacher $i$. It is assumed that the choice of rater influences neither the type of teaching practice demonstrated during a session nor the type of teacher being observed. Since differences among raters are a nuisance feature, it is advantageous to consider the marginal probabilities

$$P(X_{ijk} \mid Y_{ij}) = \int P(X_{ijk} \mid Y_{ij}, u)\, f_W(u)\, d(u), \tag{9}$$

which bring us back to equation (5). Implication of this modeling strategy are discussed, including strategies for deployment of raters and how to evaluate bias among raters.

**Usefulness / Applicability of Method:**

The utility of the proposed method is illustrated with a secondary analysis of the Measures of Effective Teaching (MET) longitudinal database (www.metproject.org), focussing on middle school English language arts teachers in the first year of the study and using the PLATO rating rubric (Grossman et al., 2014). The accuracy with which teachers in the sample were classified is shown as a function of the number of observation sessions in Figure 1. The consistency with which teachers were classified is shown for a subset of nine teachers in Figure 2. These analyses are preliminary and further analyses are to follow.

**Conclusions**

Many school districts across the nation are taking seriously the call to multiple measures of teacher effectiveness, and central to these efforts is an increased utilization of multi-rater observation systems. This is a good example of the kind translation between knowledge and practice that is the theme of the present conference.

While multi-rater systems have been motivated by research about how to improve the reliability of scores obtained from rating rubrics, there still remains much work to be done in terms reporting the standard error of measurement when making consequential decisions about individual teachers' proficiency levels, using the information provided by instruments to make inferences about the specific strengths and weaknesses of individual teachers' practices, obtaining optimal strategies for rater deployment both in terms of the total number of observation sessions required per teacher and the identification of bias among raters, and doing all of this in a statistical framework that is compatible with well-established multidimensionality of teacher observation data, yet feasible to implement. The approach developed in this research addresses these issues and thereby facilitates the continued improvement of multi-rater systems as they are currently used in practice.

# Appendices

## Appendix A. References.

Danielson, C. (2013). *The framework for teaching evaluation instrument, 2013 instructionally focused edition* Princeton, NJ: The Danielson Group

Grossman, P. Cohen, J. Ronfeldt, M. & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher valued-added on multiple types of assessment. *Educational Researcher, 43,* 293-303.

Halpin, P. F. & Kieffer, M. J. (2015). Describing Profiles of Instructional Practice: A New Approach to Analyzing Classroom Observation Data. *Educational Researcher, 44,* 263-277.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, *22*(9), 1433–1446.

Ho, A. D., & Kane, T. J. (2012). *The reliability of classroom observations by school personnel.* Bill and Melinda Gates Foundation.

Kane T. J. & Staiger, D. O. (2012) *Gathering Feedback for Teaching: Combining high quality observations with student surveys and achievement gains*. Bill and Melinda Gates Foundation.

Savitsky, T. D., & McCafrey, D. F. (2014) Bayesian hierarchical multivariate formulation with factor analysis for nested ordinal data. *Psychometrika, 79,* 275 – 302.

Vermunt, J. K. (2003). Multilevel Latent Class Models. *Sociological Methodology*, *33*, 213–239.

Vermunt, J. K. (2004). An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica*, *58*(2), 220–233.

White, T. (2014). *Adding Eyes: The Rise, Rewards, and Risks of Multi-Rater Teacher Observation Systems*. Stanford, CA: Carnegie Foundation for the Improvement of Teaching.

