Abstract Title Page

Title: Effect of Observation Mode on Measures of Teaching

Authors and Affiliations:

Daniel F. McCaffrey, RAND Corporation, danielm@rand.org

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

Abstract Body

Background:

As the education reform movement increasingly focuses on teachers and teaching, educators, policy-makers, and researchers need valid and reliable measures that can be used to evaluate individual teachers, provide guidance for improving teaching performance, and support research in ways that advance instruction and classroom dialog and practice. A new generation of classroom evaluation tools has recently been developed to support evaluation of teaching.

Live observations tend to be the standard for studies of teaching and teacher evaluations in practice. They have the benefit of the observer being in the teacher's physical classroom. This is valuable for teacher evaluations because it gives observation scores credibility among teachers. Using video provides particular affordances because they create a permanent record and teachers can review them to evaluate their own instruction as professional development (Miller, 2007; Sherin & Han, 2004; van Es & Sherin, 2010). Videos can be scored by multiple raters, which can reduce error by averaging scores. The use of video also allows for scores to be audited as a part of quality control and videos can be evaluated using multiple scoring protocols to assess the robustness of inferences to a protocol. For most of these reasons, many recent studies of classrooms have made use of videos (Bill and Melinda Gates Foundation, 2012).

Given these affordances, an important issue is to understand the comparability of the nature and quality of information created through these two observation modes. Nearly 20 years ago Jaeger (1993) identified mode of observation as potentially contributing to the psychometric properties of measuring teaching, but little research on mode effects has occurred since.

Research Questions:

This study is a first step toward rectifying the dearth in knowledge on the effects of observation mode on the psychometric properties of classroom teaching evaluations. It tests for observation mode effects on inferences about teaching, classrooms, and teachers. Specifically, we answer the research questions:

1. Do raters systematically give higher scores using one observation mode or the other?
2. Does the observation mode affect the rank ordering of scores?
3. Does the observation mode affect the size of the standard errors of measurement or the reliability of scores?
4. What are the implications of errors for inferences about the teaching in a lesson or for a classroom for a year?

Setting:

We use data collected for the study Toward an Understanding of Classroom Context (TUCC) to test for mode effects in the scores and inferences about the teaching in lessons and in classrooms. TUCC took place in middle and high schools in an urban fringe mid-Atlantic school district that serves roughly 90 percent students of color and 55 percent students who are eligible for free or reduced price meals.

Participants:

The study was designed as a validity study of the Classroom Assessment Scoring System – Secondary, (CLASS-S; Pianta, Hamre, Haynes, Mintz, & LaParo, 2007) and sought to investigate various measures of teaching quality in algebra classrooms and their relationships to one another. All participating teachers in the district taught a course that was considered some variant of Algebra 1. Of the 208 algebra teachers teaching in 56 schools identified by the district, 92 originally agreed to participate in the study. After attrition, 82 teachers in 20 middle and 20 high schools were a part of the final sample. We gathered observational data from one section (classroom) of students per teacher, resulting in data from 26 middle school and 56 high school classrooms.

Most teachers in the study (85%) held standard professional certificates of some form. All teachers had at least a bachelor's degree, whereas 53% had a master's degree or higher. Almost three fourths of teachers (73%) had taught math for 5 or more years; however, slightly more than half of teachers (53%) had taught algebra for 5 or more years. Fifty percent of the teachers are black, 26% are Asian or Asian American and 20% are white.

Intervention / Program / Practice:

The study tested the effect of mode of observation on the CLASS-S, which is organized around three domains of teacher-student interactions: Classroom Organization, Emotional Support, and Instructional Support. Each domain is associated with 3–4 specific dimensions of teacher-student interactions (Figure 1). Dimensions are scored on a 1–7 scale according to specific behavioral indicators. Domain scores are derived from their associated dimension scores.

Insert Figure 1 about here

In this study, individual lessons were divided into observation segments. A segment was defined as a 22 minute period in which the first 15 minutes were used to watch classroom interactions and take notes using observation software on a laptop. The next 7 minutes were used to assign scores for each of the 11 dimensions using the same software. Coding segments for live and video cases were identical for this study.

A classroom's lesson score on each CLASS-S dimension is the average of the scores from all segments in that lesson which, because lessons varied in length, typically included two to four segments. Scores were averaged across dimensions to obtain domain scores at the segment-level and then averaged at the lesson and classroom levels. Annual evaluations would typically use classroom level scores.

