Article

Effect of Observation Mode on Measures of Secondary Mathematics Teaching

Educational and Psychological Measurement
73(5) 757–783

DOI: 10.1177/0013164413486987
epm.sagepub.com

Jodi M. Casabianca[1], Daniel F. McCaffrey[2], Drew H. Gitomer[3], Courtney A. Bell[2], Bridget K. Hamre[4], and Robert C. Pianta[4]

Abstract

Classroom observation of teachers is a significant part of educational measurement; measurements of teacher practice are being used in teacher evaluation systems across the country. This research investigated whether observations made live in the classroom and from video recording of the same lessons yielded similar inferences about teaching. Using scores on the Classroom Assessment Scoring System–Secondary (CLASS-S) from 82 algebra classrooms, we explored the effect of observation mode on inferences about the level or ranking of teaching in a single lesson or in a classroom for a year. We estimated the correlation between scores from the two observation modes and tested for mode differences in the distribution of scores, the sources of variance in scores, and the reliability of scores using generalizability and decision studies for the latter comparisons. Inferences about teaching in a classroom for a year were relatively insensitive to observation mode. However, time trends in the raters' use of the score scale were significant for two CLASS-S domains, leading to mode differences in the reliability and inferences drawn from individual lessons. Implications for different modes of classroom observation with the CLASS-S are discussed.

[1]Carnegie Mellon University, Pittsburgh, PA, USA
[2]Educational Testing Service, Princeton, NJ, USA
[3]Rutgers, The State University of New Jersey, New Brunswick, NJ, USA
[4]University of Virginia, Charlottesville, VA, USA

Corresponding Author:
Jodi M. Casabianca, Carnegie Mellon University, Department of Statistics, Baker Hall, 232G, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA.
Email: jodicasa@andrew.cmu.edu

Keywords

mode of observation, video, classroom assessment, generalizability theory, teaching, Classroom Assessment Scoring System–Secondary (CLASS-S)

As the education reform movement increasingly focuses on teachers and teaching, educators, policymakers, and researchers need valid and reliable measures of teaching 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. 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. Renewed interest in measuring teaching and the large-scale use of observations for teacher evaluation systems has raised questions about the affordances of video capture, heightening the need for information on the comparability of scoring video and live observations. We present the first large-scale comparison of observation mode in the assessment of mathematics teaching.

Observations of teaching are viewed as very useful data sources about teaching quality because they provide assessments that incorporate not only observation of the teacher's teaching but also the level of student engagement, the cognitive complexity of student–teacher interactions, and the subject matter focus and depth of instruction (Erickson, 2006; Jaeger, 1993). Video recording of classrooms is an alternative with practical advantages (Brunvard, 2010), especially with recent technological advances in the capture and transmission of digital audio and video.

Extant research has shown very little difference in scores resulting from video and live observation. However, the studies were either not based on any rigorous evaluation (and were therefore inconclusive) or were conducted on data for nonclassroom contexts. Frederiksen, Sipusic, Gamoran, and Wolfe (1992) found live and video modes yielded scores with similar psychometric properties but evaluated just four teachers with two raters. A second study also found no differences in rater accuracy between modes of data collection (Ryan et al., 1995) but used data from on an assessment center group discussion exercise (not a typical classroom/teacher assessment). To our knowledge, there are no studies that comprehensively investigate the nature of mode differences in classroom observations.

This research considers how mode differences in both the distribution and precision of scores influence two possible inferences drawn from teacher evaluation scores: (a) inferences about the level of teaching in a classroom (an ''absolute'' reference) and (b) inferences about the ranking of teaching in classrooms (a ''relative'' reference). The first inference applies when comparing teaching to an absolute standard or cut point that relies on actual scale points. The second inference applies when considering a teacher's relative standing among other teachers within a school or district, or when considering the relative standing of schools, districts, or other institutions. Relative inferences also apply when studying the correlation between an

observation score and other measures as is done when studying the validity of measures or when using classroom observations to test for mediation effects of interventions, including professional development, on student achievement.

We make an additional distinction in the unit of measurement of classroom scores: the teaching for a single lesson or the teaching for a classroom over a school year. For example, the score on a single lesson might be used to provide very specific guidance to a teacher or scores on lessons may be of interest for studying the associations between attributes of the lesson, such as instructional topic or lesson format, and qualities of the teaching. Teacher evaluation systems, on the other hand, assess teaching for the year to provide feedback to the teacher and make human resource decisions.

This article is organized as follows: in the next section, we discuss the types of theoretical and practical differences in classroom assessment by observation mode. We then introduce the classroom measurement tool, Classroom Assessment Scoring System–Secondary (CLASS-S), which is used in this study. In the sections that follow, we describe our analytic approach and share the results of those analyses. We conclude with a discussion.

Classroom Assessment by Observation Mode: Video Versus Live

Jaeger (1993) identifies time sampling (when measurements occur during the timeline of interest), rater sampling (who evaluates the teaching as captured from any occasion), situational sampling (the sample of events occurring in a classroom that are used to assess the teaching and context), and mode (live vs. video) as potential sources of variance in the assessment of teaching.

Of particular interest is whether mode affects the psychometric quality of scores produced through observation. Live classroom observations are the conventional approach to evaluating teaching,[1] and 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, one component of validity.

Using video provides particular affordances because they create a permanent record. Video has been encouraged because teachers can review videos alone or in groups 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. 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. Of course, there are logistical and economic implications, but these are not the focus

of this study, in part because technologies and associated costs and implementation possibilities are rapidly evolving. Instead, the focus is on the quality of scores generated using two different modes of observation.

Video and live observations differ in the quality and nature of information available to an observer. One key difference between live and video observations concerns how visual information is captured. In live observation, the rater has the ability to scan the entire classroom at any time, focusing on particular aspects while also potentially being drawn to aspects of the classroom in the rater's periphery. In fact, there are no explicit scanning guidelines for an observer using CLASS-S (or other prominent observation protocols). For video, the camera setup constrains the focus so that any observer watching the same video will have the same information available in focus; fixing the view may contribute to minimizing measurement error and improving reliability.

A second key difference is audio capture. In live observation, an observer is likely to be able to hear teachers and students when in a whole-class instructional format. In addition, there is ambient audio information available to the observer. However, if a teacher is working with an individual student or small group of students, those conversations are likely lost to an observer sitting far away from them. For video observations, the ability to place a microphone on the teacher ensures that the teacher's voice will be heard regardless of instructional format, but there is much less ability to capture and attend to ambient sounds, unless additional microphones are placed around the room.

Recent research has discussed time trends in rater effects, specifically rater severity drift and changes in score scale category use (Harik et al., 2009; Leckie & Baird, 2011; Myford & Wolfe, 2009). Considering that live and video observations also differ in terms of the timing of scoring, time trends could also lead to mode differences. That is, since live observations are scored on the day of the lesson, they are confounded with effects from rater learning and experience and changes in the quality of classroom interactions over time. Videos can be scored at any time after the lesson date; while they are also susceptible to these confounds, the confounds can be mitigated since there is a gap between dates of the lesson and scoring.

