Article

Educational and Psychological
Measurement
2014, Vol. 74(3) 400–422
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164413515882
epm.sagepub.com
SAGE

The Effect of Observation Length and Presentation Order on the Reliability and Validity of an Observational Measure of Teaching Quality

Andrew J. Mashburn[1], J. Patrick Meyer[2], Joseph P. Allen[2], and Robert C. Pianta[2]

Abstract

Observational methods are increasingly being used in classrooms to evaluate the quality of teaching. Operational procedures for observing teachers are somewhat arbitrary in existing measures and vary across different instruments. To study the effect of different observation procedures on score reliability and validity, we conducted an experimental study that manipulated the length of observation and order of presentation of 40-minute videotaped lessons from secondary grade classrooms. Results indicate that two 20-minute observation segments presented in random order produce the most desirable effect on score reliability and validity. This suggests that 20-minute occasions may be sufficient time for a rater to observe true characteristics of teaching quality assessed by the measure used in the study, and randomizing the order in which segments were rated may reduce construct irrelevant variance arising from carry over effects and rater drift.

Keywords

observational measures, generalizability theory, reliability, validity

[1]Portland State University, Portland, OR, USA
[2]University of Virginia, Charlottesville, VA, USA

Corresponding Author:
Andrew J. Mashburn, Psychology Department, Portland State University, PO Box 751, Portland, OR 97207-0751, USA.
Email: mashburn@pdx.edu

Teaching observation measures are growing in popularity as a method to evaluate the quality of teaching. The increased use of these measures is largely due to recent shifts in education policy. For example, Race to the Top encourages the use of teaching observation measures in conjunction with other measures of teacher performance such as value-added (U.S. Department of Education, 2009). State and district policies are also shifting toward formal observation systems, with 24 states and the District of Columbia requiring observations as components of yearly teacher evaluations (Heitin, 2011). Teaching observations have a long history of use in education research for purposes of identifying characteristics of classroom settings that are associated with student learning (Bell et al., 2012; Bill and Melinda Gates Foundation, 2012; Mashburn et al., 2008). Observational measures are also part of new models of teacher professional development that use video observations to provide feedback and support to teachers in ways that lead to improved student learning (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Fritz & Chen, 2013; Mashburn, Downer, Hamre, Justice, & Pianta, 2010). Given the emphasis on teaching observations in policy, research, and professional development, it is no surprise that the psychometric characteristics of these measures are receiving increased scrutiny through large-scale studies such as the Measures of Effective Teaching project (see Bill and Melinda Gates Foundation, 2012).

Methods for conducting teaching observations vary across different instruments and across studies using the same instrument, and the methods that are adopted may affect the reliability and validity of scores. For example, methods of assessing the quality of teaching during a day may involve short observations with frequent ratings or long observations with infrequent ratings. Modes for collecting data may involve either live observations in the classroom or observations of videotaped lessons, and in the case of videotapes, observation segments may be presented either in sequential order as they occurred in real time or in random order. To examine the impact of length of observation and order of presentation on the reliability and validity of scores, we conducted an experimental study in which we manipulated the length and order of videotaped lessons of secondary grade classrooms to compare the reliability and validity of scores on the Classroom Assessment Scoring System–Secondary (CLASS-S; Pianta, Hamre, Hayes, Mintz, & LaParo, 2008). The goal was to identify observation procedures that maximize score reliability and the validity of score inferences.

Observational Measures in Education

Observational measures are available that assess various aspects of the quality of teaching; some of which assess the quality of teaching in a particular subject matter whereas others assess teaching more generally. For example, two content-specific measures are the Mathematical Quality of Instruction (MQI; Hill et al., 2008) instrument that assesses five domains of teaching mathematics to students in kindergarten through 8th grade classrooms, and the Protocol for Language Arts Teaching

Observation (PLATO; Grossman et al., 2010) tool that measures four components of the quality of English language arts instruction. Other observation measures assess general dimensions of teaching quality regardless of the subject area. For example, the Framework for Teaching (FFT; Danielson, 2011) measures the general quality of instruction in four domains; two of which are measureable through teaching observations, and two require additional information about a teacher's planning and professional behavior. The Classroom Assessment Scoring System (CLASS; Pianta et al., 2008) is a family of observational measures that tap into the quality of teacher–student interactions related to three domains—Emotional Support, Classroom Organization, and Instructional Support (Hamre et al., 2013). Interestingly, despite their differences, scores from subject specific and general observational measures tend to be highly correlated (Bill and Melinda Gates Foundation, 2012).

One characteristic shared by all these measures is that one or more trained raters observe a classroom or a videotaped lesson for a given period of time and provide scores reflecting the quality of teaching during that segment of time. For example, in the Measures of Effective Teaching project, raters judged videotaped classrooms using the CLASS-S and FFT (see Bill and Melinda Gates Foundation, 2012), and operational procedures for each of these measures during the project involved raters viewing and scoring the first 15 minutes of a lesson and then viewing and scoring a subsequent 15 minutes of the same lesson. However, in other studies using these measures, the protocols for dividing lessons into observation segments has varied. For example, prior use of the CLASS-S has involved dividing 40-minute lessons into two 20-minute segments (Mikami, Gregory, Allen, Pianta, & Lun, 2011), and standard protocols for the FFT involve a single rating of an entire 40- to 50-minute. Thus, the operational procedures for dividing a lesson into one or more segments vary across instruments and across studies using the same instrument. As we discuss in the subsequent section, these operational procedures regarding observation length can affect the score reliability and validity.

Framework for Reliability and Validity

Kane (1982, 2011) proposed a sampling model for validity that focuses on the accuracy of using an observed score to make an inference about a universe score. His framework derives from generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) and makes an explicit connection between reliability and validity. He explains that large amounts of random variation in observed scores (i.e., measurement error) can be reduced in three possible ways, but each one entails a tradeoff. Measurement error can be reduced through (a) more complete sampling of the universe (e.g., longer tests, more raters), (b) restricting the target universe, or (c) standardizing the measurement procedure. Of these three approaches, the last two involve a tradeoff between reliability and validity. Restricting the target universe is an extreme solution that narrows to the universe to a specific set of conditions. It improves reliability by eliminating it altogether through a narrow definition of the universe. However, this narrow universe is far from the target universe, which ultimately leads to biased scores and inaccurate inferences. Standardization is a less extreme solution in which only a few facets of the universe are fixed to specific values. It improves reliability by eliminating some, but not all, sources of measurement error. Standardization still narrows the universe and biases inferences, but to a lesser extent than a complete restriction of the universe. Thus, reliability and validity are directly connected in Kane's framework. Procedures that improve reliability often come at the expense of a loss of validity.