Either four or five lessons from each of the 82 TUCC classrooms were observed live by one or two of the five raters used by the study. Each lesson was also video recorded and two separated raters scored it by observing that recording.

Significance / Novelty of study:

Teaching and teachers are central to current education policy. Classroom observations are a key component of the new evaluation systems states and districts are implementing in response to their focus on teachers. Observations are also central to studies on teaching. For example, classroom observations are commonly used in randomized trials to assess mediation effects through teacher practices. Both live and video observations have advantages and both may be used in a variety of applications. However, there is currently no data on whether these observation modes will yield similar scores and lead to the same inferences about teaching and teachers. Because video recording cannot capture all the information available in a classroom and because the new generation of observation protocols relies on rater judgments of complex practices, we have reason to believe the modes might not yield similar results. Our study is the first to directly test for mode effects in the CLASS-S protocol and provides critical information for the field of education research and practice.

Statistical, Measurement, or Econometric Model:

The study used cross-classified hierarchical linear models for the observation scores. We used these to estimate and test difference in the score distributions (means) across modes and to estimate the correlation scores across modes accounting for measurement error. We also used this model to conduct Generalizability studies, G-studies, (Brennan, 2001) of both the live and video scores. In these G-studies the classroom effect is the construct of interest. The variance in scores from different rating of lesson segments (X_{clsr} for classroom c, lesson l, segment s¸ and rater r) can be decomposed into fifteen components:

$$\sigma^2(X_{clsr}) = \sigma_c^2 + \sigma_l^2 + \sigma_s^2 + \sigma_r^2 + \sigma_{cl}^2 + \sigma_{cs}^2 + \sigma_{cr}^2 + \sigma_{ls}^2 + \sigma_{lr}^2 + \sigma_{sr}^2 + \sigma_{cls}^2 + \sigma_{clr}^2 + \sigma_{csr}^2 + \sigma_{lsr}^2 + \sigma_{resid}^2.$$

We combined some of these components to focus on sources of greatest interest: The lesson main effect and classroom by lesson variance components are combined into the "lesson" variance component ($\sigma_{lesson}^2 = \sigma_l^2 + \sigma_{cl}^2$).the overall segment effect, the segment by lesson, and the classroom by segment are combined with the classroom by lesson by segment interaction into what we refer to as the "segment" variance component ($\sigma_{seg}^2 = \sigma_s^2 + \sigma_{ls}^2 + \sigma_{cs}^2 + \sigma_{cls}^2$). The lesson by rater and the classroom by lesson by rater components into what we refer to as the "rater by lesson" component ($\sigma_{rater\ x\ lesson}^2 = \sigma_{lr}^2 + \sigma_{clr}^2$) and the segment by rater, classroom by segment by rater, lesson by segment by rater, and the "residual" components into an overall residual error component ($\sigma_{residual}^2 = \sigma_{sr}^2 + \sigma_{csr}^2 + \sigma_{lsr}^2 + \sigma_{resid}^2$).

We decomposed the variability in segment-level scores into component sources separately for domain and mode by estimating the variance components from a cross-classified linear mixed model with random effects for classroom, lesson within classroom, segment within lesson, rater, rater by classroom, rater by lesson, and residual error. We report each source's share of the total variance.

Usefulness / Applicability of Method:

CLASS-S is a commonly used observation protocol and shares many characteristics with other widely used protocol such as the Danielson Framework for Teacher. Our methods for decomposing the variance in scores and for comparing modes will be useful for other analysts studying classroom observation protocols.

Research Design:

The study is a field trial of the CLASS-S observation protocol. Lessons were scored on the CLASS-S protocol using live in-classroom observations and video observations. Mode effects were assessed by comparing live and video scores.

Data Collection and Analysis:

One classroom taught by each of 82 sampled teachers participated in the study. The 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 scores by to separate observers.

We first study trends in scores because live and video scores occurred on different days. The study then conducted descriptive analyses of score means by observation mode (live or video) by each of the CLASS-S domains. It tested for mode difference in the means using cross-classified hierarchical models to account for the nested structure of multiple scores from lessons, nested with classrooms and scored by multiple raters who crossed with classrooms.

The study also estimated correlations between scores from two modes for the same lesson or classroom. It also corrected these correlations for measurement error.