Classroom Assessment Scoring System–Secondary

The CLASS-S framework conceptualizes classroom quality through a latent structure organizing specific teaching behaviors and student and teacher interaction patterns into dimensions tied to underlying developmental processes (Pianta, Hamre, Haynes, Mintz, & La Paro, 2007). The dimensions derive from three broad domains: *Emotional Support, Classroom Organization*, and *Instructional Support*.

CLASS-S is a modified version of CLASS, which was designed to capture aspects of Pre-K and elementary classroom interactions. CLASS-S measures similar dimensions of interaction as CLASS, but its behaviorally–anchored scale points and the

detailed descriptions of specific dimensions of classroom processes align with behaviors appropriate for supporting adolescent learning and development.

The CLASS protocol is widely used and shares many key characteristics with other observation protocols currently in use. The protocol begins with an observer developing a record of evidence from the classroom for some defined segment of time, typically without making any evaluative judgments. At the end of the segment, observers use a set of scoring criteria, or rubric, that typically includes a set of Likert scales to make both low and high inference judgments about specific dimensions of teaching based on the record of evidence. Those judgments result in numerical scores for dimensions that are aggregated to domain scores. Segment-level scores for dimensions and domains are aggregated to create lesson-level dimension and domain scores, respectively.

The measurement properties of the CLASS have been well studied (Pianta, La Paro, & Hamre, 2008). The measure has predicted relationships with student social and academic outcomes (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Bill and Melinda Gates Foundation, 2012; Burchinal et al., 2009; Hafen et al., 2012; Howes et al., 2008; Mashburn et al., 2008) and supports the proposed domain structure in empirical studies (e.g., Hamre & Pianta, 2005; La Paro, Pianta, & Stuhlman, 2004). Across these studies, researchers have documented that training and calibration procedures prescribed by CLASS can produce adequate levels of agreement between raters (e.g., Allen et al., 2011; Mashburn et al., 2008).

Mashburn, Downer, Rivers, Brackett, and Martinez (2011) conducted a generalizability study (Brennan, 2001) to explore the sources of variability in CLASS scores for elementary classrooms. They found sizable variance among raters, days, and the interaction of raters and days, making clear that a single observation by a single rater of a single day of instruction would lead to a very poor estimate of overall classroom quality. The Measures of Effective Teaching Project (Bill and Melinda Gates Foundation, 2012) decomposed the variance in a pooled sample of CLASS and CLASS-S scores for elementary and middle school mathematics and English language arts teachers from six urban school systems also finding that raters, lessons, and residual sources (including rater by lesson interactions) were large relative to the teacher so that reliability of scores from a single rating for evaluating classroom teaching was very low.

The CLASS has been used in studies with both live (e.g., Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009) and video (e.g., Allen et al., 2011; Reyes, Brackett, Rivers, White, & Salovey, 2012) observations. However, there is not yet research documenting how the mode of observation contributes systematic differences in scores or to the sources and size of measurement errors.

Research Questions

Given the two types of possible inferences made with teacher evaluation scores (i.e., level and ranking inferences), and the additional sampling consideration (one or multiple lessons), this study addresses two research questions:

1. Will the classroom assessment protocol *score* classrooms differently due to the mode of observation? Specifically,
   - Do raters use the seven-point score scale differently with live observations than with video observations?
   - How do sources of variance compare between scoring modes and what are the implications for measurement error in a score from one lesson or from the entire year?
2. Will the classroom assessment protocol *rank* classrooms differently due to the mode of observation? Specifically,
   - Do scores from different observation modes rank lessons differently? Do mean scores based on multiple lessons over a year rank classrooms differently?
   - Is the reliability of live and video observations affected differentially by various extraneous sources of variance?

Even though teaching evaluations are used to assign scores to teachers, we use the term classroom in our research questions rather than teacher as the target of inference when describing data that are summarized over lessons, since the quality of interactions is not only determined by the teacher but also by a host of contextual effects, including students, curricula, and school (Bell et al., 2012).

Method

Study Design

The study includes 82 algebra classrooms, each with a unique teacher who volunteered for the study, in a large urban fringe district that serves roughly 90% students of color and 55% students who are eligible for free or reduced price meals. Approximately two thirds of the classrooms were in high schools while the rest were middle school classrooms.

Data Collection. We collected four observations per classroom with roughly one measure per quarter for each classroom. A fifth observation was added for 80% of the classrooms ($n = 65$). Because of scheduling issues or changes of assignment, the project observed six sample classrooms fewer than the targeted four times: three classrooms were observed just one time, two were observed twice, and one was observed three times. Every observed lesson was rated by one or two live observers and video recorded.

Our time sampling of observation days captured nearly all of the school year (182 days from August to June) and observations occurred at similar times of the year for most classrooms. On average across classrooms, Observation Lesson 1 occurred on the 51st day of the school year with 50% of the lessons occurring between days 46 and 56 and all of the first observations occurring within a 2-week period. Observation Lessons 2, 3, and 4 occurred on average on the 75th, 106th, and 131st days of school with 50% of the lessons occurring within days 68 to 83, 95 to 115, and 123 to 138,

respectively, and for each lesson all observations occurred within a 30-day window. Observation Lesson 5 occurred, on average, on the 156th day of school, with 50% of the lessons occurring between days 149 and 161 of the school year and all the observations occurring within a 2-week period.

CLASS-S Scoring. CLASS-S is organized around three domains of teacher–student interactions: *Emotional Support*, *Classroom Organization,* and *Instructional Support*. Each domain is associated with three to four specific dimensions of teacher–student interactions (Figure 1). Dimensions are scored on a 1 to 7 scale according to specific behavioral indicators. Domain scores are derived from their associated dimension scores. Note that 1 of the 11 dimensions is not associated with a domain; the *Student Engagement* dimension refers to the extent to which students are actively engaged in classroom activity.

Procedures for Live and Video Scoring. 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 by domain, even though the observed domain scores tend to be moderately to highly correlated.

Raters and Training. Six raters, all former secondary public school teachers, were originally part of the study. However, very early in the study, one rater left the study, leaving the project with five raters who completed the vast majority of live observations and all of the video observations. The raters underwent extensive training including CLASS-S training, a certification test, weekly calibration tests, and conference calls to discuss calibration results.

Assignment of Raters[2]. We assigned raters to the lessons for live scoring using a design in which every pair of raters was assigned to lessons from roughly an equal number of classrooms. Loss of a rater and the addition of the fifth observation required adjustments to the initial design but the study retained the approximate balance in the rater assignments to lessons from classrooms. The design included double scoring of 20% of the live observations, which was the maximum number of double scores available given the project budget. For video coding, we again assigned raters to lesson to maintain approximate balance in the assignment of pairs of raters to the lessons from the 82 classrooms with the additional restriction that a rater would not score a video if she had rated the live observation. For both live and video scoring, the design also included one rater observing two different lessons from each

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.

classroom, which allowed for estimating a classroom by rater variance component in the generalizability study described below.