Kane (1982) viewed his sampling model and generalizability theory to be one method for providing evidence of construct validity. It is a complement to other methods and concepts in validity theory such as content and criterion-related validity, not a replacement of them. Indeed, Marcoulides (1989b) demonstrated the way convergent and discriminant validity can be explored through Kane's framework. A benefit of Kane's approach is that it makes an explicit connection between reliability and validity, and it provides a way of using generalizability theory to explore both. In the context of teaching observations, most researchers focus on reducing measurement error by sampling more of the universe. They fail to consider the way standardization improves reliability and biases inferences about the measured trait. Two features of teaching observations that are often standardized include the length of observation and the presentation order.

Observation Length and the Reliability and Validity of Scores

The length of time and frequency in which a rater observes a lesson before assigning scores can vary, and arguments can be made both for and against shorter, more frequent ratings and longer, less frequent ratings. In terms of validity, shorter ratings may be less subject to primacy and recency effects (Ebbinghaus, 1913) than longer ratings. For example, a single rating culminating at the end of a 40-minute observation will likely give undue weight to events that transpired in the first (primacy) and last (recency) 10 minutes of an observational period, and events that occurred during the middle 20 minutes are more likely to be forgotten and less likely to be incorporated into ratings. Primacy and recency effects can cause scores from a bad and subsequently good performance to be higher than scores from two good performances (Leventhal, Turcotte, Abrami, & Perry, 1983). Thus, shortening the observation from one 40-minute period to two 20-minute periods may improve validity by reducing construct irrelevant sources of variance such as primacy and recency effects.

A point of diminishing returns may be reached with respect to validity, however, if the observation time period is too short. This may be particularly salient when measuring complex, dynamic interactions between a teacher and students that are often the constructs of interest for teaching observations. In this case, an observational period may be too short for the desired classroom interaction to occur, and the resulting ratings will not permit a valid inference about the measured construct. As a result, excessively short observational periods can result in construct underrepresentation and compromise the measure's validity.

Selecting an appropriate length of observation not only affects validity but also reliability. Dividing an event into more occasions may improve reliability by increasing the number observations that can occur during a set amount of time. If one focuses strictly on reliability and chooses excessively short, but very frequent observational periods, high reliability may be achieved at the expense of valid inferences. Conversely, if one focuses on validity, the optimal level of reliability may not be attained. In sum, there may be tradeoffs between reliability and validity whereby choices that improve validity may reduce reliability, and vice versa, and a challenge is developing observational procedures that maximize both reliability and validity.

Presentation Order and the Reliability and Validity of Scores

Teaching observation measures may be used during live lessons where the raters are present in the classroom while the lesson is taking place. Or, they may occur long after the lesson has taken place through videotaped footage of the lesson. Each mode has its benefits and limitations. Live observations allow raters to hear conversations and notice interactions that would otherwise be inaudible or hidden from view in a video recording, which benefit the validity of scores; however, there are limits to the number of raters who can be physically present in the classroom, which may compromise the reliability. In contrast, with videotaped lessons, there are no limits on the number of raters who can observe and rate each lesson. Error attributable to rater effects can be mitigated by increasing the number of raters; however, a video recording of the classroom may omit or not fully convey the quality of teaching within the classroom.

Another potential advantage of observing classrooms using videotaped lessons that is relevant to the current study is that the order in which segments are presented to raters can be manipulated. Whereas live observations require that lessons be viewed and rated in sequential order such that ratings for the first part of a lesson and followed by ratings for the second part, segments from video observations may be viewed in a random order whereby parts of lessons from any day and any teacher can be randomly presented to a rater. Manipulation of the order in which raters view lesson segments may reduce other sources of construct irrelevant variance including carry over effects (Ho & Kane, 2013) and rater drift (Casabianca & Lockwood, 2013). Carryover effects occur when the scores of one segment are not independent from the scores of another segment. In the case of sequential coding, segments from a specific lesson by a specific teacher are rated in back-to-back fashion; thus, the scores from a subsequent lesson are likely to be affected by the events from or impressions left by the occurrences during the prior segments. When the ordering of segments are presented randomly, such that raters view segments randomly drawn from the lessons of all teachers, the carryover effect from one segment to the next within a lesson for a given teacher can be mitigated.

Rater drift occurs when rater performance lacks invariance over time (Congdon & McQueen, 2000) such as raters changing their use/interpretation of a scoring rubric over the course of a rating period (Casabianca & Lockwood, 2013). It is a source of construct irrelevant variance because teacher scores are systematically affected by changes in rater behavior. Casabianca and Lockwood describe a statistical model for controlling rater effects, but random presentation of segments may also reduce the influence of rater drift as well as the effect of other extraneous variables.

Study Purposes

A major question for anyone wishing to implement observations for purposes of policy or research is, ''How can I best allocate resources so as to minimize costs while producing reliable scores that permit valid inferences about teaching quality?'' Manipulating the length of observations and presentation order of segments may affect the score reliability and the degree of validity evidence without affecting the financial costs of observing and rating the quality of instructional activities. Our study aims to experimentally test the effect of observation length and presentation order on the score reliability and the degree of validity evidence supporting teaching observations. Specifically, eight trained raters were randomly assigned to rate 40-minute videotaped lessons either in one 40-minute occasion, two sequential 20-minute occasions, four sequential 10-minute occasions, or two nonsequential 20-minute occasions. The purpose of this study was to compare the reliability and predictive validity of a teaching observation measure and explore other potential threats to validity using experimental conditions that represent different ways to fix observation length and presentation order.

Method

Videotaped Lessons

For the purpose of this study, we obtained a subset of data collected from teachers and students from Grades 6 through 11 as part of the efficacy study for the My Teaching Partner professional development program for secondary school teachers (Allen et al., 2011). The study involved eight schools from the southeastern United States with random assignment of teachers within schools to either a treatment or control condition. In total, the efficacy study involved 47 teachers in the treatment condition and 43 teachers in the control condition. During the course of the efficacy study, teachers videotaped 40-minute classroom lessons on multiple days throughout the academic year and submitted videotapes to researchers during predetermined windows of time. For this study, we retained 47 of these teachers who met the following two criteria: submitted at least one videotaped lesson during each of the following three time periods: (a) September to November, (b) December to February, and (c) March to May; and had at least one lesson during each time period that was 40 minutes in length or more and without audio or video problems. In cases when a teacher had two or more 40-minute lessons available during a time period, we randomly selected one of them. The resulting sample of videotaped lessons included a total of 141 (47 teachers, 3 lessons each) videotaped lessons.

In addition to videotaped lessons from teachers, demographic information and scores on state achievement tests were available for a total of 1,366 students enrolled in study teachers' classes. Specifically, demographic characteristics of children included gender, minority status, and grade level. Achievement test scores in reading and math were collected during the prior school year and at the end of the school year during which the video lessons were collected.

