Analyzing the Factorial Structure of the Classroom Assessment Scoring System-Secondary Using a Bayesian Hierarchical Multivariate Ordinal Model

Kun Yuan[*]
Daniel F. McCaffrey
Terrance D. Savitsky

RAND Corporation

[*] Emails addresses for all authors are: Kun Yuan, kyuan@rand.org; Daniel F. McCaffrey, danielm@rand.org; Terrance D. Savitsky, savitsky@rand.org. Please contact Kun Yuan for any questions about this abstract.

Abstract Body

Background / Context:
Standardized teaching observation protocols have become increasingly popular in evaluating teaching in recent years. One of such protocols that has gained substantial interest from researchers and practitioners is the Classroom Assessment Scoring System-Secondary (CLASS-S) (Pianta, Hamre, Haynes, Mintz, & LaParo, 2007). According to the developer, CLASS-S has three domains of teacher-student interactions: Classroom Organization, Emotional Support, and Instructional Support. Each domain is associated with 3–4 specific dimensions (see Figure B.1 in Appendix B). Dimensions are scored on a 1–7 scale according to specific behavioral indicators. Domain scores are derived by averaging scores of their associated dimensions. Despite the great interest in using this tool to measure teaching, its theoretical structure is rarely tested using empirical data. Moreover, a couple of studies that did examine the structure of CLASS-S or CLASS at the kindergarten level did not account for the hierarchical multivariate ordinal nature of the scores (Paro, Pianta, & Stuhlman, 2004; Malmberg, Hagger, Burn, Mutton, & Colls, 2010; Pakarinen, et al., 2010).

Purpose / Objective / Research Question / Focus of Study:
The goal of this study is to examine the structure of CLASS-S while accounting for the hierarchical multivariate ordinal nature of data collected through this protocol.

Setting:
Data used in this study came from two sources. The first source is the Understanding Teacher Quality (UTQ) study that took place middle schools in three large school systems from the same metropolitan region in the southeastern United States. The second source is the Toward an Understanding of Classroom Context (TUCC) study conducted in middle and high schools in an urban fringe mid-Atlantic school district.

Population / Participants / Subjects:
The UTQ study includes 458 teachers teaching mathematics (n=231) or English language arts (n=227) to sixth, seventh, or eight graders. The teachers were 83% female, 56% non-Hispanic white, 36% black, and 8% Hispanic and other race and they average 9.6 years of experience in the district. On average 47% of the students in each class were eligible for free or reduced price meals, 44% were non-Hispanic black, 34% were non-Hispanic white, and 11% were Hispanic.

The TUCC study included 82 teachers in 20 middle and 20 high schools. Fifty percent of the teachers are black, 26% are Asian or Asian American and 20% are white. Roughly 90 percent of the students served by the district were of color and 55 percent students were eligible for free or reduced price meals.

Significance / Novelty of study:
This is the first study conducted to examine the structure of CLASS-S while accounting for the hierarchical multivariate ordinal nature of scores. In particular unlike other studies, our analysis acknowledges that factor structure may differ at different levels of the hierarchy. It also shows that analyst must be explicit in their specification of the inference they are making and use models that can estimate structure consistent with those inferences.

Results from this study contribute to the current understanding about the structural validity of CLASS-S. Because CLASS-S is similar to other observation protocol, the study also has implications of those measures too.

Statistical, Measurement, or Econometric Model:
We examined the structure of CLASS-S using a Bayesian hierarchical multivariate ordinal model (referred to as the BHMLOM model below) (Savitsky & McCaffrey, under review).

The observational data from CLASS-S consist scores from multiple ratings from lessons from teacher's classroom on the ten dimensions evaluated by the protocol. Each dimension is scored on a seven point scale. Let y_{ijklr} equal the vector of 10 scores from rating r conducted one of n_r raters ($n_r = 11$ for UTQ and 5 for TUCC) for time segment l of session (lesson) k, of section j, taught be teacher i.