Timing of Scoring. Live scoring occurred when the lessons took place; video scoring occurred throughout the school year and into the summer that followed. In calendar days from the start of data collection, the day on which 25th, 50th, and 75th

percentiles of live and video scores were completed are 45, 122, and 172 for live, and 130, 234, and 286 for video. Across all lessons and scorings, the average number of days between the day of the live lesson and the scoring of the respective video was 106 days. Across all lessons, the average number of days between the first and second video scoring for the same lesson was 72. Note we use calendar days since the first day of scoring rather than the day of the school year to describe the timing of scoring since video scoring was not confined to school days. All dates in the remainder of the article refer to days since the first day of scoring.

Data Analysis

To evaluate mode differences in scoring trends, we tested for time trends as a function of the scoring date and then adjusted video scores as if they were scored on the same day as the lesson to compare to the means from live observations. To compare score levels by mode, we examined differences in means and distributions of the domain scores by observation mode. We used generalizability study results to compare modes in their sources of variance and used standard error of measures from Decision (D) study results to compare modes in their precision for a lesson and over a year. To compare modes in how they rank lessons and classrooms, we estimated correlations between mode scores at the lesson and classroom levels. We also compared mode ranking precision using reliability estimates from D studies under a variety of sampling plans.

Testing and Adjusting for Time Trends. Because video scoring and live observations occurred on different days, trends over the course of the study in the use of the score scale could contribute to mode differences in scores. Scores might systematically vary over time for one of two reasons. First, the actual quality of classroom interactions might change, and therefore, changes in scores reflect true variation in classroom quality. Second, observers may change in their rating behavior because of factors associated with additional experience and/or feedback on their scoring they received through calibration sessions.

For live scoring, it is impossible to disambiguate these two potential sources of score variation as they are completely confounded. When raters score lessons later in the school year they are also more experienced. These effects can be distinguished with video however.

To separate the effects of the timing of scoring from the different uses of the scale for live and video scores assigned on the same day, we first tested for trends in scores as a function of the day they were scored (the scoring date) and then used the model to estimate the mean scores for videos had they been observed on the day the lessons occurred rather than at a later date. We then compared the raw means from live observations to the adjusted video score means.

Testing for trends. To test for trends, we modeled lesson mean scores by domain and mode as functions of classroom and rater fixed effects and trends in the day the lesson occurred (live and video) and the day the lesson was scored (video only). We modeled the trends in lesson and score date using flexible nonparametric spline

smoothing via generalized additive models (Hastie & Tibshirani, 1990), which fit the data better than polynomial models for the trends. Specifically, letting y_{ilk} and y_{ivk} equal mean scores on a lesson from live or video scoring, our models are the following:

$$LIVE: \; y_{ilk} = \mu_{lk} + \gamma_{ilk} + \beta_{j(i,l)lk} + f_l(lesson\ date_i) + \varepsilon_{ilk},$$

$$VIDEO: \; y_{ivk} = \mu_{vk} + \gamma_{ivk} + \beta_{j(i,v)vk} + f_v(lesson\ date_i) + g_v(score\ date_{ij(i,v)}) + \varepsilon_{ivk},$$

where k denotes the three domains, μ_{lk} and μ_{vk} are overall means, γ_{ilk} and γ_{ivk} and $\beta_{j(i,l)lk}$ and $\beta_{j(i,v)vk}$ are classroom and rater fixed effects, $j(i, l)$ and $j(i, v)$ denote a rater who scored the lesson live or by video, f_l and f_v are smooth (nonparametric) functions of the date the lesson occurred (lesson date), g_v is a smooth function of the day the video was scored (score date) by rater j, and ε_{ilk} and ε_{ivk} are error terms. All days were defined as the number of calendar days since the first day of scoring for the study. The models include fixed effects for classroom and raters to improve the precision of estimates. Live observations occurred on the lesson date, so the live model only includes a term for lesson date. For video observations we can distinguish between trends in the teaching and trends in the scoring. To test for trends, we fit the models above with the smooth functions of lesson date or score date and compared them to reduced models that excluded the smooth functions for lesson or score date using a likelihood ratio test.

Adjusting video scores. We adjusted the video scores as if they were scored on the same day as the lesson. To obtain the adjusted means for the video scores, we again fit a generalized additive model for videos to the segment scores excluding the classroom and rater fixed effects ($y_{ivk} = \mu_{vk} + f_v(\text{lesson date}_i) + g_v(\text{score date}_{ij(i,v)}) + \varepsilon_{ivk}$). Using the results of this model, we calculated the expected (predicted) value for each video score by using the date of its corresponding lesson and its actual scoring date, or $E(y|\text{true lesson date, true score date}) = \hat{y}_{vk1}$. We subtracted this value from the actual observed score to obtain a residual for the score, $(y_{vk} - \hat{y}_{vk1})$.

We also calculated the expected (predicted) value for each video score had it been scored on the date its corresponding lesson occurred by using the model to estimate $E(y|\text{true lesson date, score date} = \text{lesson date}) = \hat{y}_{vk2}$. We added the residual to this estimated expected value to estimate the video score that the lesson would have received had it been scored on the same day that the live scoring occurred, the day the lesson occurred. We call this the adjusted video score, $y_{vk(adj)} = (y_{vk} - \hat{y}_{vk1}) + \hat{y}_{vk2}$, and we compared the mean of the adjusted video score to the raw mean of the live scores to check for the sensitivity of our estimates of mode effects to the difference in the timing of the scoring.

Testing for Mode Effects on the Use of the Score Scale. To test if raters used the scale score differently when conducting live observations than they did when doing video observations, we compared the distributions of scores on each of the 11 CLASS dimensions that raters assigned to segments using live observations to the corresponding distributions from video observations. We used scores on the dimensions

assigned to segments because these were the units at which raters used the score scale. We tested for overall mode differences in the score distributions for each dimension using a Cochran-Mantel–Haenszel test (Agresti, 2002) with segments as strata, which restricted the sample to only those segments scored under both modes.

We also tested for mode differences in the distribution of domain scores for segments using a Kolmogorov–Smirnov test. To account for matching by segment, we used a permutation test (Efron & Tibshirani, 1993) in which the mode labels of scores from the same segment were randomly permuted and the Kolmogorov–Smirnov statistic was recalculated using the permuted scores. We repeated the permutations 1,000 times to create the distribution of test statistics under the null distribution of no mode effects and estimated our p value for each domain as the proportion of the permutation sample that was greater than the statistic for the actual observed sample.

We estimated and tested for mode differences in the mean domain scores using a linear model fit to the pooled segment-level score data from both scoring modes. The model included an indicator for mode and segment fixed effects. We tested the null hypothesis that the coefficient on the indicator for mode equaled zero using a two-tailed test and repeated the test separately for each dimension score.

Generalizability Studies. Generalizability, ''G,'' theory uses an analysis of variance approach to partition a score into an effect for each facet or source of variability. 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). For inferences about teaching in a classroom, it would be preferable if classrooms accounted for a substantial proportion of score variation and factors like raters or specific lessons, and their interactions, did not. Other factors accounting for score variation might be temporal—when during the week or school year a lesson was observed or even the number of hours a given rater has spent scoring observations. Variation on such factors does not inform us about the general level of teaching in a classroom, and thus, we consider temporal sources as error and classrooms as the signal of interest. We use G theory to assess these various sources of variance. Because lessons have differing numbers of segments, we analyzed segment scores and included terms for the additional sources of variance in those scores.