Measure of Teaching Quality

The CLASS includes three observational measures that span pre-kindergarten through 12th grade. We used CLASS–Secondary (CLASS-S; Pianta et al., 2008) in this study to assess the quality of teacher–student interactions in middle and high school grades. The measure comprises 11 dimensions (i.e., items) that tap into three domains: (a) emotional support (EMSUP), (b) classroom organization (CLORG), and (c) instructional support (INSUP). Each dimension contributes to scores on one domain only and is rated on a 7-point scale. Anchor point descriptions for each dimension guide raters in selecting an appropriate score level. Factor analysis studies support the three domain structure of the measure (Bell et al., 2012; Malmberg, Hagger, Burn, Mutton, & Colls, 2010). However, the domains tend to be highly related, and models that take into account the nested structure of rating data suggest that a three factor or single factor model at the teacher level are plausible (McCaffrey, Yuan, Savitsky, Lockwood, & Edelsen, 2013; Savitsky & McCaffrey, 2013).

CLASS-S raters went through a formal training and certification period that required them to reach 80% agreement with master benchmark ratings. Certified raters return for additional training and calibration at a later date. The official protocol for CLASS-S requires raters to view a lesson for about 15 minutes, provide ratings, view the next 15 minutes of the same lesson, and provide another set of ratings. The ratings for each 15-minute segment are averaged to produce a score for the lesson. Rater agreement tends to be high in operational scoring (Bell et al., 2012) and generalizability studies indicate an index of dependability (i.e., phi-coefficient) that ranges between 0.5 and 0.63 when scores are averaged over four lessons (Bill and Melinda Gates Foundation, 2012).

Experimental Conditions

We designed an experiment to study the effect of observation length and order of presentation on the reliability and validity of CLASS-S scores. We randomly assigned two raters to one of four conditions. The first condition had raters judge a single 40-minute lesson. We call this condition the 1 × 40 condition. The second condition divided the lesson in half and raters observed the first 20 minutes, rated the quality of teacher–student interactions, observed the next 20 minutes of that lesson, and provided another rating. We refer to this condition as the 2 × 20 ordered condition. The third condition also involved two 20-minute segments of videotapes. However, in this case we randomized the order of presentation; raters watched 20-minute segments in random order and rated each segment. Randomization was at the segment level, not the teacher level. As such, it was unlikely for raters to observe all 40 minutes of a classroom in consecutive order. They may have watched the second 20 minutes of a lesson and then seen videotapes from other teachers before rating the first 20 minutes. We refer to this condition as the 2 × 20 random condition. Finally, we included a condition with four 10-minute segments assigned to raters in an ordered fashion. This fourth condition was like the 2 × 20 ordered condition, but it involved shorter and more frequent segments. We refer to this condition as the 4 × 10 ordered condition.

According to Kane's (1982) sampling model of validity, these experimental conditions could be considered facets of the target universe and evaluated as variance components in a generalizability study. This would be the desired approach if these experimental conditions represented random facets. However, teaching observations typically treat observation length and presentation order as fixed facets. Therefore, we conducted a separate generalizability theory analysis for each condition. To determine if our experimental conditions represent important sources of variance, we compared results from each condition using bootstrap procedures.

Analysis

Generalizability theory (Cronbach et al., 1972) provides the analytic framework for examining reliability of scores for each of the four study conditions. Generalizability theory is well-suited to teaching observations that are influenced by multiple sources of variance (e.g., raters, segments, lessons, and teachers), and it has been applied to observational measures of education settings in order to specify and estimate salient sources of variance in observed scores and to identify observation procedures that minimize sources of error and optimize the reliability of the measure (e.g., Erlich & Borich, 1979; Hintze, 2005; Marcoulides, 1989a; Mashburn, Downer, Rivers, Brackett, & Martinez, 2013; Meyer, Henry, & Mashburn, 2011).

In applying generalizability theory, variance components are estimated in a generalizability study (G-study), and these components are combined to estimate error variance and reliability in a decision study (D-study). In the design of this G-study, we treat teachers as the object of measurement, with lessons (l) nested within teachers ($l : t$). Arguments for this design instead of a crossed design may be found in Hill, Charalambous, and Kraft (2012) and Erlich and Borich (1979). Note that we obtained each lesson from one of three different time periods. Therefore, the lesson facet not only represents different lessons but also three different points throughout the academic year. We assume that lessons are exchangeable within each teacher, regardless of when the lesson was observed. In three of the experimental study conditions, we further subdivided each lesson into multiple segments. This design resulted in another level of nesting such that we had segments nested within lesson within teacher ($s : l : t$).

In the three study conditions for which we created multiple segments for each lesson (the two 2 × 20 conditions and the 4 × 10 condition), two raters (r) observed every segment of each lesson for every teacher. Therefore, raters were crossed with the other facets of the design. Considering the object of measurement (teachers) and all three facets, the universe of admissible observations corresponds to a design with raters crossed with segments nested within lessons within teacher or a $r\times(s : l : t)$ design. For the 1 × 40 condition in which there was a single 40-minute lesson, segment is not a facet and the design reduces to $r\times(l : t)$.

We estimated variance components for the universe of admissible observations using the MIVQUE estimation method as implemented in SAS version 9.2. In the D-Study, we then computed relative error[1] variance and the generalizability coefficient (i.e., a reliability estimate) for the universe of generalization.

The generalizability coefficient is $E\rho^2 = \sigma^2(\tau)/[\sigma^2(\tau)+\sigma^2(\delta)]$, which depends on the specification of relative error variance, $\sigma^2(\delta)$. In the $r\times(S : L : t)$ design, relative error variance is given by

$$\sigma^2(\delta) = \frac{\sigma^2(tr)}{n'_r} + \frac{\sigma^2(l : t)}{n'_l} + \frac{\sigma^2(s : l : t)}{n'_s n'_l} + \frac{\sigma^2(r\times[l : t])}{n'_r n'_l} + \frac{\sigma^2(t\times[s : l : t])}{n'_r n'_s n'_l}, \qquad (1)$$

where the symbol n denotes the decision study sample size for the facet indicated by the subscript. This expressions reduces to

$$\sigma^2(\delta) = \frac{\sigma^2(tr)}{n'_r} + \frac{\sigma^2(l : t)}{n'_l} + \frac{\sigma^2(r\times[l : t])}{n'_r n'_l}$$

for the $r\times(L : t)$ design.

We can determine the best approach for minimizing error variance in the $r\times(S : L : t)$ design by evaluating Equation 1 in the limit of each facet. The limits for increasing lessons, segments, and raters are

$$\lim_{n_l \to \infty} \sigma^2(\delta) = \frac{\sigma^2(tr)}{n'_r}, \qquad (2)$$

$$\lim_{n_s \to \infty} \sigma^2(\delta) = \frac{\sigma^2(tr)}{n'_r} + \frac{\sigma^2(l : t)}{n'_l} + \frac{\sigma^2(r\times[l : t])}{n'_r n'_l}, \qquad (3)$$

and

$$\lim_{n_r \to \infty} \sigma^2(\delta) = \frac{\sigma^2(l : t)}{n'_l} + \frac{\sigma^2(s : l : t)}{n'_s n'_l}. \qquad (4)$$

Assuming that all variance components are greater than zero, the following statements can be made about the effect of facet sample size on relative error variance.

