Abstract Title Page

Title: Rater Drift and Time Trends in Classroom Observations

Authors and Affiliations:

Jodi M. Casabianca, Carnegie Mellon University, jodicasa@andrew.cmu.edu

J. R. Lockwood, RAND Corporation, lockwood@rand.org

Background / Context:
Classroom observation protocols, in which observers rate multiple dimensions of teaching according to established protocols (either live in the classroom, or post-hoc from lesson videos), are increasingly being used in both research and policy contexts. Researchers are studying how these measures relate to other potential measures of teacher quality such as those based on student outcomes (Bill and Melinda Gates Foundation, 2012), and states and districts are increasingly introducing these measures into their teacher accountability systems. In both contexts, scores from typically only a few lessons are aggregated to derive measures of an individual teacher's teaching.

Classroom observation protocol scores have many sources of error as measures of the attributes of an individual teacher and/or their teaching, even when these attributes are narrowly defined within a given school year. Day to day variation in classroom activities, and variation in curricula and context across different classrooms a teacher may teach (e.g. middle school teachers who teach multiple sections or courses each year), mean that a small sample of lessons could vary widely on scores even when scored by a common rater. Variation among raters and changes in raters' use of the rubric score scale over time also contribute to error. Despite efforts to retrain raters with calibration sessions and frequent feedback, rating judgments may still change over time (Bock, 1995; Congdon & McQueen, 2000; McKinley & Boulet, 2004). Recent research has discussed rater effects specifically rater severity drift, central tendency, and rater experience/learning (Leckie & Baird, 2011; Myford & Wolfe, 2009), and how they relate to ratings and changes in score scale category use over multiple ratings. Casabianca and colleagues (2012) showed that two of three domain scores from video observations of the Classroom Assessment Scoring System, Secondary (CLASS-S) followed a downward trend during the scoring period and then later in the period raters increased their scores; during the scoring period there was a 1-point decrease and then a 1-point increase, on a 1-7 Likert scale. Understanding the nature and magnitude of rater variation and time trends in scoring is critical to designing measurement systems for teaching that are minimally impacted by these sources of variance and preventing them from manifesting as bias – i.e. systematic errors in scores for teachers that would replicate under repeated realizations of the measurement process.

Purpose / Objective / Research Question / Focus of Study:
The purpose of this study is to use scores from video observations from multiple teaching observation protocols administered as part of the Understanding Teaching Quality (UTQ) study to investigate and understand the nature of rater variation due to rating severity, including overall time trends and variations among raters. The scoring times for the videos were designed to be unrelated to the actual date of the lesson, so that any time trends in the scores reflect time trends in how the raters were using the scales over the course of the rating period rather than changes in quality of instruction over the course of the school year. Our analysis is structured to answer three research questions:

(1) What happens in terms of individual and overall rater severity over the duration of the video scoring period?

(2) How much variance in scores is attributable to time trends in rater effects, and how much could the reliability of teacher scores improve if the rater drift could be removed?

(3) Given that UTQ uses the same raters for four different classroom observation protocols, how do the protocols compare in terms of Research Questions (1) and (2)?

Setting:
The UTQ study took place in middle schools in three large school systems from the same metropolitan region in the southeastern United States.

Population / Participants / Subjects: The study includes 458 teachers teaching mathematics (n=231) or English language arts (n=227) to sixth, seventh, or eight graders. Two classes (or sections) of students were observed for each of two lessons for each teacher for a total of 916 sections and 1,829 lesson scored. Thirty-four percent of the sections were grade 6, 29 were grade 7, 36 were grade 8 and the rest were mixed grade. 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.

Significance / Novelty of study:
This research will analyze rater severity drift for four different classroom observation protocols and determine its effect on the error variance. Understanding the effects of time trends on scores can inform researchers and practitioners so that they can design teacher evaluation systems more effectively. The statistical model used in this study will be useful to other researchers interested in modeling rater effects over time. Additionally, because this study includes an analysis of four classroom observation protocols, there is a replication aspect to our research that will permit us to provide a rich discussion of our research questions using different research instruments. This directly aligns with the theme of the spring 2013 SREE Conference, *Capitalizing on Contradiction: Learning from Mixed Results.*

Statistical, Measurement, or Econometric Model:
The study addresses the three research questions by augmenting a traditional approach to disentangling the variation in the scores. Generalizability, "G", theory, is an analysis of variance approach to partition a score into an effect for each facet or source of variability. Generalizability (G) studies (Brennan, 2001) have been used to evaluate sources of variance in classroom assessments for decades (see Erlich & Borich, 1979; Frederiksen, Sipusic, Sherin, & Wolfe, 1998; Hill, Charalambous & Kraft, 2012; Meyer, Cash, & Mashburn, 2012; Newton, 2010; Shavelson & Dempsey-Atwood, 1976).