We used a basic model[3] for decomposing the CLASS-S score $X_{clsr,dm}$ from a rating of one classroom (c) on one lesson (l), for one segment of the lesson (s) by one rater (r) for domain d, and mode m, live or video (Shavelson & Webb, 1991). For clarity of presentation, we drop the domain and mode subscript but we fit a separate model to the scores from each domain and both modes. To decompose the sources of variance in the segment scores, we fit the model

$$X_{clsr} = \mu + \mu_c + \mu_{l(c)} + \mu_{s(l)} + \mu_r + \mu_{cr} + \mu_{lr} + \varepsilon_{clsr},$$

where μ is the grand mean, μ_c is a random effect for the classroom, $\mu_{l(c)}$ is a random effect for the lesson nested within the classroom, $\mu_{s(l)}$ is a random effect for the segment nested with the lesson, μ_r is a random rater main effect, μ_{cr} is a random rater by classroom effect, μ_{lr} is a random rater by lesson within classroom effect, and ε_{clsr} is a residual error effect that includes rater by segment within lesson effects and unexplained error not captured in the other terms.[4] We model all the effects as random to estimate the contributions of variance from the various sources.

The classroom effect is the construct of interest. In G theory terminology, the classroom effect is the universe score or the average score for the classroom across all other sources of variance. The lesson within classroom effect captures variability among the average scores across ratings and segments for lessons from the same classroom and the segment within lesson effect captures the variability among average scores across ratings from segments from the same lesson. The rater effect captures the tendency of some raters to rate classrooms higher or lower than other raters. The classroom by rater interaction captures the tendency of a rater to judge the classroom differently from other raters accounting for the rater's main effect and the general level of teaching within the classroom. Another key component is the rater by lesson interaction that captures how raters differentially evaluate a specific lesson for a classroom given all the other tendencies for scores to be relatively high or low. Large differences in these means would suggest trouble in raters agreeing on the score of the same teaching.

Each of the seven components in the equation corresponds to a potential source of observed score variance that can be decomposed:

$$\sigma^2(X_{clsr}) = \sigma^2_c + \sigma^2_{l(c)} + \sigma^2_{s(l)} + \sigma^2_r + \sigma^2_{cr} + \sigma^2_{lr} + \sigma^2_{res}.$$

We decomposed the variability in segment-level scores into component sources separately for domain and mode by estimating the variance components from a 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.

To test for mode effects in the decomposition of variability, we pooled the data from both modes and fit linear mixed models with all the same random effects used in modeling the modes separately. The model included separate random effects for each source of error by mode but constrained the variance components to be equal across modes. We used a likelihood ratio to test the null hypothesis of equal distribution of sources of variance across modes by comparing the constrained model against a model that allowed for mode differences in variance components.

Decision Studies. D studies (Brennan, 2001) provide estimates of reliability using various potential scoring designs involving differing numbers of raters and lessons for each classroom. A D study[5] estimates reliability as the ratio of universe score variance (''true score'' variance of teaching among classrooms) to the total variance of the average of scores from multiple measurements (the universe score plus the

error variance for the average). For inferences about classrooms, we assume scores will be the average over multiple ratings by different raters on each of multiple lessons with multiple segments scored by each rater in each lesson. Hence the error variance equals:

$$\sigma^2_{error,\,class} = \frac{\sigma^2_l}{n_l} + \frac{\sigma^2_s}{n_l n_s} + C\left(\frac{\sigma^2_r}{n_r} + \frac{\sigma^2_{cr}}{n_r}\right) + \frac{\sigma^2_{lr}}{n_l} + \frac{\sigma^2_{res}}{n_l n_s}$$

where n_l is the number of lessons observed for the classroom, n_r is the number of unique raters who scored the classroom, n_s is the number of segments per lesson, and C is a constant that equals one when all raters observe the same number of lessons and equals 1.25 for the design in which one rater scores three lessons but another scores one (Design 4.2 described below).[6] The formula assumes each lesson will be scored only one time, which is true in all the designs we consider for assessing the teaching in a classroom, and for our calculations, we assume three segments for each design.

For inferences about a single lesson, we assume scores will be the average across all the ratings of the lesson so that error variance equals:

$$\sigma^2_{error,\,lesson} = \frac{\sigma^2_r}{n_r} + \frac{\sigma^2_{cr}}{n_r} + \frac{\sigma^2_{lr}}{n_r} + \frac{\sigma^2_{res}}{n_r n_s}$$

where n_r equals the total number raters who score the lesson. The true score variance for a lesson equals $\sigma^2_c + \sigma^2_l + \sigma^2_s / n_s$. Again we assume n_s equals three for all lessons for all designs.

We conducted a D study with four possible scoring designs for inferences about classrooms. We use the standard error of measures, the square root of $\sigma^2_{error,\,class}$, to evaluate the precision of estimates of the score level, and the reliability to assess the precision of rankings of classrooms. The candidate scoring designs for observations are the following: (a) two lessons observed each by one rater, the same rater scores both lessons (2.1); (b) four lessons observed each by one rater, the same rater scores all lessons (4.1); (c) two lessons observed each by one rater, one rater scores one lesson and a separate rater scores the other one (2.2); and (d) four lessons observed each by one rater, one rater scores one lesson and a separate rater scores the other three (4.2). Scoring designs 2.1 and 4.1 are most likely to occur in a traditional school setting where a principal, mentor, or peer observes a teacher's classroom multiple times over an academic year. Scoring designs 2.2 and 4.2 are most likely to occur in research studies that use multiple raters to improve reliability or in school settings that use both principal and peer mentors as raters.

For a single lesson, we calculated standard error of measures and reliabilities for each domain and mode combination assuming there were one to eight raters scoring the lesson (Scoring Designs 1.1 to 1.8).

To test for mode differences in D study reliabilities, we used a jackknife estimate of the standard error of the estimated reliability. That is, we removed all scores for

one classroom from both live and video samples, reestimated variance components by fitting mixed models to the reduced samples, and reestimated the reliabilities using the resulting variance component estimates, repeating this for each classroom. We estimated the difference between mode reliabilities using the reliabilities from each jackknife replicate. The estimated standard error in the difference in mode reliabilities equals the square root of the variability across the jackknife replicates in the estimates of this difference. We tested the null hypothesis of no difference in reliabilities across modes with a t test using the jackknife estimate of the standard error. We used a similar procedure for testing for mode differences in the standard error of measures.

Correlations. Although the relative magnitude of scores given in different modes is of interest, many current policy initiatives and research efforts are concerned with the ordering of classrooms. Therefore, we examined whether modes tended to order classrooms similarly by estimating the average domain scores for each lesson and each classroom (over a year of lessons) by mode and estimating the Pearson correlation coefficients between the scores from the two modes. We repeated the analysis using the adjusted lesson-level scores to ascertain the effects of differences in scoring date on our conclusions about correlations between mode scores.