First, notice that Equation 3 contains the variance component in Equation 2 plus two additional components. Therefore, increasing the number of lessons will always result in lower relative error variance than increasing the number of segments. Second, increasing the number of lessons will produce lower relative error variance than increasing the number of raters when

$$\frac{\sigma^2(tr)}{n'_r} < \left(\frac{\sigma^2(l:t)}{n'_l} + \frac{\sigma^2(s:l:t)}{n'_s n'_l}\right). \tag{5}$$

Finally, increasing the number of segments will produce lower relative error variance than increasing the number of raters when

$$\left(\frac{\sigma^2(tr)}{n'_r} + \frac{\sigma^2(r \times [l:t])}{n'_r n'_l}\right) < \frac{\sigma^2(s:l:t)}{n'_s n'_l}. \tag{6}$$

These inequalities are useful when interpreting the results of a decision study for a $r \times (S:L:t)$ design. They also provide a context for understanding the results of our study.

To study the influence of our experimental conditions on validity, we analyzed the data in a variety of ways. First, we conducted two unplanned auxiliary analyses based on results from the G-study to explore reasons for observing changes in variance components across the conditions of our study. We focused on rater and carryover effects as possible sources for these changes. Next, we evaluated the impact of our conditions on CLASS-S domain scores through ANOVA methods, follow-up procedures, and correlations. Finally, we evaluated predictive validity through a multilevel model of general education student test scores that were standardized within subject and grade level. The first level of the model included a random intercept and nonrandom effects that accounted for gender, minority status, grade level, and prior achievement. The second level model for the intercept included a CLASS-S domain score, which was the main effect of interest. We conducted a separate multilevel analysis separately for each CLASS-S domain, given the high correlation among the three scales. We used the xtmixed command in Stata Version 12 to conduct the multilevel analysis.

Results

G-Study

Table 1 presents sources of variance estimated for each study condition for each domain of the CLASS. Across all conditions, teachers are the largest source of variance in CLORG scores. Teacher variance is also the largest source of variance in EMSUP scores in the 1 × 40 condition. This result is desirable as teacher variance in the G-study becomes universe score variance in the D-study. In all other conditions, the largest source of variance is due to raters, the rater by lesson within teacher

Table 1. Variance Component Estimates (Percentage of Total in Parentheses) for Each Condition.

Condition	Component	EMSUP	INSUP	CLORG
1 × 40	Teacher, *t*	.208 (34.21)	.183 (18.58)	.420 (47.46)
	Rater, *r*	.032 (5.26)	.261 (26.50)	.040 (4.52)
	l : *t*	.118 (19.41)	.235 (23.86)	.174 (19.66)
	r×*t*	.067 (11.02)	.072 (7.31)	.053 (5.99)
	r×(*l* : *t*)	.183 (30.10)	.234 (23.76)	.198 (22.37)
2 × 20 random	Teacher, *t*	.195 (29.24)	.215 (16.85)	.279 (37.40)
	Rater, *r*	.032 (4.80)	.196 (15.36)	.020 (2.68)
	l : *t*	.064 (9.60)	.087 (6.82)	.039 (5.23)
	r×*t*	.024 (3.60)	.076 (5.96)	.027 (3.62)
	s : *l* : *t*	**.106 (15.89)**	**.291 (22.81)**	**.144 (19.30)**
	r×(*l* : *t*)	**.027 (4.05)**	**.000 (0.00)**	**.005 (0.67)**
	r×(*s* : *l* : *t*)	.219 (32.83)	.411 (32.21)	.232 (31.10)
2 × 20 ordered	Teacher, *t*	.141 (25.41)	.065 (9.50)	.198 (28.99)
	Rater, *r*	.015 (2.70)	.006 (0.88)	.066 (9.66)
	l : *t*	.042 (7.57)	.011 (1.61)	.101 (14.79)
	r×*t*	.000 (0.00)	.023 (3.36)	.040 (5.86)
	s : *l* : *t*	**.020 (3.60)**	**.071 (10.38)**	**.032 (4.69)**
	r×(*l* : *t*)	**.233 (41.98)**	**.344 (50.29)**	**.122 (17.86)**
	r×(*s* : *l* : *t*)	.104 (18.74)	.164 (23.98)	.124 (18.16)
4 × 10 ordered	Teacher, *t*	.059 (7.06)	.050 (8.39)	.145 (22.48)
	Rater, *r*	.412 (49.28)	.017 (2.85)	.083 (12.87)
	l : *t*	.042 (5.02)	.082 (13.76)	.063 (9.77)
	r×*t*	.038 (4.55)	.043 (7.21)	.031 (4.81)
	s : *l* : *t*	**.018 (2.15)**	**.107 (17.95)**	**.040 (6.20)**
	r×(*l* : *t*)	**.154 (18.42)**	**.119 (19.97)**	**.139 (21.55)**
	r×(*s* : *l* : *t*)	.113 (13.52)	.178 (29.87)	.144 (22.33)

component, or the residual. Rater variance is not much of a concern in the current study because it does not contribute to relative error variance in the D-study. However, the rater by lesson within teacher component and the residual component are problematic. They both contribute to relative error variance in the D-study as do all other variance components that are not teachers or raters.

For two of these variance components in Table 1, there is a discernible and interesting pattern across conditions. Specifically, in the 2 × 20 and 4 × 10 ordered conditions, the rater by lesson within teacher component accounts for a large portion of variance in EMSUP, INSUP, and CLORG scores, but the segment within lesson within teacher component does not. The opposite result is evident in the 2 × 20 random condition; segment within lesson within teacher accounts for a large portion of variance, but the rater by lesson within teacher only accounts for a small portion. We highlighted this pattern with bold font in Table 1. These results suggest that when segments are viewed and rated in immediate succession, raters' scores on the second segment are rarely different from their scores in the first segment. Consequently, there is little segment variance but substantial variance in the rater by lesson within teacher component. In contrast, this effect diminishes when raters view segments in a random order that may be separated by days or months of time. In this case, raters are more likely to base their score on the actual segment being viewed and change their score accordingly when viewing segments in a random order. As a result, there is very little rater by lesson within teacher variance and a substantial amount of segment within lesson within teacher variance when segments are randomized.