## Appendix B. Tables and Figures

*Figure 1.* Precision of Classification of Teachers as a Function of Number of Observations

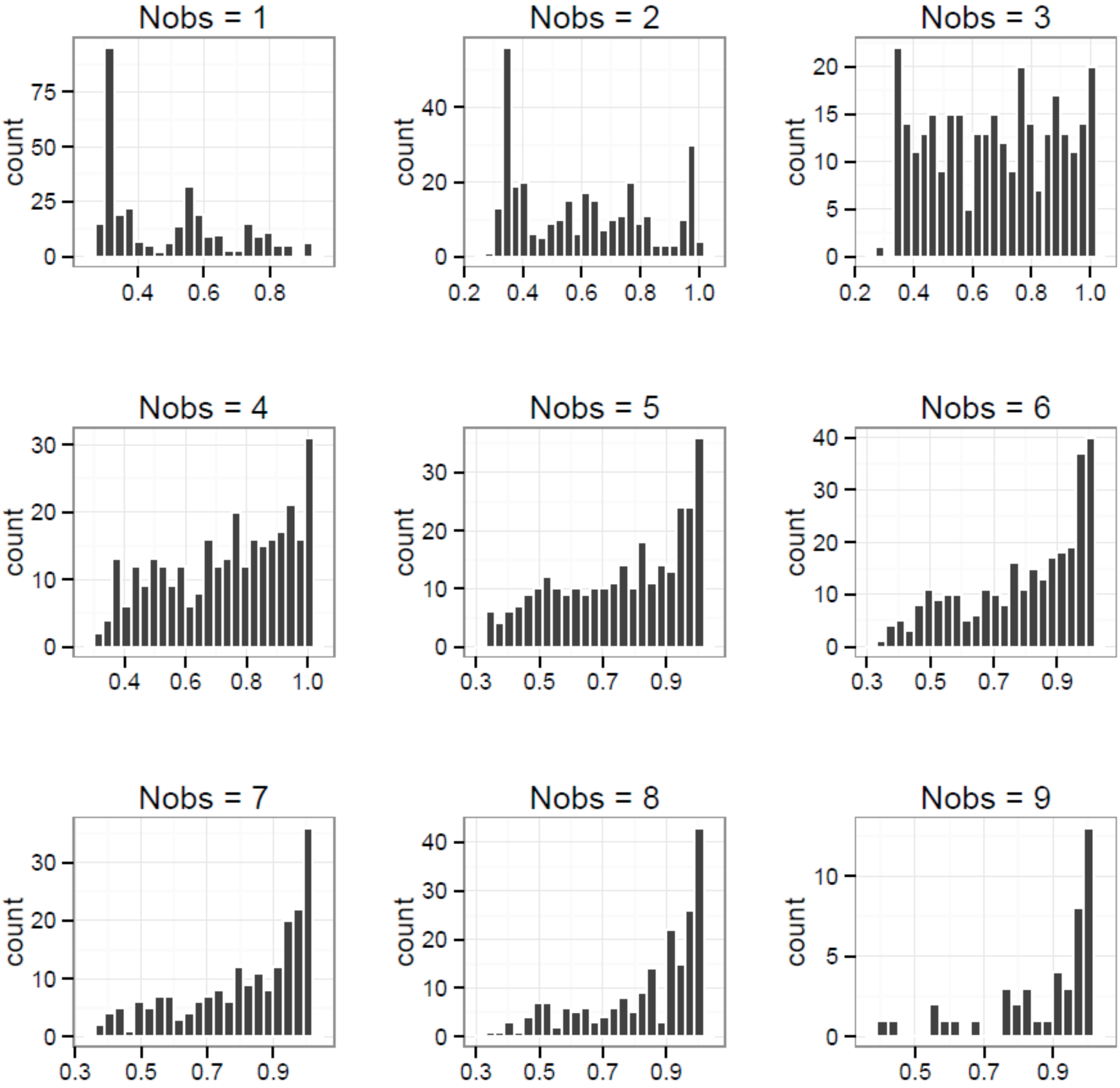
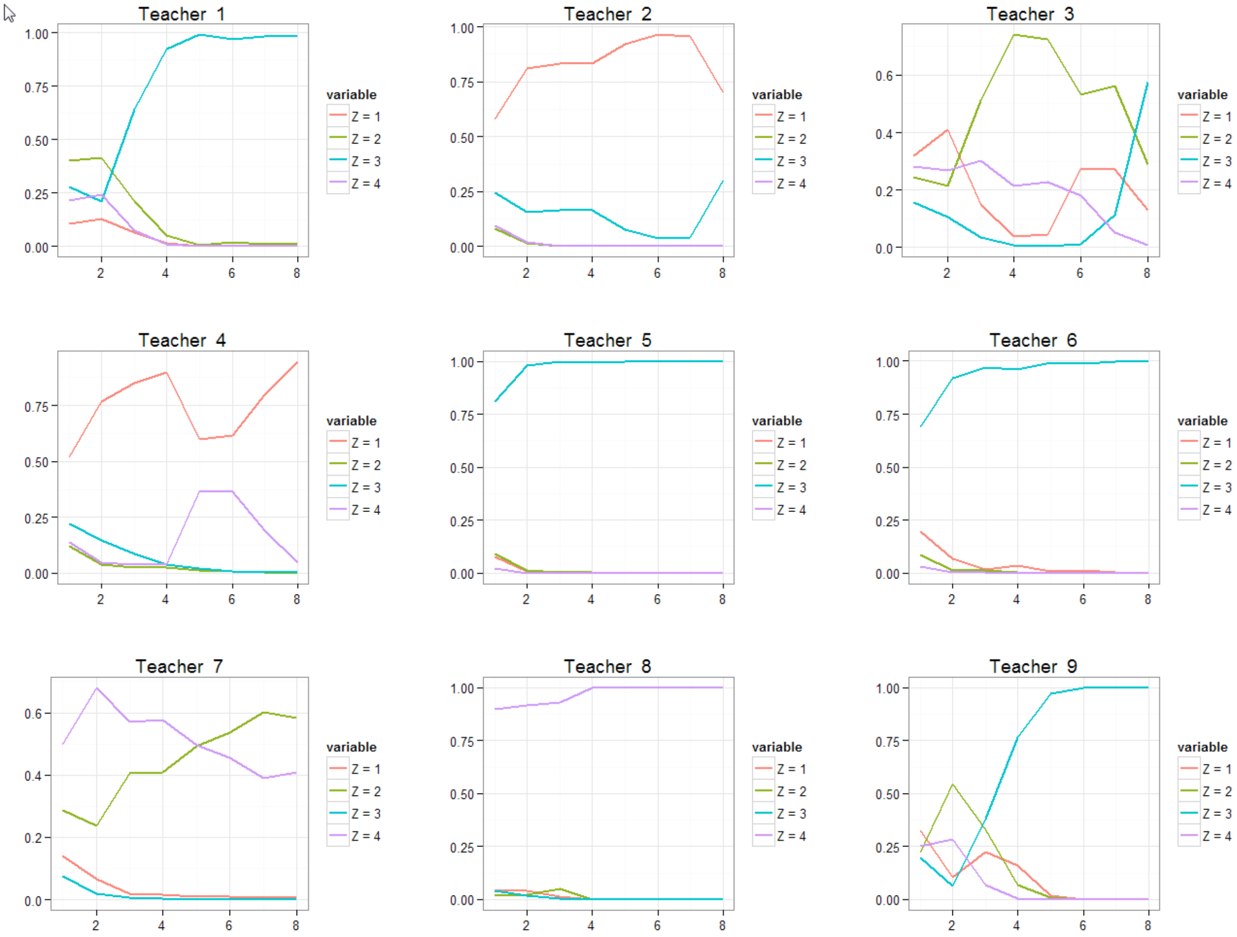


*Figure 2*. Consistency of Classification of $N = 9$ Teachers as a Function of Number of Observations.


Teacher 1
Teacher 2
Teacher 3
Teacher 4
Teacher 5
Teacher 6
Teacher 7
Teacher 8
Teacher 9
variable
Z = 1
Z = 2
Z = 3
Z = 4