Because of measurement error, two distinct sets of scores obtained using the same observation mode will have a Pearson correlation less than one. Our goal is to understand how observation mode further reduces the correlation. We do this by estimating the ''disattenuated'' correlation or the correlation between perfectly reliable scores obtained from each observation mode. To estimate the disattenuated correlation for each domain, we fit a linear mixed model to the individual scores from both modes including random effects for classroom, rater, lesson, segment, and interactions of all terms with mode plus a residual term. The model also included fixed main effects for mode. For inferences about classrooms, the disattenuated correlation equals the ratio of the variance component for classroom to the sum of the variance components for the classroom and classroom by mode. For inferences about lessons, the disattenuated correlation equals the ratio of the sum of the variance components for classroom and lesson to the sum of the variance components for the classroom, lesson, classroom by mode, and lesson by mode.

Results

Trends in Scoring

Figure 2 shows trends in domain scores by lesson date for live (Panel A) and video observations (B) and scoring date for video observations (C). There are notable trends in both live and video scores for *Emotional* and *Instructional Support* with scoring trending downward early in the school year and then leveling off. For both observation modes, *Classroom Organization* scores trend weakly upward across lesson dates but this trend is not significant. The trend in scoring day is similar, with scores trending downward early in the year but with a pronounced rise in *Emotional*

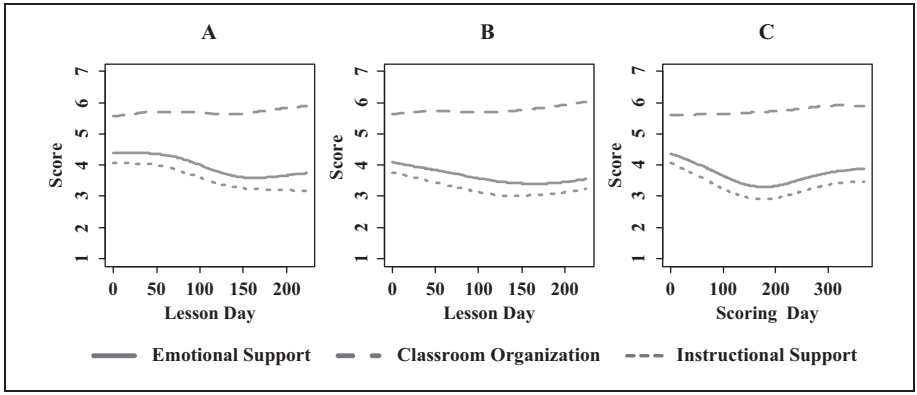

Figure 2. Time trends relative to the first day of data collection, by domain.
Note. Emotional Support, solid line; *Classroom Organization*, dashed line; and *Instructional Support*, dotted line. (A) Live observation scores by lesson date, (B) video observation scores by lesson date, and (C) video observation scores by date scored.

and *Instructional Support* scores for videos observed after about the 200th day of scoring.

Using the model that distinguishes the two time trends, we find significant trends in the date when videos were scored (scoring date) for *Emotional* and *Instructional Support* domains ($\chi^2 = 22.9$, $p = .001$, and $\chi^2 = 24.1$, $p = .003$, for *Emotional* and *Instructional Support*, respectively) but no significant trends for when the lesson actually occurred (lesson date). Neither scoring date nor lesson date trends were significant for *Classroom Organization* scores.

Raters are systematically changing how they use the score scale as they become more experienced as raters over time. It is likely that the changes by raters are similar for live and video scoring given the similarity in the trends for lesson date across modes. Therefore, given that video scoring decoupled lesson and scoring date while live scores did not, trends in the use of the score scale might contribute to differences in scores between observation modes. Mode effects may reflect, in part, differences in observer experience when the scores were assigned. We examine this possibility in the next set of analyses.

Level-Based Inferences

Mode Differences in Mean Scores and Distributions. Figure 3 shows the means and distributions of segment dimension scores for the full video ($N = 2{,}017$) and live ($N = 1{,}625$) samples.[7] Overall, the distributions are generally similar across modes, although mean scores for live observations were typically a little higher. Tests of significance for distribution and mean differences between modes were significant for all dimensions and domains except for *Negative Climate* and *Student Engagement*.

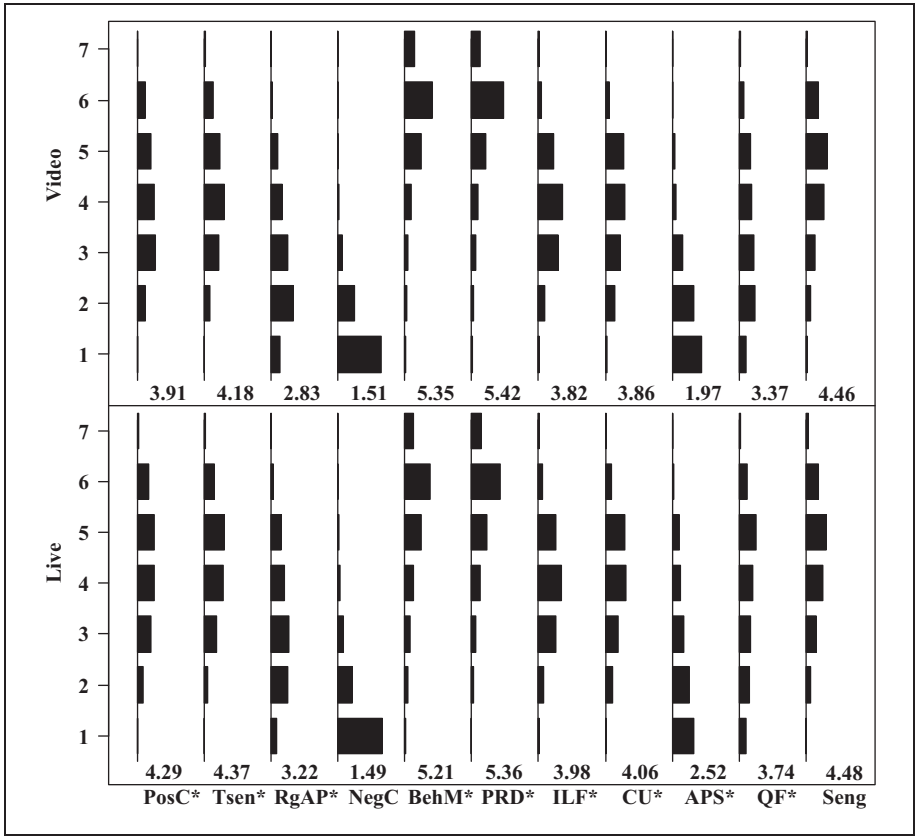

Figure 3. Distributions of scores by scoring mode (live observation vs. video lessons) and dimension.

Note. Bars show the proportion of the sample of segments scored at each of the seven scale points. Mean scores for each dimension are given below the bars. Dimension abbreviations are as follows: PosC = Positive Climate; Tsen = Teacher Sensitivity; RgAP = Regard for Adolescent Perspectives; NegC = Negative Climate; BehM = Behavioral Management; PRD = Productivity; ILF = Instructional Learning Formats; CU = Content Understanding; APS = Analysis and Problem Solving; QF = Quality Feedback; Seng = Student Engagement. * indicates that differences between modes in the means or score distributions are statistically significant. The difference in means and the difference between distributions are significant for same set of dimensions.

The domain scores were also significantly higher for live than video scoring in the *Emotional Support* (3.69 vs. 3.64) and *Instructional Support* domains (3.58 vs. 3.26) but differences on the *Classroom Organization* (5.69 vs. 5.75) were not significant. Adjusting for the timing and time trends in video made the means for video scoring slightly lower (3.53, 3.21, and 5.69 for *Emotional Support, Instructional Support*, and *Classroom Organization*, respectively) and did not change our conclusions about

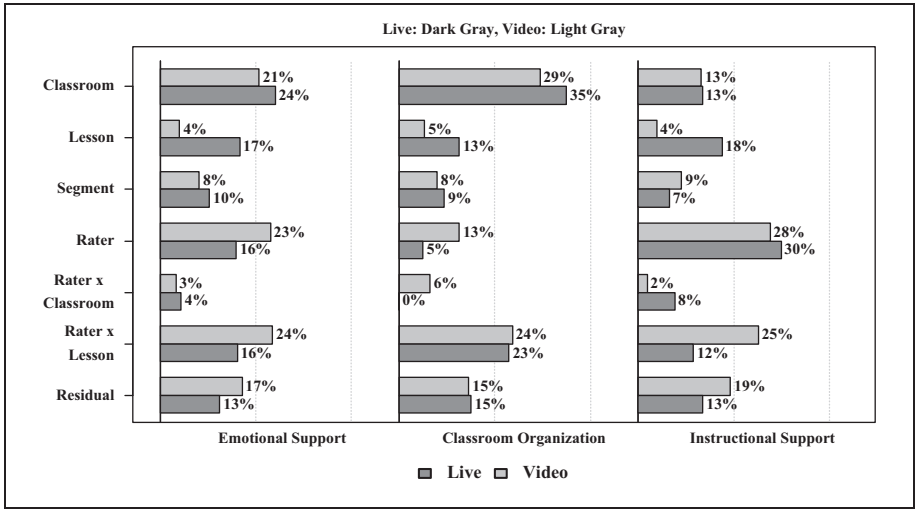

Figure 4. Decomposition of variability of scores into different sources by domain and mode.
Note. Bars represent the share of variability attributed to each source. Gaps between domains represent 50%, so bars that cover the entire range would indicate a source of variance accounting for 50% of the variability. Gray dotted lines indicate 20% and 40%. Actual percentages are given next to each bar.

mode differences: the effects of mode are generally small ranging from −.06 to .41 on a score scale that ranges from 1 to 7.

G Study Results. Figure 4 provides the decomposition of variability of domain scores for live and video observations.[8] Results show that variation in *Emotional Support* video scores was driven by rater main effects and interactions, while variation in live scores was mainly driven by classroom main effects. *Classroom Organization* video and live scores had similar variance decompositions with variation largely driven by classroom effects and rater interactions. Variation in *Instructional Support* video and live scores was driven by large rater main effects. Video scores also had a larger share of rater interactions and residual error while live scores had a larger share of lesson main effects. Likelihood ratio tests found significant mode differences in sources of variability for *Emotional Support* (LR = 21.6, p = .003) and *Instructional Support* (LR = 25.8, $p < .001$), but not *Classroom Organization*.

For all domains, variation attributable to lesson-level effects was larger for live scores than video scores. In addition, variation from rater interaction effects, specifically, rater by lesson effects (which are a combination of lesson by rater and the classroom by lesson by rater components), were always larger for video scores.

D Study Results. Table 1 provides standard error of measures (SEMs) for the four scoring designs for evaluating classrooms. The SEMs in scores for classrooms will be very large if only a single rater observes a teacher twice during the year (Scoring Design 2.1). It exceeds 0.8 for *Instructional Support* for scores from live observations

Table 1. D Study Estimates of Standard Error of Measures (SEMs) and Reliability Estimates for Classroom Measures from Four Designs.

	Design 2.1 (two observations conducted by the same observer)				Design 4.1 (four observations conducted by the same observer)			
	SEMs		Reliability		SEMs		Reliability	
	Live	Video	Live	Video	Live	Video	Live	Video
Emotional Support	0.69	0.74	0.37	0.32	0.60	0.66	0.44	0.37
Classroom Organization	0.49	0.57	**0.57**	**0.44**	**0.37**	**0.50**	**0.69**	**0.51**
Instructional Support	**0.83**	**0.73**	0.19	0.21	0.76	0.65	0.22	0.25
	Design 2.2 (two observations conducted by two separate observers)				Design 4.2 (four observations conducted by two observers, one conducts 3 the other conducts 1)			
	SEMs		Reliability		SEMs		Reliability	
	Live	Video	Live	Video	Live	Video	Live	Video
Emotional Support	0.60	0.62	0.44	0.40	0.50	0.53	0.54	0.48
Classroom Organization	0.46	0.49	0.59	0.51	0.34	0.41	**0.72**	**0.60**
Instructional Support	**0.67**	**0.61**	0.27	0.28	**0.60**	**0.52**	0.32	0.35

Note. Bold numbers indicate significant ($p < .05$) mode differences.

and is about 0.5 or greater for all domains on either mode. This SEM is very large relative to the 7-point scale; scores could easily move across scale points due to the errors. Under this scoring design, live scores have smaller SEMs for *Emotional Support* and *Classroom Organization* and a larger SEM for *Instructional Support*; only the difference for *Instructional Support* is significant ($p = .03$). However, the differences between modes are small relative to the overall large sizes of the SEMs. Increasing the number of lessons scored or using different raters to score some of the lessons reduces the SEMs but does not change the direction of mode differences and the overall SEMs remain large. There were also statistically significant mode differences in SEMs for *Instructional Support* for Scoring Designs 2.2 and 4.2 and a large difference (0.13, $p = .018$) in *Classroom Organization* SEMs for Scoring Design 4.1.

For inferences about lessons the SEMs are a function of the number of raters who score the lesson, assuming each rater will score it only one time. For one lesson scored by one rater observing a video (Scoring Design 1.1), the estimated SEMs were .83, .65, and .82, for *Emotional Support, Classroom Organization*, and *Instructional Support*, respectively. For live observations, the corresponding SEMs were .70, .54, and .82. With the addition of a second rater (Scoring Design 1.2), the SEMs fall to .59, .46, and .58, for *Emotional Support, Classroom Organization*, and *Instructional Support* for video observations, and to .49, .38, and .58 for live observations. Even with the addition of a rater, SEMs remain large relative to the score scale with live observations yielding somewhat more precise measures. All the estimated SEMs have large standard errors but the general patterns are stable—the SEMs are large and improve modestly with each additional rater. The mode differences in SEMs were statistically significant for the *Emotional Support* domain, for all scoring designs. However, mode differences are small relative to the large errors. It would require four raters using live observation and five using video observation for the SEMs for all three domains to be under .5.

Ranking-Based Inferences

Mode Differences in Correlations. Scores from raters using different observation modes result in large differences in the ordering of teaching across lessons. Pearson correlations between video and live domain scores for lessons were moderate: $r(333) = .48$ for *Emotional Support,* $r(333) = .63$ for *Classroom Organization*, and $r(333) = .33$ for *Instructional Support*. After adjusting video scores, correlations increased only slightly: $r(333) = .52$ for *Emotional Support,* $r(333) = .65$ for *Classroom Organization*, and $r(333) = .39$ for *Instructional Support*. However, scores from different observation modes order classrooms more consistently with Pearson correlations between video and live domain scores of: $r(82) = .80$ for *Emotional Support,* $r(82) = .86$ for *Classroom Organization*, and $r(82) = .74$ for *Instructional Support*.

Much of the observed instability at both the lesson and classroom level is due to the measurement error in scores obtained using either observation mode. After disattenuating these correlations, the relationships between live and video scores are

almost perfect; disattentuated correlations are either equal to, or just below, 1.0. Thus, variability between scores for a classroom or lesson across modes is about equal to what the variability would be for multiple scores using the same mode.

D Study Results. Table 1 also presents the reliability of classroom scores for the four scoring designs we considered in our D study. Consistent with the estimated SEMs, the reliabilities tend to be slightly higher for live observations than video scoring for *Emotional Support* and *Classroom Organization*, but not for *Instructional Support.* Scoring Designs 2.1, 4.1, and 4.2 had significant mode differences (0.12-0.17) in *Classroom Organization* reliabilities. For both modes the reliabilities tended to be low and the differences in the modes are small relative to the increases needed to obtain desired levels of reliability.

The reliability of inferences about the teaching for a single lesson was significantly higher for live observations than video observations for all three domains. For one lesson and one rater (Scoring Design 1.1), the estimated reliabilities were .33, .44, and .25 (video observations) and .52, .61, and .38 (live observations) for *Emotional Support, Classroom Organization*, and *Instructional Support*, respectively. With the addition of a second rater (Scoring Design 1.2), the estimated reliabilities increase to .50, .61, .40, for video observations, and .69, .76, and .55, for live observations. With four raters, the reliability of live scores exceeds .8 for *Emotional Support* (.81) and *Classroom Organization* (.86) whereas it is .71 for *Instructional Support.* To achieve these levels of reliability with video observations from this study would require eight raters!

Discussion

The need for high-quality measures of teaching is great. Policymakers and educators have increased their focus on teachers and teaching, and teacher evaluation systems in many states and districts now call for using scores from observations made using standardized protocols to support high-stakes decisions. The two available modes of capturing observation data each have affordances and limitations, as we have discussed. The question remains whether these modes yield measures with similar psychometric properties.

Our results suggest the scores from the two alternative observation modes do not have identical properties. Live observations yielded slightly higher scores and more reliable scores on two of the domains for inferences about classrooms and all three domains for inferences about individual lessons.

However, observations conducted on the same day by either mode yielded inferences that are highly similar. They rank ordered classrooms the same except for measurement error; the constructs they measure will have equal correlation with other measures and so they provide similar information. Live scores on the *Emotional* and *Instructional Support* domains were slightly higher than those from video observations but the difference was inconsequential. There was more sensitivity due to the raters when they used video observations and less variability due to the classroom,

so live scores were somewhat more reliable. However, both methods had large errors and low reliability for making inferences about the teaching in a classroom unless a large number of ratings was conducted on multiple lessons from multiple raters. The increase in reliability from using live observations did little to alleviate the shortcomings in the reliability of the scores. Given the conditions under which this study was conducted, none of these differences are likely to be substantial enough to influence the choice of observation modes over other concerns such as credibility, feasibility, costs, and so on. For example, even though video observations yield less reliable scores than live observations for the same designs, additional ratings of recorded videos is most likely less costly than observing additional lessons. Consequently, video observations may be more cost effective for achieving a specified level of reliability.

The Real Difference in Modes: Time Trends in Scoring

Live and video scoring, however, did have one difference that had implications for inferences about the teaching in individual lessons: live scoring must occur on the day of the lesson whereas video scoring can be decoupled from the day of the lesson. This affordance of video scoring was important in our study because raters changed how they used the score scale over the course of our study. For live observations, these changes in the scoring are conflated with true variation in teaching across lessons. Inferences about lessons from live observations will be distorted by the trend in raters' use of the scale. For video observations, raters scored the lessons at different times of the year so that trends in scoring were not conflated with the lesson.

Changes across the study in the raters' use of the score scale contribute to the lesson-to-lesson variance in teaching in addition to the true variability in teaching. But because the raters on any day would be consistent in their use of the scale, trends in the use of the scale do not contribute to rater-to-rater variability in live scoring. Hence for live scoring, changes in the use of the score scale inflate variability among lessons but do not affect rater variability creating reliable but inaccurate scores.

For video scoring, changes in the use of the score scale contribute to rater-to-rater variability in the scores for the same lesson since ratings occur on different days when the use of the score scale differs. But changes in the use of the score scale do not contribute to the lesson-to-lesson variance. Consequently, more ratings are needed to achieve reliable scores using video scoring than with live scoring. However, the reliability of the live scores comes with the cost of distorted measures. Statistical adjustments like the ones we used can further remove the effects of trends from video scores but they would not be possible with live scores. The trend in ratings did not affect classroom inference in large part because our study design observed nearly all participating classroom evenly across the school year.

Other observation efforts, including research studies or teacher evaluation programs, in which a cohort of raters starts with limited experience and their ratings evolve over time, may introduce time trends into their live observation systems. Such

observation efforts would benefit from the use of video scoring provided videos can be evaluated irrespective of the timing of the lessons.

Possible Sources of Trends in Ratings

What might be the causes for the observed scoring trends? Certainly, raters gain experience in scoring more observations, but clarifying the nature of that experience is critical. While we have limited data to investigate the scoring day trend, we hypothesize the trend is the result of two influences. First, our raters were former teachers. In general, teachers have not seen a lot of teaching practice outside their own classrooms. Therefore, some of the changes in score scale use may be the result of the raters renorming their underlying views of high-quality teaching. Scores in this study generally decreased over scoring days. This is consistent with raters indicating they were becoming more stringent in their views of good instruction over the study duration. A second possible influence may come from raters learning through repetition how to apply the scoring criteria to a range of different topics, instructional formats, activities, and learning goals. That *Classroom Organization* scores exhibited relatively high levels of reliability and were resistant to trends in scoring supports conclusions by Gitomer et al. (in press) that raters judge certain aspects of instruction more consistently than others and therefore raters stabilize in their scoring more quickly for the *Classroom Organization* domain.

Importantly, observers received ongoing feedback about the quality of ratings throughout scoring. This feedback occurred through calibration sessions once a week in which raters scored an observation also coded by a master rater and then focused on discrepancies in a discussion that was led by one of the study investigators. Thus, these observers did not simply score more videos, they received continuous feedback that was intended to facilitate observer learning.