As a follow-up test of the idea that raters rarely change their score when segments are viewed sequentially, we computed the correlation between segments in the 2 × 20 ordered condition and the correlation between segments in the 2 × 20 random condition. EMSUP, INSUP, and CLORG between segment correlations are 0.76, 0.64, and 0.74, respectively, in the 2 × 20 ordered condition, but they are only 0.47, 0.40, and 0.48 in the 2 × 20 random condition. Correlations in the random condition are all significantly lower than those in the ordered condition at a significance level of .001.

We also explored the idea that the large rater by lesson within teacher variance component in the 2 × 20 ordered condition was due to rater drift. To study this effect, we averaged segment scores to obtain a score for each lesson. We then computed the difference in days between the start of scoring for the study and the date a rater scored a lesson. We plotted the results for each domain and added a local linear regression line for each rater. Figure 1 shows a prominent decline in EMSUP and INSUP scores for both raters in the 2 × 20 ordered condition. Conversely, rater scores at the beginning and end of the scoring period in the 2 × 20 random condition are much more similar, although there are fluctuations over the time period. In both conditions, rater scores for CLORG are similar and stable over time. These results suggest that some of the rater by lesson within teacher variance in the 2 × 20 ordered condition could be due to rater trends. It seems that raters become more severe in their ratings as time goes on, perhaps due to fatigue. As a result, viewing lessons in order means that some teachers have both video tapes viewed when raters tend to give low ratings. The random condition counteracts this effect because teachers likely have at least one lesson viewed when raters are lenient and one rated when they are severe.

To summarize the results of the G-study, our results provide evidence that support three findings. First, raters rarely change their scores from one segment to another when segments are viewed in sequential order and this contributes to rater by lesson within teacher variance. Second, raters also appear to trend downward as the rating period progresses when segments are viewed sequentially. Finally, randomizing segments seems to prevent the carryover from one segment to another and it also appears to mitigate rater trends over time.

D-Study

Variance components from the random condition and two ordered conditions (Table 1) mainly differed on the segment within lesson within teacher component and the

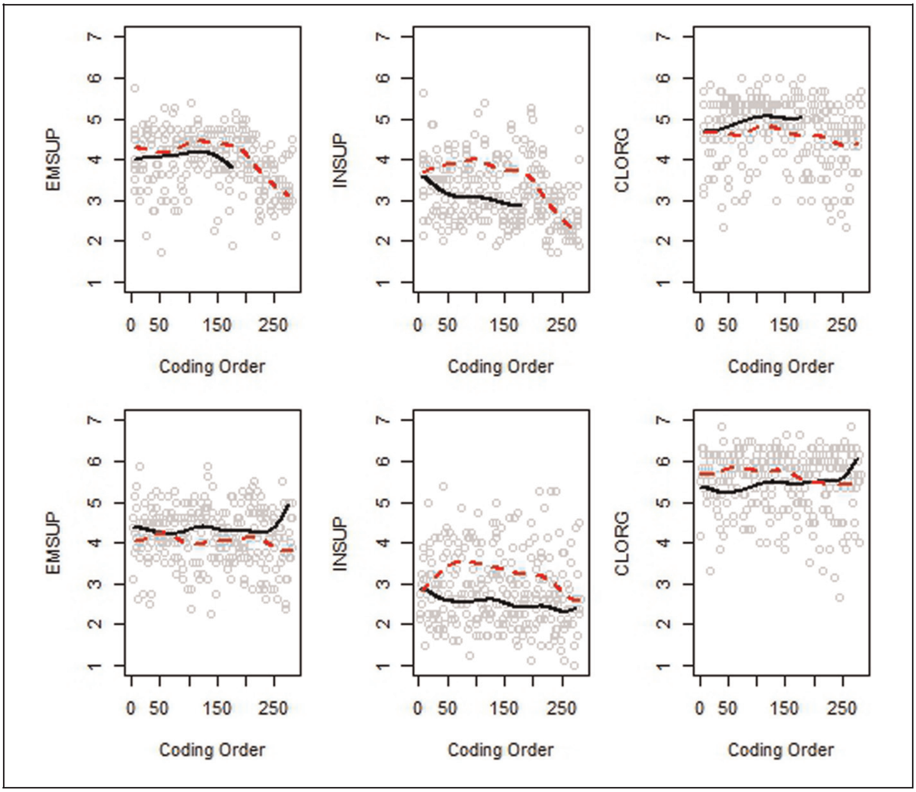

Figure 1. Rater trends over the duration of the scoring period. The *x*-axis of each plot is the rank of the number of days between the time scoring begins and when a video is scored. The top two panels are for the 2 × 20 ordered condition and the bottom two panels are for the 2 × 20 random condition.

rater by lesson within teacher component. As illustrated in the previous subsection, the way these two sources of variance changed in each condition yields insight into two rater processes—carryover effects and rater drift—that are not intended to influence scores. However, both these variance components are part of relative error and the change in magnitude across conditions may not make much of a difference in terms of relative error and the generalizability coefficient.

To investigate the impacts of these variance components on relative error and the generalizability coefficients, D-study results in Table 2 indicate that relative error variance for EMSUP and INSUP scores are lower in the 2 × 20 and 4 × 10 ordered conditions, but relative error variance for CLORG is lowest for the 2 × 20 random condition. However, universe score variance follows this same pattern, which leads the 2 × 20 random condition to have the highest generalizability coefficient estimates. Whereas the lowest generalizability coefficient was for the 4 × 10 ordered

Table 2. Decision Study Results for all Conditions.

Statistic	Condition	EMSUP	INSUP	CLORG
Universe score variance, $\sigma^2(\tau)$	1 × 40	0.208	0.183	0.420
	2 × 20 random	0.195	0.215	0.279
	2 × 20 ordered	0.141	0.065	0.198
	4 × 10 ordered	0.059	0.050	0.145
Relative error variance, $\sigma^2(\delta)$	1 × 40	0.103	0.153	0.118
	2 × 20 random	0.064[a]	0.140	0.063[e]
	2 × 20 ordered	0.057[b]	0.090	0.082
	4 × 10 ordered	0.061	0.082[c,d]	0.065[f]
Generalizability coefficient, $E\rho^2$	1 × 40	0.67	0.54	0.78
	2 × 20 random	0.75	0.61	0.81
	2 × 20 ordered	0.71	0.42	0.71
	4 × 10 ordered	0.49	0.38	0.69

Note. Facet sample sizes are the same as those in the generalizability study.

a. Significantly lower than error variance for the 1 × 40 condition; 95% confidence interval is (.009, .104).

b. Significantly lower than error variance for the 1 × 40 condition; 95% confidence interval is (.002, .109).

c. Significantly lower than error variance for the 1 × 40 condition; 95% confidence interval is (.033, .197).

d. Significantly lower than error variance for the 2 × 20 random condition; 95% confidence interval is (.026, .170).

e. Significantly lower than error variance for the 1 × 40 random condition; 95% confidence interval is (.013, .137).

f. Significantly lower than error variance for the 1 × 40 random condition; 95% confidence interval is (.010, .162).

condition. Bootstrap confidence intervals[2] for the pairwise comparison of relative error variances between conditions indicate that the multiple segment conditions tend to have significantly lower relative error variance than the 1 × 40 condition. This result varies by domain, but it is consistent enough to suggest that considerable reductions in relative error variance are achieved by using multiple segments. However, large error variance in the 1 × 40 condition is also accompanied by large universe score variance leading the generalizability coefficients that were never the smallest in any condition. The results also suggest that length of observation is an important source of variance that could be included as a facet in the universe. Instead of narrowing the universe to inferences about 10- or 20-minute observations, observation length could be treated as random and incorporated into the measurement procedure.