We then separately decomposed variance in scores from live and video observations via G-studies as described above. We used the results of the G study to estimate the standard error of measurement and the reliability of scores for lessons assuming they receive from 1 to 8 ratings and for classroom scores for a year under four possible data collections schemes: two lessons each scored one time by the same rater; two lessons each scored one time by two different raters; four lessons each scored one time by the same rater; and four lesson each scored one time with one rater scoring one lesson and separate rater scoring the other three.

Findings / Results:

Figure 2 shows there are distinct trends in scores across the school year. The trends are due to changes in the ratings as raters gain experience with the protocol.

Insert Figure 2 about here

Figure 3 shows that there were differences in the means across the domains with live scores tending to be somewhat higher for two CLASS-S domains. The differences persist even when we account for the differences in the scoring dates and the scoring trend.

Insert Figure 3 about here

The G study results suggest that lesson scores are more reliable for live scores for all three domains. For classrooms scores, live scores are more reliable for two of the domains but not for the Instructional Support domain.

Conclusions:

There are small mode effects on score means but they are relatively small and most likely inconsequential. Modes do not rank order teachers differently; it is the measurement error that results in differences between live observations and video scoring in the ordering of classrooms or lessons. Differences in the decomposition of variance across modes are a result of the differences in scoring dates.

Scoring trends and the differences in timing across the modes are the only significant difference between modes. They have important implications for studies using classroom observations. Live scoring will confound rater learning with lessons and video scoring can avoid this confound.

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.

Jaeger, R. M. (1993, April). Live vs. Memorex: Psychometric and practical issues in the *collection of data on teachers' performanc*es in the classroom. Paper presented at the meeting of the American Educational Research Association, Atlanta, GA. ERIC Document Reproduction Service No. ED360325. Retrieved December 21, 2011, from http://www.eric.ed.gov/PDFS/ED360325.pdf

Miller, K. (2007). Learning from classroom video: What makes it compelling and what makes it hard. In R. Goldman, R. Pea, B. Barron & S. J. Derry (Eds.), Video research in the learning sciences (pp. 321–334). Mahwah, NJ: Lawrence Erlbaum.

Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S. L., & La Paro, K. M. (2007). Classroom assessment scoring system manual, middle/secondary version. Charlottesville: University of Virginia.

Sherin, M., & Han, S. Y. (2004). Teacher learning in the context of a video club. Teaching and Teacher Education, 20, 163–183.

Van Es, E. A., & Sherin, M. G. (2010). The influence of video clubs on teachers' thinking and practice. Journal of Mathematics Teacher Education, 13, 155–176.

Appendix B. Tables and Figures

Domain	Dimensions	Dimension Description
Emotional Support	Positive Climate	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	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	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	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	encompasses the teacher's use of effective methods to encourage desirable behavior and prevent and redirect misbehavior
	Productivity	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	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	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	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	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 1. CLASS-S domains and dimensions.

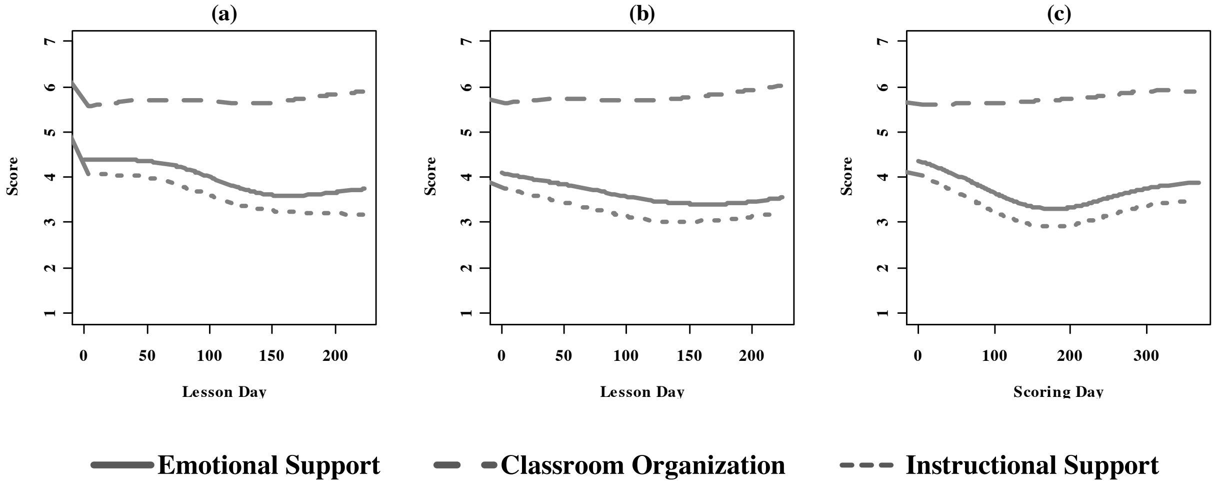

Figure 2. Time trends relative to the first day of data collection, by domain. (Emotional Support, solid line, Classroom Organization, dashed line, and Instructional Support, dotted line). Subplot (a) live observation scores by lesson date, (b) video observation scores by lesson date, and (c) video observation scores by date scored.