A basic model for decomposing the score X_{iclsr} for a dimension of a scoring protocol from a rating of one teacher (*i*), in a classroom (*c*) on one lesson (*l*), for one segment (several minutes) of the lesson (*s*) by one rater (*r*) is as follows. Let θ_i be the teacher effect, η_{ic} the classroom within teacher effect, γ_{icl} the lesson within classroom (within teacher) effect, ω_{icls} the segment within lesson (within classroom within teacher) effect, β_r the rater effect, ϕ_{iclr} the rater by lesson effect, and ε_{iclsr} is residual error. Then the G study equation

$$X_{iclsr} = \mu + \theta_i + \eta_{ic} + \gamma_{icl} + \omega_{icls} + \beta_r + \phi_{iclr} + \varepsilon_{iclsr},$$

specifies the different sources of observed score variance that can be decomposed with the available data. For inferences about teaching and the effectiveness of individual teachers, it

would be preferable if teachers accounted for a substantial proportion of score variation and factors like raters, specific classrooms or specific lessons, and their interactions, did not. That is, ideally, the variance component for θ_i would be large relative the variances of all other terms.

We wish to compare results from this variance components model, which does not account for lesson scoring time, to one that does. We thus consider the following augmented model in which $\mathbf{p}(\mathbf{t})$ is a polynomial or other basis function of the scoring time t :

$$X_{iclsrt} = \mathbf{p}(\mathbf{t})'\boldsymbol{\mu} + \theta_i + \eta_{ic} + \gamma_{icl} + \omega_{icls} + \mathbf{p}(\mathbf{t})'\boldsymbol{\beta}_{\mathbf{r}} + \phi_{iclr} + \varepsilon_{iclsr},$$

That is, this equation augments the G study model in the previous equation for decomposing video scores rated at a specific time (t) to include an overall time trend and rater-specific time-trends. We estimate $\boldsymbol{\mu}$ and then model $\boldsymbol{\beta}_{\mathbf{r}}$ as a multivariate random effect.

For each dimension score from each scoring protocol, we will estimate the base model and then the augmented model for different choices for $\mathbf{p}(\mathbf{t})$ including linear, quadratic, cubic and quartic. We will use likelihood ratio tests to test whether the models that incorporate scoring time provide a better fit to the data, and if so, whether nonlinear time trends are required as suggested by previous research (Casabianca et al., 2012). We will obtain estimated variance components from the base model, and from the best-fitting time-augmented model, to conduct Decision (D) studies (Brennan, 2001) to quantify how much reliability of teacher scores could be improved under different measurement scenarios if the time trends in the scoring could be mitigated (e.g. through improved rater training and calibration). Unmodeled time trends could manifest as either rater variance or rater by lesson variance, so we anticipate that to the extent that there are time trends, the variance component corresponding to ϕ_{iclr} will decrease under the augmented model relative to the base model, and so may that for the intercept portion of $\boldsymbol{\beta}_{\mathbf{r}}$. We will quantify how much any reductions in these sources of variance increases the reliability of teacher measures under different hypothetical measurement scenarios similar to what are currently being used in practice. We will also consider how much the relative rankings of teachers' score could change depending on when in the rating period they were rated and by whom. Finally, we will compare findings across dimensions both within and across protocols to see if different dimensions have markely different time trends, and also to examine the consistency of any rater-specific time trends across dimensions.

Usefulness / Applicability of Method:
Our study uses four classroom observation scoring protocols to demonstrate a statistical model that incorporates a time-varying component into the traditional G study model to account for scenarios where raters rate over a scoring period that is not related to the date of the actual lesson. Time trends in ratings and rater effects have been studied in recent literature (Leckie & Baird, 2011; Meyer, Cash, & Mashburn, 2011; Myford & Wolfe, 2009); our modeling approach is applicable in these and similar situations.

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 three observation protocols: Classroom Assessment Scoring System, Secondary, CLASS-S (Pianta, Hamre, Haynes, Mintz, & LaParo, 2007) and Danielson's

Framework for Teaching, FFT, (Danielson, 2007) and either the Protocol for Language Arts Teaching Observations, PLATO (Grossman et al., 2010), for English language arts classes, or the Mathematical Quality of Instruction, MQI (Hill et al., 2008), for mathematics classes. 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. For each protocol raters observed a lesson or a time segment of it and then scored it according to the protocol specifications. Lessons are evaluated on between 10 and 20 dimensions of teaching depending on the protocol. Each dimension receives a score on a 1-3, 1-4 or 1-7 for each dimension depending on the protocol according to descriptive anchors provided in the protocol.

Raters received multiple days of training on each rubric and proved able to score in agreement with master codes before starting classroom or video observations. Raters also conducted weekly calibration exercises with project staff for the entire study 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.

Findings / Results:

We have not yet fit the augmented variance decomposition model to these data. We present preliminary findings based the CLASS-S protocol scores, which consists of 10 dimension scores organized into three conceptual domains: *Classroom Organization* (CO), *Emotional Support* (ES), and *Instructional Support* (IS). We averaged the dimension scores to the domain level, and then averaged the domain scores across segments within lessons. Appendix B provides 11 plots of the scores versus scoring time, one for each of the raters participating in the study. Trendlines on the plots are based on a fourth degree polynomial model fit by domain and by rater.

The plots provide evidence of time trends that appear to vary by domain and by rater. The ES and IS domains appear more sensitive to time, while the CO domain appears to stabilize quickly after the start of scoring. This was true for most raters. Most raters also assigned higher scores to the CO domain. However, from rater to rater, there was a lot of variation in the trends over scoring days. For example, rater 343 started scoring around 2.5 on IS, 3.2 on ES, and 5 on CO. Around scoring days 70-75, rater 343 had assigned scores about 1 full point higher, and then for the rest of the scoring period duration this rater's scores decreased to a level lower than the beginning of the scoring period. Rater 315's trend was relatively stable; typical scores at the start of the scoring period (3 on IS, 3.5 on ES, and 6 on CO) remained at the same level. Another distinct time trend is for rater 322, who started with IS and ES scores around 4 and CO scores around 5, showed a steep decrease in scores for IS, but increases for ES and CO. Then this rater did not assign scores for almost 100 days during the scoring period, but when they began scoring again, their scores were at the same level or higher from when they stopped. For the remainder of the scoring period scores for all domains decreased, but never returned to the original level. Plots for the remaining 8 raters also show variation in terms of trends and rater severity.

Conclusions:

The descriptive analyses suggest time trends that vary by domain and by rater, suggesting that scoring time is consequential. Thus the formal analyses proposed above are warranted and will allow us to quantify the impacts that these trends have on teacher scores.

Appendix A: References

Bill and Melinda Gates Foundation, Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Seattle: The Bill and Melinda Gates Foundation. Downloaded from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf, January 8, 2012.

Brennan, R. L. (2001). Generalizability theory. New York: Springer-Verlag.

Bock, R. D. (1995). Open-ended exercises in large-scale educational assessment. In L. B.

Resnick & J. G. Wirt (Eds.), *Linking school and work: Roles for standards and assessment* (pp. 305–338). San Francisco: Jossey-Bass.

Casabianca, J. C., McCaffrey, D. F., Bell, C. A., Gitomer, D., Hamre, B., & Pianta, R. (2012). The effect of observation mode on measures of mathematics secondary mathematics teaching. Submitted for publication.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*, 163–178.

Danielson, C. (2007) *Enhancing Professional Practice: A framework for teaching* Alexandria, VA: Association for Supervision and Curriculum Development.

Erlich, O. & Borich, G. (1979). Occurrence and generalizability of scores on a classroom interaction instrument. *Journal of Educational Measurement, 16*, 11-18.