Although reliability estimates are highest in the 2 × 20 random condition, the question arises as to whether increasing the number of lessons, segments, or raters will produce reliability estimates that are more favorable for other conditions. We return to the inequalities discussed earlier to answer this question. We showed that increasing lessons is always better than increasing the number of segments. Using

Table 3. Correlations and Descriptive Statistics.

		Condition			
Scale		(1) 1 × 40	(2) 2 × 20 Random	(3) 2 × 20 Ordered	(4) 4 × 10 Ordered
EMSUP	1	1.00			
	2	0.90	1.00		
	3	0.84	0.84	1.00	
	4	0.77	0.79	0.73	1.00
	Mean	4.11	4.18	3.97	3.93
	SD	0.56	0.51	0.46	0.36
INSUP	1	1.00			
	2	0.87	1.00		
	3	0.72	0.65	1.00	
	4	0.79	0.78	0.66	1.00
	Mean	3.21	2.83	3.15	2.55
	SD	0.58	0.60	0.42	0.37
CLORG	1	1.00			
	2	0.90	1.00		
	3	0.91	0.88	1.00	
	4	0.85	0.80	0.84	1.00
	Mean	5.18	5.51	4.72	4.74
	SD	0.73	0.59	0.55	0.46

the inequality in Equation 5 and the variance components in Table 1, it is also always better to increase the number of lessons instead of the number of raters.

The inequality in Equation 6 is particularly important with respect to our experimental conditions because this inequality involves the rater by lesson within teacher component and the segment within lesson within teacher component. The relative magnitude of these two components differed in the 2 × 20 random and 2 × 20 and 4 × 10 ordered conditions. Thus, the decision to increase the number of segments or raters depends on the estimated variance components. Using the inequality in Equation 6 and the variance components in Table 1, increasing the number of segments in the random condition leads to lower relative error variance than increasing the number of raters. The opposite is true in the 2 × 20 ordered and 4 × 10 ordered conditions; increasing the number of raters leads to lower relative error variance. Thus, lower relative error variance in the 2 × 20 random condition is achieved by increasing the number of 20-minute segments, but in the 2 × 20 and 4 × 10 ordered conditions, lower relative error variance is achieved by increasing the number of raters. The implication is that the method for presenting segments to raters affects the choice of which facet sample size to increase.

To briefly recap the D-study findings, our analysis supports four important results. First, significantly lower relative error variances are frequently achieved in all domains by rating multiple segments instead of rating a single 40-minute lesson.

Table 4. Bonferroni Adjusted Confidence Intervals for Pairwise Comparisons.

Comparison	Adjusted 95% confidence interval		
	EMSUP	INSUP	CLORG
2 × 20 random–1 × 40	(−0.06, 0.19)	(−0.53, −0.23)*	(0.20, 0.47)*
2 × 20 ordered–1 × 40	(−0.26, −0.02)*	(−0.21, 0.09)	(−0.60, −0.33)*
4 × 10 ordered–1 × 40	(−0.30, −0.06)*	(−0.81, −0.51)*	(−0.58, −0.31)*
2 × 20 ordered–2 × 20 random	(−0.32, −0.08)*	(0.17, 0.47)*	(−0.93, −0.66)*
4 × 10 ordered–2 × 20 random	(−0.37, −0.12)*	(−0.44, −0.13)*	(−0.91, −0.64)*
4 × 10 ordered–2 × 20 ordered	(−0.16, 0.08)	(−0.76, −0.46)*	(−0.11, 0.16)

Note. Type I error adjustment was 0.05/6/2 = .0042.
*Statistically different from zero.

Table 5. Descriptive Statistics for Student Math and Reading Scores by Grade.

Subject	Test year	Grade	*n*	Mean	*SD*
Math	Prior	6	103	513.29	77.46
		7	56	368.29	62.61
		8	64	492.11	63.27
		9	47	496.85	70.06
		10	43	489.88	68.11
		11	5	490.6	37.6
	Current	6	103	448.32	93.10
		7	56	401.55	71.12
		8	64	479.14	54.88
		9	47	485.66	59.40
		10	43	487.35	63.40
		11	5	487.80	16.78
Reading	Prior	6	31	486.10	58.70
		7	59	469.12	55.43
		8	59	493.76	51.60
		11	82	469.73	50.45
		12	3	409.33	84.83
	Current	6	31	487.48	54.12
		7	59	475.24	52.45
		8	59	495.07	50.15
		11	82	512.61	56.75
		12	3	434.67	41.77

Second, rating sequential 10-minute segments produced the lowest generalizability coefficient. This suggests that a 10-minute observation may not be sufficient for an observer to notice and rate true characteristics of teacher–student interactions. There simply may be inadequate time to see a complete and scorable interaction during a 10-minute period. Third, randomizing the order of segment presentation leads to

Table 6. Multilevel Model Estimates for CLASS Scales Predicting student Math Test Scores.

Scale	Condition	*B*	*SE*	*p* Value	Level 1 variance	Level 2 variance
EMSUP	1 × 40	.20	.12	.097	.496	.023
	2 × 20 random	.39	.11	<.001	.498	.003
	2 × 20 ordered	.29	.13	.032	.495	.019
	4 × 10 ordered	.54	.18	.003	.497	.006
INSUP	1 × 40	.12	.18	.506	.496	.030
	2 × 20 random	.19	.19	.327	.496	.029
	2 × 20 ordered	.18	.18	.308	.496	.029
	4 × 10 ordered	.56	.31	.072	.498	.018
CLORG	1 × 40	.17	.09	.068	.496	.021
	2 × 20 random	.25	.11	.026	.496	.017
	2 × 20 ordered	.26	.12	.032	.496	.018
	4 × 10 ordered	.33	.21	.119	.497	.022

Note. Fixed effects for grade, prior achievement, minority status, study year, and gender are not listed.

higher, albeit not significantly different, reliability estimates. Finally, increasing the number of 20-minute segments is a better choice than increasing the number of raters when segments are presented in random order but the opposite is true when segments are presented sequentially.