The BHMLOM specifies an associated latent vector μ_{ijklr} such that the probability that the d = 1, …, 10 component of y_{ijklr} (the score on dimension d) receives a score of s= 1, …, 7, equals the probability that the corresponding element of μ_{ijklr} is in specified interval:

$$\Pr(y_{ijklr,d} = s) = \Pr(\gamma_{d-1,s} < \mu_{ijklr,d} \leq \gamma_{d,s})$$

The vector μ_{ijklr} follows a multivariate hierarchical model such that

$$\mu_{ijklr} = \alpha + \beta_{(r)} + \theta_i + \psi_{ij} + \lambda_{ijk} + \varphi_{ijk} + \xi_{ijk}$$

where α is a vector of dimension specific means, $\beta_{(r)}$ is a vector of rater specific, dimension specific fixed effects to allow for some raters to be more or less lenient on each dimension than the average rater, θ_i is a vector of teacher-level Gaussian random effects with mean zero and variance $\Sigma_{teacher}$ and ψ_{ij}, λ_{ijk}, φ_{ijk}, and ξ_{ijk} are also multivariate Gaussian random effects with mean zero and separate variance covariance matrices for the sections, sessions, ratings, and residual variance. To study the structure of the multivariate scores, we use separate factor analytic specification for the variance-covariance matrix at each level of the hierarchy. Savitsky and McCaffrey (under review) provide prior distributions for all the model parameters and details on identifying assumptions for the factor analytic specifications. They also discuss model parameterization to improve the convergence of posterior estimates.

We also model rating level averages using a traditional multivariate hierarchical linear model with random effects for teachers, sections, sessions, and ratings and fixed effects for raters (Goldstein, 1995). These models yielded estimates of the variance-covariance matrix at each level of the hierarchy which we then modeled using factor analytic specifications to recover the structure at each level.

Usefulness / Applicability of Method:
This study demonstrates the applicability of using Bayesian hierarchical multivariate ordinal models for uncovering structure among the dimension scores for observation protocol. It provides a sharp contrast between models that separate structure from the multiple sources of variance in scores such as teachers, classroom, lessons, and ratings, and models that do not.

Construct irrelevant sources of variance (e.g., ratings) may have different correlational structure that can mask relevant structures and this study demonstrates the utility of the Bayesian hierarchical multivariate ordinal model for distinguishing among these structures.

Data Collection and Analysis:
UTQ collected data over two school years with roughly half of the teachers participating in each year. For each classroom two lessons were video recorded and then scored using CLASS-S. Video scoring started during the first school year and continued through the second year. Twenty percent of the videos were double coded by two separate raters. Half of the lessons from year one were also scored via live observation using the protocol.

The TUCC project observed four lessons per classroom with roughly one measure per quarter for each classroom. A fifth lesson was added for 80% of the classrooms (N=65). Every lesson was observed by one rater and video recorded. A second rater conducted an additional live observation for 20 percent of lessons and all videos were scored by two separate observers.

Trained raters observed a lesson or a time segment of it and then scored it according to the CLASS-S specifications. Lessons were evaluated on all 10 dimensions of CLASS-S. Each dimension received a score of a 1-7 for each dimension according to descriptive anchors provided in the protocol.

The hierarchical structure of data from UTQ includes four levels:

- Teacher -- the average teaching for a teacher across sections and all the lessons in a given year;
- Section – the average teaching for the entire year for a section or classroom of students receiving instruction together as unit during the school year;
- Session – the teaching during a specific observation session or lesson; and
- Rating – the score provided by a single rater observing a specific lesson.

Because TUCC included only one section for each teacher, data from that study data include three levels: teachers/section, session, and rating.

Raters received multiple days of training on the CLASS-S rubric and proved able to score in agreement with master codes before starting observations. Raters also conducted weekly calibration exercises with project staff for the entire study period until all scoring was complete. In these exercises raters scored training videos and compared their results with master codes. Project staff then reviewed the scores with raters and provided additional training when there were disagreements between the scores from the project observers and the master codes.

We conducted exploratory factor analyses (EFA) at each level of the hierarchical structure for each set of data. In addition to using the BHMLOM and HLM models, we conducted EFA using mean dimension scores at different levels. This approach has been used in previous studies and it ignores both the hierarchical multivariate structure and the ordinal nature of data. We also conducted confirmatory factor analyses (CFA) at the teacher level for both sets of data to examine how well the three-factor theoretical structure of CLASS-S fits with the observed data.