These conditions have important implications and caveats for generalizing to the practice of evaluating teaching in accountability systems. First, many studies use a similar design with all observers starting with limited background and being trained and gaining experience as a cohort. Hence, our experience may be common in research. Second, we observe scoring trends that occur under conditions of experience *and* feedback. Whether we would have observed such trends in the absence of ongoing calibration activities is uncertain. Third, it is important to understand that scoring trends did appear to stabilize. Therefore, scoring trends may or may not be as influential over time given experienced raters in established evaluation systems. Finally, the observers in this study were completely independent of the teachers they were observing, a condition that does not exist in routine evaluation practices. All these differences mean that the approaches taken in this study need to be replicated under conditions of implementation in functioning evaluation systems.

What does this imply for potential differences in scores produced by different modes? The effect of scoring trends on mode differences may be most pronounced in research studies where raters have similar experience and training so that they are

all at the same point in the trend on every day of live observations but not for video observations. Confounding of rating trends and lesson-level scores could severely degrade studies measuring intermediate effects of various educational interventions; for this reason, video scoring might be preferable.

The implications are less clear for evaluation systems. In some systems, evaluators working at any given time may have varied levels of exposure to scoring so that experience and observation date are no more confounded in the live observations than in video observations. However, the variability in rater experience will remain a source of error and could result in lower reliability (under either observation mode) than what we estimate with our results.

For other evaluation systems, raters may have similar levels of experience. For example, states are rolling out observation systems with large-scale principal training sessions and follow-up calibrations that will result in principals with similar levels of experience and training, at least during the early years of the program. Peer evaluation systems like those used in Cincinnati, Toledo, and some other districts have plans for rotating peer evaluators. Depending on the plans for rotation, such programs might also create cohorts of raters with similar experiences that could confound experience with observation date and make the scores from live and video observations distinct.

Limitations

This study had some limitations related to the sample of classrooms and protocol that may limit the generalization of findings. First, the classrooms are a minority of the algebra classrooms in a single district, and they participated on a volunteer basis, though we found the sample and overall population of eligible classrooms to be very similar in terms of their characteristics and those of their students. Second, though algebra is viewed as a critical course for students' long-term academic and career success, the generalizability of our results to other courses remains unknown. Third, the study used a single observation protocol, CLASS-S, and so, how these findings generalize to other protocols is not yet known. Last, we scored only one class for each teacher and results for measuring teachers rather than a class per teacher may be different; however, it has been shown that section-to-section variance tends to be small (Bill and Melinda Gates Foundation, 2012).

Authors' Note

The opinions expressed are those of the authors and do not represent views of the Spencer Foundation, the William T. Grant Foundation, or the Institute of Education Sciences, U.S. Department of Education (the funding agencies).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by grants from the William T. Grant Foundation (9622), the Spencer Foundation (200900181), and the Institute of Education Sciences, U.S. Department of Education (R305B1000012 to Carnegie Mellon University).

Notes

1. We refer to teacher evaluation processes that are required as routine and administratively codified employment practice. There have been other efforts that have examined teaching using video (e.g., National Board for Professional Teaching Standards [NBPTS], Performance Assessment for California Teachers [PACT], Connecticut Beginning Educator Support and Training [BEST]), but they have been used to support other kinds of decisions (e.g., advanced certification and licensure).
2. A more detailed description of the procedure for assigning raters is provided online at http://cmart.stat.cmu.edu/suppmat_tucc.pdf
3. The full G study model equation is available online at http://cmart.stat.cmu.edu/suppmat_tucc.pdf
4. The lesson within classroom effect confounds lesson order main effects with the lesson by classroom interactions. The segment within lesson effect confounds the segment main effect with the segment by classroom interactions. Similarly, the lesson by rater effects confounds rater by lesson and rater by lesson by classroom interactions. We used this specification because the separate effects of lesson order, segments or their interactions would not affect our D study results and were not of interest for mode comparisons. Our study design does not yield multiple ratings from the same rater on a segment of instruction for a specific classroom so we cannot separately estimate rater by segment within lesson effects as it is confounded with the residual error term.
5. Traditional D study estimation holds certain elements constant under the assumption that some facets would be common to all classrooms—for instance, if all classrooms were reviewed by every member of the same panel of raters. However, we do not think that in practice classrooms will be scored using all common raters or lessons so we did not remove any of the sources of variance from our D study reliability calculations.
6. When four separate raters score the lessons for a classroom, the average of the scores for a classroom includes the average of four independent rater and rater by classroom effects so the variance of those effects is the variance component for each effect divided by four, the number of raters. When one rater scores three lessons and the other rater scores one lesson, the average score for the classroom includes three times the rater and classroom by rater effects for one rater and the rater and classroom by rater effect for the second rater all divided by four. The variance of these effects is 10 times the variance components for raters and classroom by raters divided by 16, which equals 1.25/2 or 1.25 divided by the number of raters. Because each lesson is scored by just one rater, the denominator for the residual variance is the number total number of scored segments equal to $n_l n_s$.
7. A table of mean differences and significance test results for dimensions and domains is provided online at http://cmart.stat.cmu.edu/suppmat_tucc.pdf
8. Decompositions of variability of dimension scores from live and video observation follow the patterns found for the corresponding domain scores and are available online at http://cmart.stat.cmu.edu/suppmat_tucc.pdf

References

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: John Wiley.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*, 1034-1037.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*, 62-87.

Bill and Melinda Gates Foundation. (2012, January). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

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

Brunvard, S. (2010). Best practices for producing video content for teacher education. *Contemporary Issues in Technology and Teacher Education*, *10*, 247-256.

Burchinal, M., Kainz, K., Cai, K., Tout, K., Zaslow, M., Martinez-Beck, I., & Rathgrab, C. (2009). *Early care and education quality and child outcomes*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Human and Health Services, and Child Trends.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Erickson, F. (2006). Definition and analysis of data from videotape: Some research procedures and their rationales. In J. L. Green, G. Camilli, & P. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 177-192). Mahwah, NJ: Erlbaum.

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., Gamoran, M., & Wolfe, E. W. (1992). *Video portfolio assessment: A study for the National Board for Professional Teaching Standards*. Berkeley, CA: Educational Testing Service.

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.

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (in press). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record.*

Hafen, C. A., Allen, J. P., Mikami, A. Y., Gregory, A., Hamre, B., & Pianta, R. C. (2012). The pivotal role of adolescent autonomy in secondary school classrooms. *Journal of Youth and Adolescence*, *41*, 245-255.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development*, *76*, 949-967. doi:10.1111/j.1467-8624.2005.00889.x

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, *46*, 43-58.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London, England: Chapman & Hall.

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 study. *Educational Researcher*, *41*, 56-64.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R. M., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, *23*, 27-50.

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

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *Elementary School Journal*, *104*, 409-426. doi:10.1086/499760

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.

Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2011). *Improving the power of an experimental study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes*. Unpublished manuscript.

Mashburn, A. J., Pianta, R. C., Hamre, B., Downer, J., Barbarin, O., Bryant, D., . . . Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*, 732-749.

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.

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.

Myford, C. M., & 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.

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.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system, manual, K-3*. Baltimore, MD: Brookes.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology, 104,* 700-712.

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology*, *45*, 958-972.

Ryan, A. M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T., & McCormick, S. (1995). Direct, indirect, and controlled observation and rating accuracy. *Journal of Applied Psychology*, *6*, 664-670.

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

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

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.