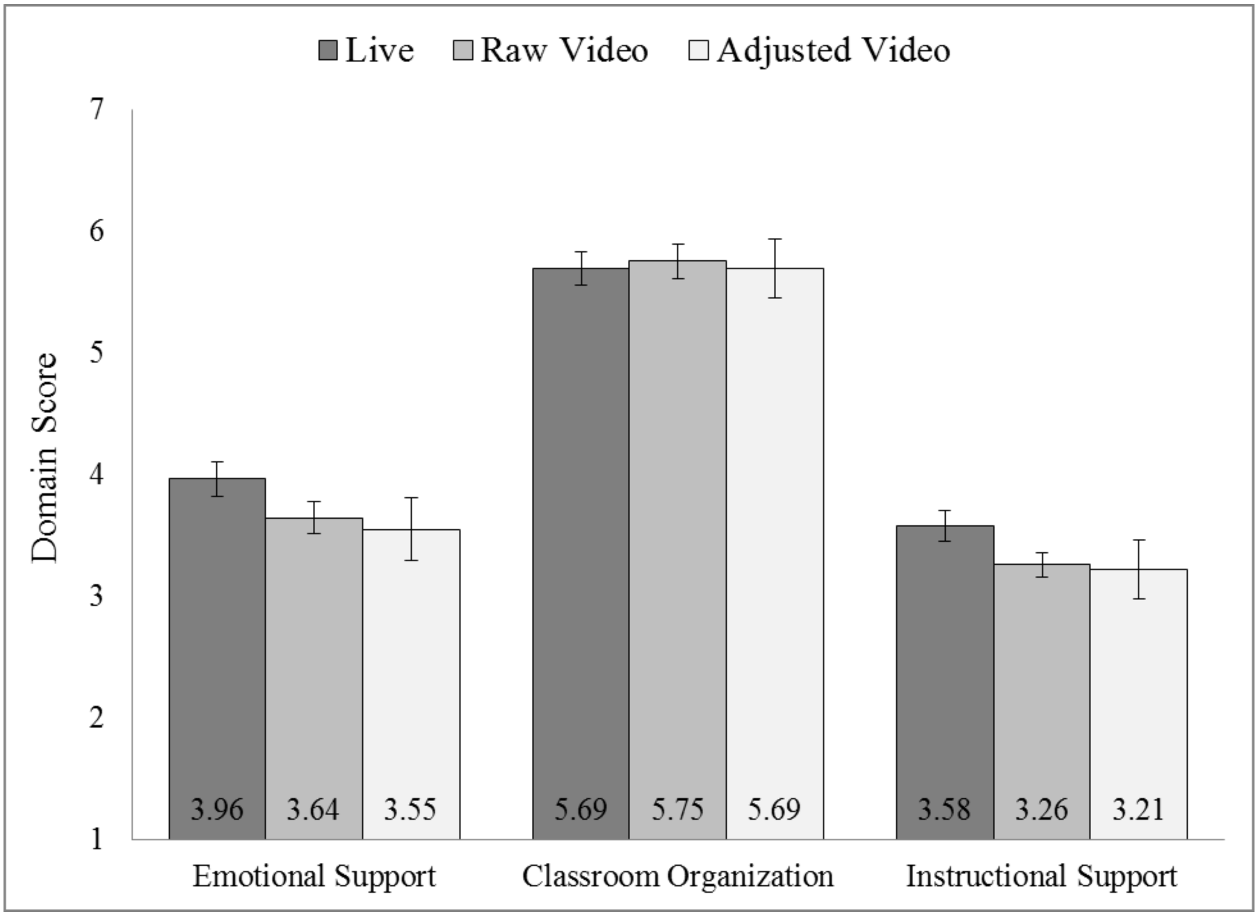

Figure 3. Means of domain scores by scoring mode: live scores, video scores, and adjusted video scores. Mean scores for each domain are given at the base of each bar and 95% confidence intervals are shown at the top of the bars.