Findings / Results:
Results showed different factorial structure of CLASS-S at different levels (see Figure B.2. and B.3. in Appendix B). At the teacher level, EFA results suggested a one-factor model. At the section (UTQ) and session (UTQ and TUCC) levels, the data did not yield clear factor structure. At the rating level, EFA results from both sets of data suggest a two-factor model, with measures of classroom organization loading on a second factor, and all other dimensions loading on one a main teaching capacity factor.

Factor analysis using correlation matrix from the HLM model revealed the same structure of CLASS-S as that using the BHMLOM model at the teacher level. Results from the HLM model were different from those using the BHMLOM model at other levels. Factor analysis results using mean dimension scores showed factorial structures of CLASS-S different from those identified by the BHMLOM model at all levels.

EFA and CFA results using three models and two data sets do not support the three-factor theoretical model of CLASS-S in general. The three-factor model only had a moderately good fit to rating-level data. It did not fit the data at any other level even moderately well. This suggests that this three-factor model may have resulted from the rating process.

Conclusions:
To obtain structurally valid inferences about teaching using the multivariate nested ratings from standard teaching observation protocols such as CLASS-S, we need to account for the hierarchical multivariate ordinal nature of scores. Results from this study showed that accounting for the hierarchical multivariate ordinal nature of data does matter for understanding the structure of CLASS-S scores. Results provided little support for the three-factor theoretical model and suggested that rating process might introduce additional structure into the scores.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., & Hamre, B. K. (under review). Effects of observation mode on measures of secondary mathematics teaching. Journal of Educational Measurement.

Goldstein, H. (1995). Multilevel Statistical Models, Second Edition. New York: John Wiley & Sons Inc.

Malmberg, L., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. Journal of Educational Psychology, 102(4), 916-932.

Pakarinen, E., Lerkkanen, M., Poikkeus, A., Kiuru, N., Siekkinen, M., Rasku-Puttonen H., & Nurmi, J. (2010): A Validation of the Classroom Assessment Scoring System in Finnish Kindergartens, Early Education & Development, 21(1), 95-124.

Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. The Elementary School Journal, 104(5), 409-426.

Pianta, R.C., Hamre, B. K., Haynes, N. J., Mintz, S. L. & L Paro, K. M. (2009). Classroom Assessment Scoring System (CLASS), Secondary Manual. Charlottesville, VA: University of Virginia Center for Advanced Study of Teaching and Learning.