Relationships Among CLASS-S Domain Scores

Table 3 shows that domain score means are rather variable across conditions. Indeed, significant and moderately sized differences exist for EMSUP ($F_{3,131} = 12.20, p < .001; Cohen's f = 0.2$), INSUP ($F_{3,132} = 60.15, p < .001; Cohen's f = 0.53$), and CLORG ($F_{3,131} = 118.25, p < .001; Cohen's f = 0.56$).Table 4 shows the pairwise comparisons among domain score means and most of them are statistically significant. EMSUP and CLORG scores are highest in the 2 × 20 random condition, but INSUP scores are highest in the 2 × 20 ordered condition. The lowest scores for each domain are in the 4 × 10 ordered condition. The differences in means by condition are particular important for absolute decisions that are made from CLASS-S domain scores as each condition would lead to different decisions about teachers. However, in the context of relative decisions (the type of decision that is the focus of this study), the correlation among scores across conditions is more relevant.

For EMSUP and INSUP, the highest correlations are between the 1 × 40 and 2 × 20 random conditions (see Table 3). The correlation is also high (0.90) between these conditions for CLORG but not the highest. Correlations among the other conditions are also high for each domain. Indeed, no correlation is less than 0.65. In each domain, the lowest correlations typically involve the 4 × 10 condition. These results mean that rank ordering of teachers on the basis of observed scores is fairly similar

Table 7. Multilevel Model Estimates for CLASS Scales Predicting Student Reading Test Scores.

Scale	Condition	*B*	*SE*	*p* Value	Level 1 variance	Level 2 variance
EMSUP	1 × 40	.58	.13	<.001	.524	<.001
	2 × 20 random	.53	.15	.001	.530	.005
	2 × 20 ordered	.55	.23	.018	.530	.015
	4 × 10 ordered	1.18	.52	.023	.529	.019
INSUP	1 × 40	.69	.15	<.001	.519	<.001
	2 × 20 random	.47	.17	.005	.530	.010
	2 × 20 ordered	.79	.26	.002	.531	.008
	4 × 10 ordered	.14	.79	.859	.529	.040
CLORG	1 × 40	.51	.12	<.001	.526	<.001
	2 × 20 random	.30	.36	.403	.530	.035
	2 × 20 ordered	1.31	.29	<.001	.521	<.001
	4 × 10 ordered	.81	.33	.014	.530	.015

Note. Fixed effects for grade, prior achievement, minority status, study year, and gender are not listed.

for all domain scores between the 1 × 40 and 2 × 20 conditions, but slightly different for all domain scores in the 4 × 10 condition.

Predictive Validity Analysis

The predictive validity study focused on math and reading teachers that had student level data. Math teachers taught in Grades 6 through 11. Almost all math teachers taught a single class, but one teacher taught three. Among the 14 teachers with student-level math test data, class sizes ranged from 2 to 36, with 20 students being the typical class size. Table 5 lists descriptive statistics for math test scores by grade. Reading teachers with student-level data taught Grade 6, 7, 8, 11, and 12. Class sizes ranged from 1 to 36 students. The teacher with only one student also taught a second class with 17 students. Another teacher with only 3 twelfth-grade students also taught a class of 36 eleventh-grade students. Like math, the typical class size was about 20 students. Table 5 also lists descriptive statistics for reading scores.

For the combined sample, a majority of students are female (55.6%) and most of them represented White (73.24%), Black (19.94%), Hispanic (3.89%), and Asian (1.76%) backgrounds. Native Americans, Hawaiians, and biracial students each represent less than 1% of the sample. The teacher sample consists of males (44.4%) and females (55.6%) who are White (91.1%), biracial (6.7%), or Black. They taught an average of 8.25 years. Sixty percent have a bachelor's degree or 1 year beyond the bachelor's level, 31% have a master's degree, and the remaining teachers have an Educational Specialist of Doctor of Philosophy degree.

Table 6 lists results from the multilevel analysis of math test scores. The analysis takes into account demographic variables and prior achievement, but Table 6 only lists the effects of interest. Across all conditions, the 1 × 40 condition has the smallest coefficients, whereas the 4 × 10 always has the largest coefficients. For all but the 4 × 10 condition, the coefficients appear to be fairly similar. In terms of class domains, EMSUP and CLORG are significant predictors of math achievement in almost all conditions. EMSUP is not a significant predictor of achievement in the 1 × 40 condition, and CLORG is not a significant predictor of math achievement in the 1 × 40 or 4 × 10 condition. INSUP is never a significant predictor of math achievement.

Results for reading achievement are more variable across conditions (Table 7). The same patterns seen in the math results are not present in the reading coefficients. Moreover, the coefficients are more variable across conditions for reading scores than they were for math scores. CLASS-S domain scores are significant predictors of reading achievement in all but two conditions. INSUP is not a significant predictor of reading achievement in the 4 × 10 condition and CLORG is not a significant predictor in the 2 × 20 random condition.

Overall, presentation order and segment length do not appear to diminish or enhance predictive validity results. Coefficients do appear to be larger for math scores when lessons are broken into multiple segments, but multiple segments do not seem to make much of a difference in the coefficients for reading scores. Note that the observed lessons are not tied to a specific subject, but the outcomes of interest in the predictive validity study are test scores in a particular subject. It is possible that CLASS-S scores would be more predictive of student achievement when observations are limited to the subject of interest.

Discussion

Teaching observations are increasingly being used in education policy, research, and professional development. As a result, there is a need to understand and improve the psychometric properties of these measures, particularly if scores are to be used for high-stakes purposes, such as evaluating teachers' performance. This study examined how various procedures for using the CLASS-S (Pianta et al., 2008)—a commonly used observational measure of the quality of teachers' interactions with children in 6th grade to 12th grade classrooms—affected the reliability and validity of scores. Specifically, from three 40-minute videotaped lessons collected from 47 teachers, we manipulated the length of observation and order of presentation of the lessons in four different ways, and we randomized two raters to observe and rate all lessons from all teachers.

A generalizability study, decision study, and additional analyses of the validity of scores were conducted to contrast the reliability and validity for each study condition. The generalizability study estimated multiple sources of variance in scores related to rater, teacher, lesson, and for the three study conditions that decomposed lessons into multiple occasions, segment. Although there are no appreciable differences in the financial costs of implementing the four different operational procedures under study, there were notable differences in some aspects of the reliability and validity of scores related to segment length and/or order of presentation.

Specifically, results indicated that lessons rated in the shortest and most frequent manner (4 × 10 minute segments) produced the lowest generalizability coefficients for all three domains of the CLASS-S. Although this condition produces the highest number of occasions of measurement per lesson, which is favorable in reducing error attributable to segments that is part of the relative error estimate, results indicated that this condition was characterized by the lowest universe score variances and lowest generalizability coefficients. This suggests that 10 minutes may not be adequate time to observe the specific indicators of teaching quality that inform raters' judgments about the quality of teaching. As a result, raters likely use extraneous information when judging the quality of a lesson, which subsequently reduces universe score variance and the generalizability coefficient.