Frederiksen, J. R., Sipusic, M., Sherin, M., & Wolfe, E. W. (1998). Video portfolio assessment: Creating a framework for viewing the functions of teaching. *Educational Assessment, 5*, 225-297.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., et al. (2010, May). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores* (NBER Working Paper No. 16015). Retrieved from http://www.nber.org/papers/w16015

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26,* 430–511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability. *Educational Researcher, 41,* 56-64.

Leckie, G. & Baird, J. A. (2011). Rater effects on essay scoring: A multi-level analysis of

severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48,* 399-418.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McKinley, D.,& Boulet, J. R. (2004). Detecting score drift in a high-stakes performance-based assessment. *Advances in Health Sciences Education*, *9*, 29–38.

Meyer, J. P., Cash, A. H., & Mashburn, A. (2012). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment, 16,* 227–243.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from a sparse matrix sample of item responses. *Journal of Educational Measurement, 29,* 131–154.

Myford, C. M., and Wolfe, E. M. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale use. *Journal of Educational Measurement, 46,* 371-389.

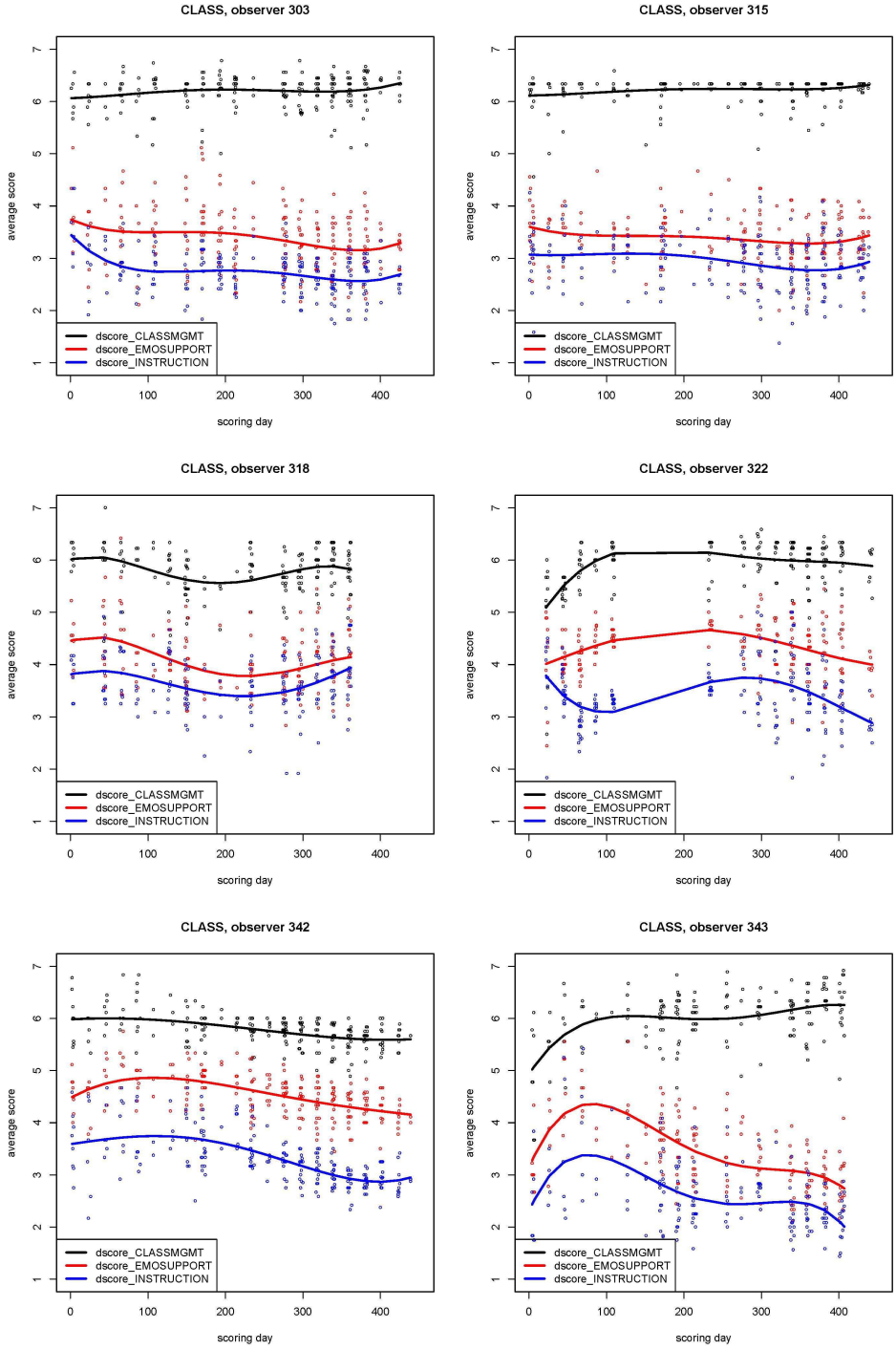
Newton, X. A. (2010). Developing indicators of classroom practice to evaluate the impact of district mathematics reform initiative: A generalizability analysis. *Studies in Educational Evaluation, 36*, 1-13.

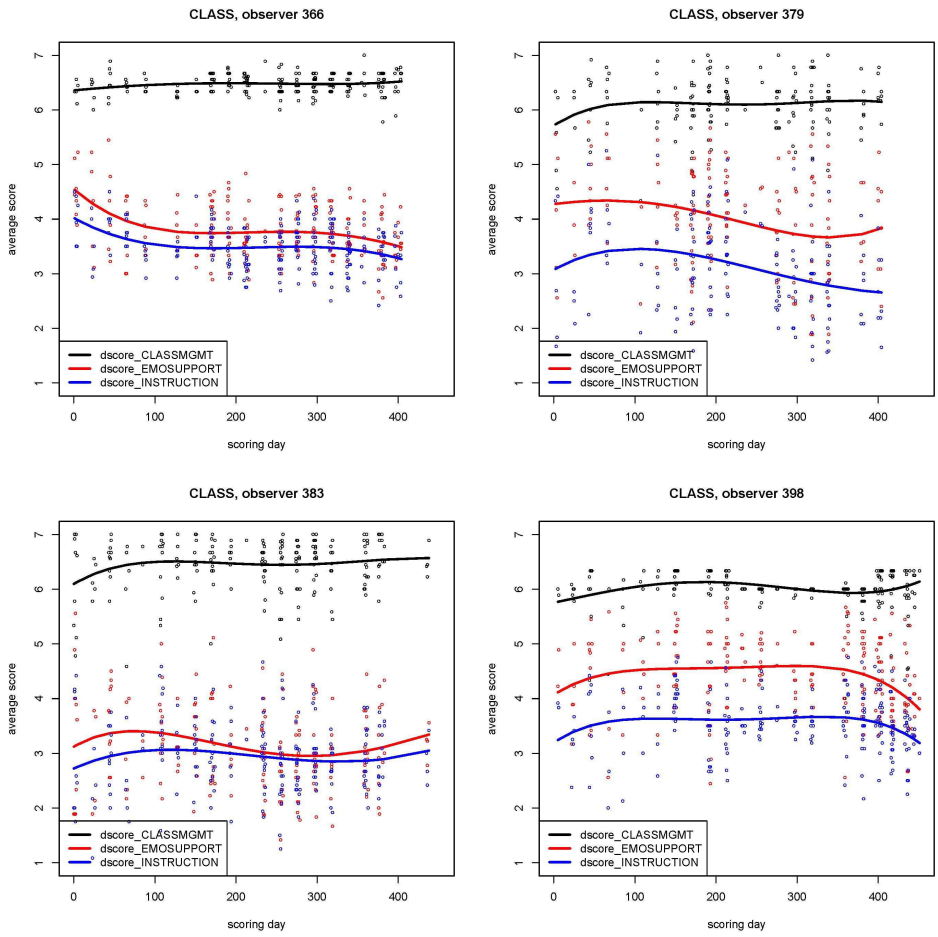
Patz R. J., Junker B. W., Johnson, M.S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341-384.

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.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research, 46*, 553-611.

Appendix B. Tables and Figures




CLASS, observer 366
CLASS, observer 379
CLASS, observer 383
CLASS, observer 398
average score
scoring day
dscore_CLASSMGMT
dscore_EMOSUPPORT
dscore_INSTRUCTION