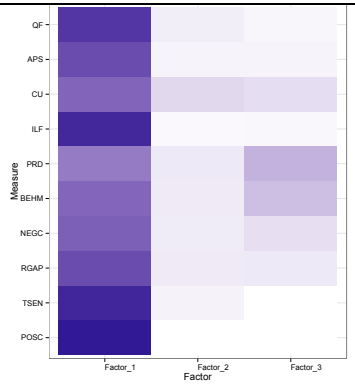
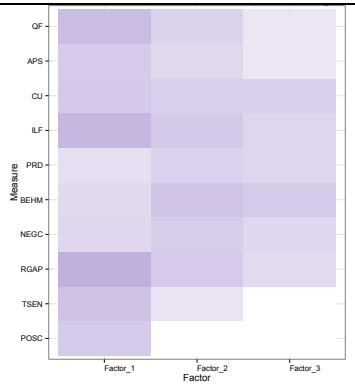
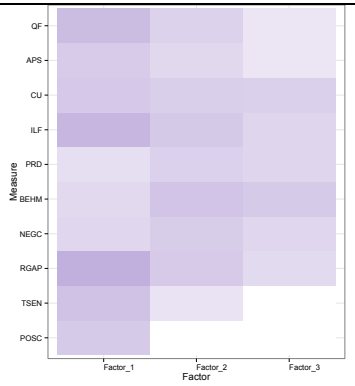
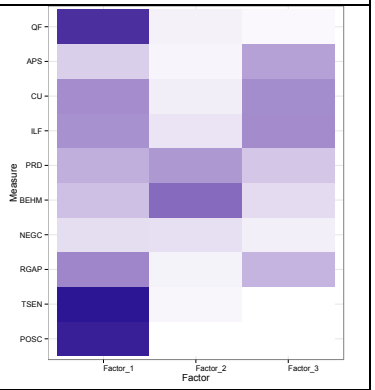
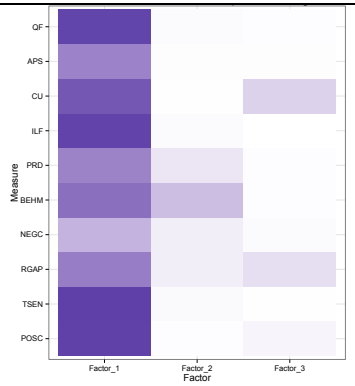
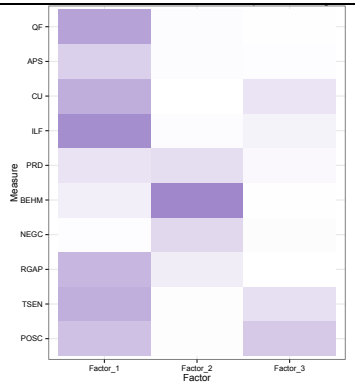
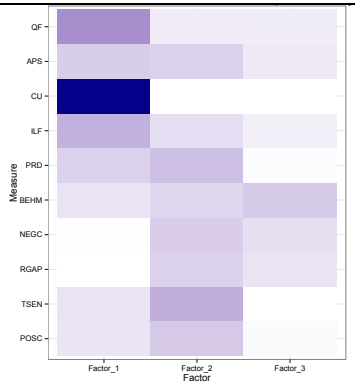
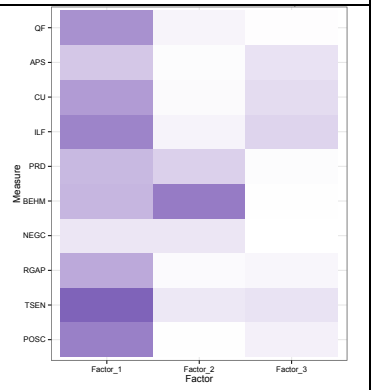
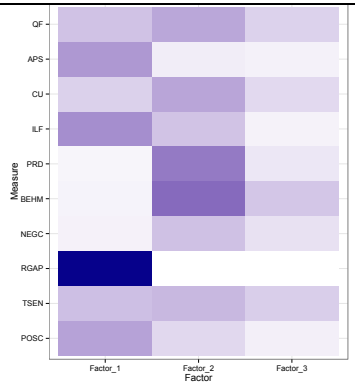
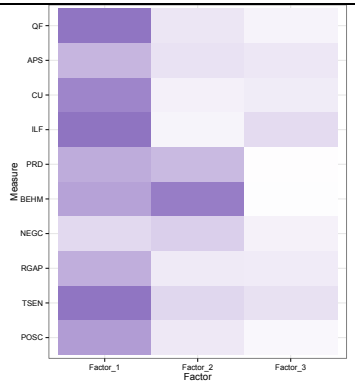
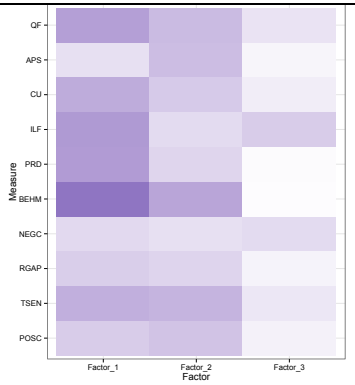
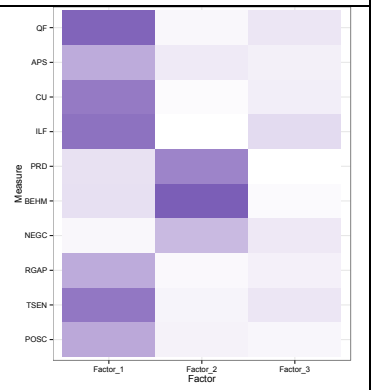
Savitsky, T.D. and McCaffrey, D. F. (under review) Bayesian hierarchical multivariate formulation with factor analysis for nested ordinal data. Psychometrika.

Appendix B. Tables and Figures

Figure B.1. Theoretical Structure of CLASS-S

Domain	Dimensions	Dimension Description
Emotional Support	Positive Climate (POSC)	reflects the emotional connection and relationships among teachers and students, and the warmth, respect, and enjoyment communicated by verbal and non-verbal interactions
	Teacher Sensitivity (TSEN)	reflects the teacher's responsiveness to the academic and social/emotional needs and developmental levels of individual students and the entire class, and the way these factors impact students' classroom experiences
	Regard for Adolescent Perspectives (RGAP)	focuses on the extent to which the teacher is able to meet and capitalize on the social and developmental needs and goals of adolescents by providing opportunities for student autonomy and leadership; also considered are the extent to which student ideas and opinions are valued and content is made useful and relevant to adolescents
Classroom Organization	Negative Climate (NEGC)	reflects the overall level of negativity among teachers and students in the class; the frequency, quality, and intensity of teacher and student negativity are important to observe
	Behavior Management (BEHM)	encompasses the teacher's use of effective methods to encourage desirable behavior and prevent and redirect misbehavior
	Productivity (PRD)	considers how well the teacher manages time and routines so that instructional time is maximized; captures the degree to which instructional time is effectively managed and down time is minimized for students; it is not a code about student engagement or about the quality of instruction or activities
Instructional Support	Instructional Learning Formats (ILF)	focuses on the ways in which the teacher maximizes student engagement in learning through clear presentation of material, active facilitation, and the provision of interesting and engaging lessons and materials
	Content Understanding (CU)	refers to both the depth of lesson content and the approaches used to help students comprehend the framework, key ideas, and procedures in an academic discipline; at a high level, refers to interactions among the teacher and students that lead to an integrated understanding of facts, skills, concepts, and principles
	Analysis & Problem Solving (APS)	assesses the degree to which the teacher facilitates students' use of higher level thinking skills, such as analysis, problem solving, reasoning, and creation through the application of knowledge and skills; opportunities for demonstrating metacognition, i.e., thinking about thinking, also included
	Quality of Feedback (QF)	assesses the degree to which feedback expands and extends learning and understanding and encourages student participation; in secondary classrooms, significant feedback may also be provided by peers; regardless of the source, focus here should be on the nature of the feedback provided and the extent to which it "pushes" learning

Figure B.2. Factor Loadings of Exploratory Factor Analysis at Four Levels Using Data from the UTQ Study

Model\Level	Teacher	Section	Session	Rater
BHMLOM				
HLM				
Means				

NOTES: Each plot shows factor loadings from a three-factor exploratory factor analysis. POSC = Positive Climate; TSEN = Teacher Sensitivity; RGAP = Regard for Adolescent Perspectives; NEGC = Negative Climate; BEHM = Behavior Management; PRD = Productivity; ILF = Instructional Learning Formats; CU = Content Understanding; APS = Analysis & Problem Solving; QF = Quality of Feedback. The darkness of color represents the value of factor loading. The following legend applies to figures in all cells.

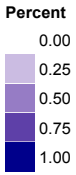
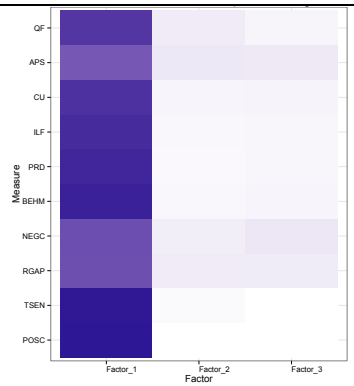
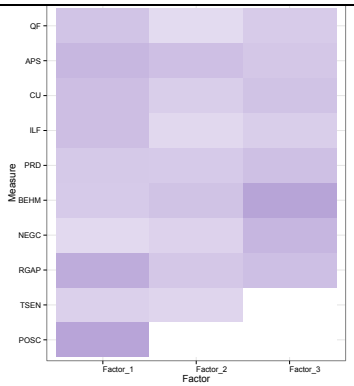
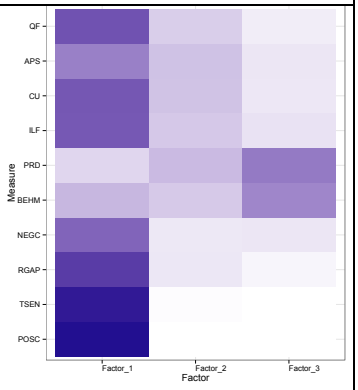
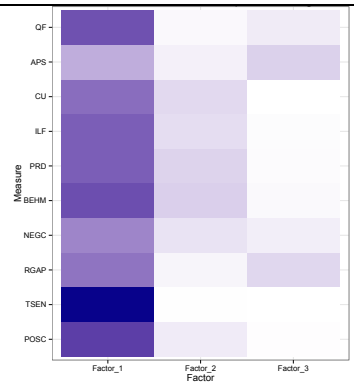
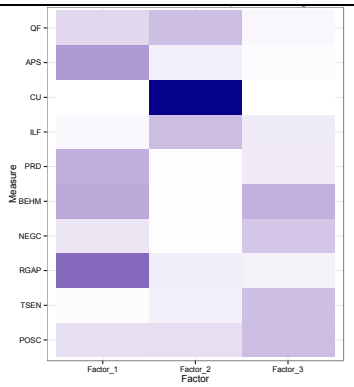
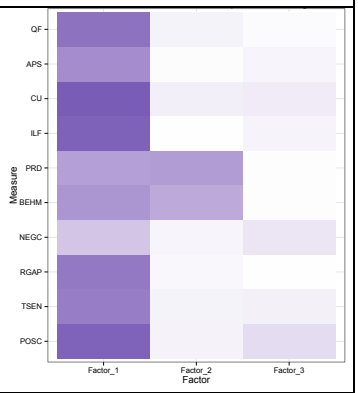
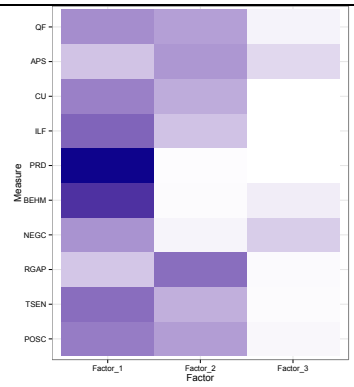
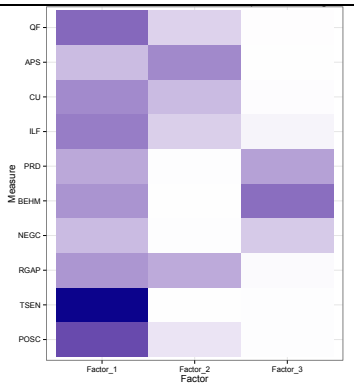
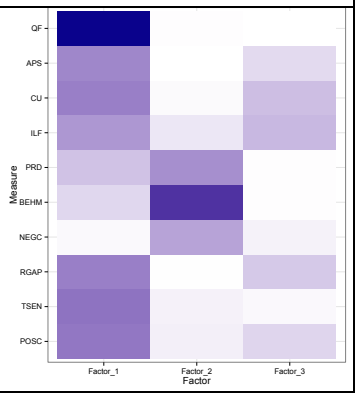


Figure B.3. Factor Loadings of Exploratory Factor Analysis at Three Levels Using Data from the TUCC Study

Model\Level	Teacher	Session	Rater
BHMLOM			
HLM			
Means			

NOTES: Each plot shows factor loadings from a three-factor exploratory factor analysis. POSC = Positive Climate; TSEN = Teacher Sensitivity; RGAP = Regard for Adolescent Perspectives; NEGC = Negative Climate; BEHM = Behavior Management; PRD = Productivity; ILF = Instructional Learning Formats; CU = Content Understanding; APS = Analysis & Problem Solving; QF = Quality of Feedback. The darkness of color represents the value of factor loading. The following legend applies to figures in all cells.