Procedures involving single ratings made following 40-minute lessons resulted in domain scores that were highly correlated with scores from the other three observation conditions, and in generalizability coefficients that are similar to those from observations of 20-minute segments. However, scores from the 1 × 40 condition tended to suffer from large relative error variance and low predictive validity coefficients. Among the two conditions in which raters assigned scores following 20-minute segments, there was no evidence that the order in which the 20-minute segments were presented to raters significantly affected the reliability of scores. However, randomly presenting 20 minute segments to raters from the entire pool of 282 segments per teacher had the advantage of reducing sources of construct irrelevant variance by reducing carry over effects and rater drift.

It is important to consider some limitations with these results. Table 1 shows that a large source of variance is often attributed to the residual effect. This result suggests that conditions of measurement outside of those studied in this article are influencing scores in a notable way. We have heard raters report teaching methods and classroom management features that affect their ability to provide ratings. For example, they may view a classroom while students are completing a worksheet leaving little opportunity to observe and rate teaching. These characteristics are not easily classified by teaching observation measures and likely contribute to a large residual variance term. Unfortunately, our study was unable to address this possibility, but it is an area worth further study. Until more is known about conditions that contribute to large residual variance terms, practitioners should ask raters to provide annotations and notes about challenges encountered when using the rating scales. This could inform a revision to the measures or a standardization that reduces this source of variance.

In sum, results indicate that operational procedures related to length of observation and order of presentation can impact the reliability and validity of scores, while adding few financial costs for conducting teaching observations. Given the growing importance of teaching observations, further research is needed to understand the tradeoffs between reliability and validity related to the operational procedures under conditions of different instruments, frequencies of observations, ordering of presentation, and modality of collecting data (live and videotaped).

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.

Notes

1. Generalizability theory allows for two types of error variance. Relative error reflects all sources of variance that affect the rank ordering of examinees, whereas absolute error represents all sources of error that affect the relative standing and overall location of examinees. Absolute error variance is typically used to represent the error in criterion-referenced testing, where an examinee's score is compared to a passing score. Relative error best represent error in norm-referenced test where the purpose is to distinguish between examinees at various score levels. The reliability coefficient for relative error is referred to as the generalizability coefficient, whereas the reliability coefficient for absolute error is referred to as the index of dependability or phi-coefficient (Brennan, 1992). As described in the text, we applied generalizability theory to classroom observation scores. We focused on relative error and the generalizability coefficient.
2. We also computed bootstrap confidence intervals for the generalizability coefficients. However, these intervals were excessively wide as they involved a ratio of variances. These intervals did not reveal any significant differences.

References

Allen, J., Pianta, R., Gregory, A., Mikami, A., & 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. doi:10.1080/10627197.2012.715014

Bill and Melinda Gates Foundation. (2012). *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_Practioner_Brief.pdf

Brennan, R. L. (1992). Generalizability theory: An NCME instructional module. *Educational Measurement: Issues and Practice*, *11*, 27-34.

Casabianca, J., & Lockwood, J. R. (2013, March). *Rater drift and time trends in teaching observations. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness*, Washington, DC.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Danielson, C. (2011). *The framework for teaching evaluation instrument*. Princeton, NJ: Danielson Group.

Ebbinghaus, H. (1913). *On memory: A contribution to experimental psychology*. New York, NY: Teacher's College Press.

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

Fritz, M., & Chen, K. (2013, February). *Federal grants, rise of charter schools expand teacher evaluations*. Retrieved from http://www.pbs.org/newshour/rundown/2013/02/teacher-evaluations.html

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores* (National Bureau of Economic Research Working Paper No. 16015). Cambridge, MA: National Bureau of Economic Research.

Hamre, B., Pianta, R., Downer, J., Hamagami, F., Mashburn, A., Jones, S., . . . Brackett, M. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *Elementary School Journal*, *113*(4), 461-487.

Heitin, L. (2011, October 26). Study: States' teacher-evaluation policies are A-changin'. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/teacherbeat/2011/10/study_states_teacher_evaluatio.html

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

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

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review*, *34*, 507-519.

Ho, A. D., & Kane, T. J. (2013). *The reliability of teaching observations by school personnel*. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6*, 125-160.

Kane, M. T. (2011). The errors of our ways. *Journal of Educational Measurement*, *48*, 12-30.

Leventhal, L., Turcotte, S. J., Abrami, P. C., & Perry, R. P. (1983). Primacy/recency effects in student ratings of instruction: A reinterpretation of gain-loss effects. *Journal of Educational Psychology*, *75*, 692-704.

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

Marcoulides, G. (1989a). The application of generalizability analysis to observational studies. *Quality & Quantity*, *23*, 115-127. doi:10.1007/BF00151898

Marcoulides, G. A. (1989b). Performance appraisal: Issues of validity. *Performance Improvement Quarterly*, *2*, 3-12.

Mashburn, A. J., Downer, J. D., Hamre, B. K., Justice, L. M., & Pianta, R. C. (2010). Teaching consultation and children's language and literacy development during pre-kindergarten. *Applied Developmental Science*, *14*, 179-198.

Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2013). Improving the power of an efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science*. Advance online publication. doi:10.1007/s11121-012-0357-3

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D, . . . Howes, C. (2008). Measures of pre-k quality and children's development of academic, language and social skills. *Child Development*, *79*, 732-749. doi:10.1111/j.1467-8624.2008.01154.x

McCaffrey, D., Yuan, K., Savitsky, T., Lockwood, J. R., & Edelen, M.O. (2013, February). *Using latent hierarchical estimation to uncover multivariate structure in teaching observation protocol data. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness*, Washington, DC.

Meyer, J. P., Henry, A. E., & Mashburn, A. J. (2011). Occasions and the reliability of teaching observations: Alternative conceptualizations and methods of analysis. *Educational Assessment Journal*, *16*, 227-243. doi:10.1080/10627197.2011.638884

Mikami, A., Gregory, A., Allen, J. P., Pianta, R. C., & Lun, J. (2011). Effects of a teacher professional development intervention on peer relationships in secondary classrooms. *School Psychology Review*, *40*, 367-385.

Pianta, R. C., Hamre, B., Hayes, N., Mintz, S., & LaParo, K. M. (2008). *Classroom Assessment Scoring System–Secondary (CLASS-S)*. Charlottesville, VA: University of Virginia.

Savitsky, T. T., & McCaffrey, D. F. (2013, February). *Bayesian hierarchical multivariate formulation with factor analysis for nested ordinal data. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness*, Washington, DC.

U.S. Department of Education. (2009). *Race to the top executive summary*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf