When Seeing is Believing: Generalizability and Decision Studies for Observational Data in Evaluation and Research on Teaching

Timothy J. Weston[1], Charles N. Hayward[2], and Sandra L. Laursen[2]

[1]National Center for Women & IT (NCWIT), College of Engineering and Applied Science, University of Colorado, Boulder

[2]Ethnography & Evaluation Research, School of Arts and Sciences, University of Colorado, Boulder

Author Note

Timothy J. Weston https://orcid.org/0000-0003-2982-4874

Sandra L. Laursen https://orcid.org/0000-0002-4327-9887

Charles N. Hayward https://orcid.org/0000-0003-2733-6025

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Timothy J. Weston, National Center for Women & IT (NCWIT), 320 UCB, University of Colorado, Boulder, Boulder CO 80309-0320. Email: westont@colorado.edu

Abstract

Observations are widely used in research and evaluation to characterize teaching and learning activities. Because conducting observations is typically resource intensive, it is important that inferences from observation data are made confidently. While attention focuses on interrater reliability, the reliability of a single-class measure over the course of a semester receives less attention. We examined the use and limitations of observation for evaluating teaching practices, and how many observations are needed during a typical course to make confident inferences about teaching practices. We conducted two studies based in generalizability theory to calculate reliabilities given class-to-class variation in teaching over a semester. Eleven observations of class periods over the length of a semester were needed to achieve a reliable measure, many more than the one to four class periods typically observed in the literature. Findings suggest practitioners may need to devote more resources than anticipated to achieve reliable measures and comparisons.

Keywords: observation, generalizability studies, reliability, teaching practice, social science research methods, higher education

When Seeing is Believing: Generalizability and Decision Studies for Observational Data in Evaluation and Research on Teaching

Direct observation is a widespread practice in the evaluation community, especially in the evaluation of teachers (Wragg, 2013). Evaluators use observations across numerous settings, including medical education where direct observation is used to assess diagnostic, basic care, and surgical skills (Iobst et al., 2010; Naeem, 2012). Others use observations in a diverse range of non-educational areas such as evaluating the effectiveness of assistive technology for dementia patients (Nolan et al., 2002), the quality of parent-child interactions (Gardner, 2000), and consumer behavior (Carins, 2016). Teachers use direct observations to assess student skills, performance and understanding as an alternative form of testing (Mertler, 2016). In K-12 education, direct observation is used in teacher evaluation as principals and peer teachers visit classrooms to assess teachers and college faculty (Darling-Hammond et al., 2012). In many cases, assessments drawn from observations have consequences for employment (Cohen & Goldhaber, 2016). While observations are used as a component of a wider system that employs student outcomes and other measures, their use in practice is often problematic due to the lack of reliability in these measures (Marcoulides, 1989).

In higher education, structured classroom observations are also an increasingly important tool for evaluation of teaching in higher education, the focus of our study (Smith et al., 2013). This is common in the evaluation of science, technology, engineering, and mathematics (STEM) teaching (American Association for the Advancement of Science [AAAS], 2013), and professional development for college faculty (Ebert-May et al., 2011). Observational protocols are used to describe teaching and learning activities in classrooms, especially in situations where instructors implement new teaching methods (Smith et al., 2013; Laursen et al., 2014).

Observations are also used extensively in the evaluation of professional development programs and interventions in teaching in higher education (Stains et al., 2015; Pilburn et al., 2000).

Unlike surveys, observations have the advantage of not depending upon self-reports of behavior; instead observers directly witness what teachers and students do in the classroom (Stains et al., 2018; Ebert-May et al., 2011). While observations provide a way around dependence on self-report, they also have problems of their own. Most are time and resource intensive and are prone to their own kind of observer subjectivity, especially when it comes to rater agreement (Waxman & Padron, 2004). Additionally, the representativeness of observations can be in doubt when claims are made about teaching style from low numbers of observations, as is often the case in research and teacher evaluation (Cohen & Goldhaber, 2016; Hill et al., 2012).

As these examples show, evaluators, researchers, and professional development experts all use observation to describe and assess teaching, and thus need accurate and reliable descriptions of what instructors do in college classrooms (Cohen & Goldhaber, 2016; West, et al., 2013). Observations are used in evaluation and research studies to assess interventions such as teacher-scientist collaboration (Campbell et al., 2012), co-teaching (Beach et al., 2007), and faculty development workshops (Adamson et al., 2003; Ebert-May et al., 2011; Stains et al., 2015). Observations also play a part in understanding the relationship between faculty demographics and teaching style (Budd et al., 2013) and the efficacy of pedagogies on student engagement (Lane & Harris, 2015). Others use observation scores as independent variables for comparing teaching methods with their effects on student outcomes (Bowling et al., 2008; Budd, et al., 2010). In all of these efforts, direct observation studies make descriptions or comparisons of teaching practice, and often make evaluative judgments about the efficacy of interventions based on observations.

The impetus for developing observational protocols for STEM (and other classrooms comes in part from educational reform efforts meant to improve teaching in classrooms using research-based instructional methods (Blanchard et al., 2010; Michael, 2006). Recent reports (Freeman et al., 2014; Brewer & Smith, 2011) present multiple studies supporting the efficacy of reform-based teaching methods and advocate their use in undergraduate STEM classes. Most of these methods emphasize less lecture and more student participation using broadly defined approaches such as active learning (Chi & Wiley, 2014), inquiry-based learning (Laursen et al., 2014), active-inquiry teaching (Hake, 1998), and interactive engagement (Turpen & Finkelstein, 2009). More research is needed about these emerging practices, and observations are one tool for conducting rigorous and meaningful studies.

Moreover, evaluators and researchers are increasingly using observations to make consequential claims: Did teachers change their practices in response to professional development? Do instructors accurately report their teaching practices on surveys when compared to observations of actual teaching? Are individual teachers incorporating new instructional methods in their practices? Making claims about a teacher's practice, especially over a semester or term, depends upon using measures from observations in an appropriate and valid manner. However, in many cases claims based on observational data are made without sufficient evidence (Hill et al., 2012).

This study examines the use and limitations of observational protocols for evaluating teaching practices in undergraduate courses. We pay particular attention to how many observations are needed during a typical course to make confident inferences about teaching practices. We then define statistical criteria for characterizing measurement error for entire courses drawn from a set of observed classes and test these criteria against an empirical data set

from college mathematics courses. We also discuss the implications of our findings for the design of studies that seek to use observations to draw conclusions about teaching and learning.

Review of Observation Protocols and Practices

In this section we review the characteristics of observation protocols and their use in practice in education research and evaluation, emphasizing considerations that affect the confident use of observations to make inferences about teachers' practice.

In the literature, several terms are sometimes used interchangeably or have ambiguous meaning; here the differences are important. We use *class* to refer to an individual meeting or class session, and *course* to refer to the set of all class sessions in a course. A course may consist of 15-60 classes, depending on the length of the term and the number and duration of weekly class meetings. Some studies classify instructors by their *teaching profile* or *teaching style,* which is a broad concept referring to patterns of practice across multiple courses.

Measurement Characteristics of Observational Protocols

Observations have long been used in anthropology and sociology to observe cultural interactions, religious rituals, and customs, and events; with observations on a continuum from full participant observation to the "fly-on-the-wall" observer who has no interaction with participants (Tashakkori & Teddy, 2010). Observers in the ethnographic tradition capture descriptions in field notes and devote extensive time to observations and their analysis; typically these researchers are not using their data to make quantitative comparisons. In educational research, observations commonly straddle qualitative and quantitative methods (Wragg, 2013).

However, the products of structured teacher observations are decidedly quantitative (Pianta & Hamre, 2009). These data are subject to the same measurement standards around

sampling, inference, and reliability that are inherent in any use of numerical data used to describe or compare (Cohen & Goldhaber, 2016). An observation protocol specifies the procedure for gathering data: what is to be recorded, when, and how, in order to standardize data gathering as much as feasible. Existing classroom observation protocols are each designed to fit different needs for research, professional development, or evaluation, but can be classified according to common characteristics; current protocols exemplify these characteristics in many different combinations.

One fundamental design variation in observational instruments is between holistic and segmented observations (Lund et al., 2015). *Holistic* protocols ask observers to rate teachers or teaching at the end of a class session by answering a series of survey-like questions. In one of the earliest holistic protocols, the Teaching Behavior Inventory (TBI) (Murray, 1991), the observer is asked to make ratings at the end of class noting the presence or absence of specific teaching qualities such as "clarity" or "enthusiasm." The Reformed Teaching Observation Protocol (RTOP) (Pilburn et al., 2000; Sawada et al., 2002) also uses a holistic protocol where observers make judgments about the quality of the teaching using 25 Likert-like scale items that rate lesson design and classroom culture. The RTOP is now widely used and provides a scale that classifies teachers as more or less student-centered.

Segmented observational protocols are more granular and divide a class into timed segments with the observer observing teachers and students and recording behaviors within short intervals (e.g., two minutes). The Teaching Dimensions Observation Protocol (TDOP) (Hora & Ferrare, 2013) is a popular segmented instrument and the model for similar protocols. Other segmented protocols include the Flanders Interaction Analysis (FIA), the VaNTH Observation System (VOS), and the Classroom Observation Rubric; each designed for different classroom

contexts and focusing on different activities and behaviors (AAAS, 2013). Segmented protocols provide copious numerical data that lend themselves to sophisticated quantitative analytic techniques such as cluster and latent class analyses (Lund et al., 2015; Halpin & Kieffer, 2015).

Protocols differ also on how their items are designed (Lund et al., 2015). Some observational items ask for evaluative judgments of the quality of teaching, while others are more descriptive using simple observations of behaviors. Quality ratings are more subjective and call on observers' expert knowledge about pedagogy to make judgments of a teacher's efficacy, while descriptive items simply mark the presence or absence of practices. Yet even descriptive protocols seek to reduce subjectivity in coding by providing codebooks that differentiate similar codes and standardize the use of individual codes across different raters (Hora et al., 2013). Protocols can measure the same things in different ways based on whether they are segmented or holistic, evaluative or descriptive. For example, both the TDOP and the RTOP have been used to assess active learning (Sawada et al., 2002; Hora & Ferrare, 2014). The TDOP uses a segmented and descriptive approach, asking observers to mark the presence or absence of specific activities in 2-minute intervals throughout the class, including things such as "small group work" or "desk work." The RTOP uses a holistic and evaluative approach. Observers take structured notes throughout the class, then, at the end of the class, rate the class on items such as "There was a high proportion of student talk and a significant amount of it occurred between and among students" or "Active participation of students was encouraged and valued." Observers are told to use their judgment to assign a value from 0, (never occurred in class) to 4, (very descriptive of the lesson).

Items also differ in the amount of inference observers need to make about non-observable teacher or student phenomena such as cognition, motivation, or engagement. Some items from

the TDOP, for example, ask the observer to infer students' level of interest or engagement from their general affect or other behaviors (Hora et al., 2013). Research has shown that independent observers experience more difficulty agreeing on ratings for both quality and inferential items than for descriptions of behavior, although observers also vary substantially in their agreement across different behaviors (Cohen & Goldhaber, 2016; West et al., 2013; Amrein-Beardsley & Popp, 2012).

Obstacles to Confident Use of Observations in Research

Practitioners who use observational instruments for evaluation, research, or professional development confront several obstacles in their operational use. One oft-discussed obstacle is establishing interrater reliability, not only as a characteristic of the instrument itself, but for the instrument as used for a specific purpose. In fact, many discussions of measurement rigor in observations focus exclusively on reducing this error (Cash et al., 2012). Instrument developers publish the reliability of their measures to provide information about interrater agreement (Hora et al., 2013; Sawada et al., 2002). More sophisticated analyses of reliability use generalizability theory (Webb et al., 2006), which take into account different sources of variability that affect agreement and assesses how agreement varies across different item types, settings, teachers, and occasions.

While the confident use of a protocol does depend upon the characteristics of that instrument, the actual interrater reliability of raters *in practice* depends upon the amount and type of training given to raters, and on ongoing calibration of raters over time (Cash et al., 2012). In many settings, raters are trained to meet a standard of agreement, and to meet an expert standard of accurate observation (Gittomer et al., 2014) by practicing on video-recorded classes or in classrooms until adequate agreement is attained. For instance, when Hora et al. (2013)

established rater agreement for the TDOP, "the researchers participated in an extensive three-day training process" (p. 6), using the protocol with video-recorded classes and discussing the meaning of specific codes before rating classes independently. The developers of the Real-time Instructor Observing Tool (RIOT) (West et al., 2009) established agreement with two researchers who met three times, then observed in the same classroom each week. The Classroom Observation Protocol for Undergraduate STEM (COPUS) (Smith et al., 2011) was designed purposefully to reduce the training raters needed: participants met for 1.5 hours of video training and discussion, then visited pilot classrooms to practice. For all these observational studies, establishing interrater reliability for a specific context was an important first step before actual measurement could be conducted.

The practical logistics of observational studies pose additional obstacles, so implementing observations is logistically daunting when compared to surveys or test data (Hill, et al., 2012). Conducting a single classroom observation requires significant planning, time, resources, and coordination. Instructors and their institutions may limit access to classrooms or require cumbersome consent procedures. Then, observers must physically get to the site, observe for an hour or more, and then collate data from multiple observers and sites (Cash et al., 2012). Video-recording can allow for more observations and less time traveling, but raters must still watch and code the videos (Lee et al., 2017). A fixed video camera may also miss student activities and more subtle interactions taking place in a classroom. Overall, observation is one of the least efficient of all social science data collection methods. For these reasons, it is important to know whether and how observation data can be used confidently to make comparisons in research, professional development, and evaluation.

Reliability of Observations Over Time

Interrater reliability and logistical considerations are commonly recognized as significant challenges for the confident use of observational protocols. Less attention is given to the reliability of observations over time, sometimes called "occasions" in reliability studies. If instructors are observed only once or twice, they may be tempted to put their best foot forward for the observer, leading to "dog and pony" or Hawthorne effects (Weisberg et al., 2009). Instructors may not act naturally when they are being watched (Hill et al., 2012; Cook et al., 1979), or the observer may attend on a day with an unusual class activity, such as a guest presenter, demonstration, or an assessment. Some activities critical to evaluating teaching style and quality, such as group work or interactive discussion, may vary in frequency from class to class (Grossman, et al., 2015). For these reasons, observing only a handful of class sessions may not give a true representation of the course or the instructor's teaching style. Characterizing the average frequency of specific activities and variation from class to class may necessitate multiple observations.

When observations are not representative, and no inference is attempted about an individual's teaching over the course of a term, the data are suitable for certain purposes only. Low numbers of observations can be used if the observations are not linked to consequential decisions, or if they are used to classify or cluster teaching styles seen in a larger population of teachers (Hora & Ferrare, 2014). Stains et al. (2018) provides a good example of how different types of teaching can be categorized and profiled with relatively few observations per teacher. This study used between one and four observations in over 2000 courses and showed the continuing predominance of didactic teaching across STEM disciplines.

Less frequent observations can be also valuable for formative evaluation and professional development when the data are fed back to instructors to improve teaching (Amrein-Beardsley & Popp, 2012). Kane et al. (2012) argued that low numbers of observations can be used effectively for low-stakes feedback that is meant to improve practice, and that different evidentiary standards are appropriate given the differing consequences of decisions.

However, using small numbers of observations is less defensible when observations are used to make evaluation or research claims, inferences and comparisons, or to make decisions based on teachers' classroom performance about retention, compensation, or promotion (Hill et al., 2012; Van der Lans et al., 2016). Yet very low numbers of observations seem to be the norm for many studies. Ebert-May et al. (2011) compared college teachers' self-report of their teaching methods to their self-reported activities from surveys using two observations per semester, as did Lewin et al. (2016) in a similar study. West et al. (2013) used two observations to make conclusions about how graduate students implemented reform-based curriculum over multiple sections of the same class. Smith et al. (2014) observed two classes over a term to characterize teacher and student behavior in large and small classrooms, as did Sawada et al. (2002), who evaluated a program for encouraging the adoption of new teaching methods. Auerbach & Schussler (2016) observed classes once per month during a semester (4-5 times) to compare instructors using alternative teaching methods, and Nadelson et al. (2013) used one observation per semester to make a pre/post comparison about the frequency of teaching practices. Stains et al. (2015) sampled one week of a college class for their study of the impacts of professional development on teaching practices.

Low numbers of observations are the norm even for high-stakes decisions about teacher retention and promotion, such as those required by recent accountability efforts in K-12

education. School principals may be required to make just a handful of observations, or only one (Hill et al., 2012). Cohen & Goldhaber (2016) report that the average number of class observations for non-tenured K-12 teachers was 3.4 over a school year, with large variations among schools in different states. While other researchers (Pianta & Harmer, 2009) have called for more observations, with some exceptions (Darling-Hammond et al., 2012), few researchers, states, or colleges have consistently implemented this practice.

Generalizability Studies

This review of literature shows that the question of how many observations are needed to confidently characterize teaching is both open and important. This issue is usually addressed through generalizability studies (G-studies) (Webb et al., 2006; Brennan, 1992; Marcoulides, 1989), which quantify the reliability of a measure (in this case generalizability) over multiple facets such as raters, teachers, or items. As noted above, many observational studies examine measurement error for interrater agreement, and this is often treated as a feature of the protocol that should be checked when using the protocol in a new study. Some studies also consider the reliability of observations over multiple occasions; essentially asking how many observations are needed for a reliable measure (Hill et al., 2012). However, few if any science education researchers have applied these methods to evaluation and research on teaching and professional development in higher education.

As stated, observations are an almost universal method of teacher evaluations (Darling-Hammond, 2015). Researchers using G-studies in K-12 contexts have found that teachers vary what they do in classrooms across observations, with the duration and quality of specific teaching activities changing from class-to-class. In all these studies, reliability is based on generalizability coefficients. Hill et al. (2012) found that even with four raters rating four

lessons, the observations did not meet their high evidentiary standard of 90% reliability for summative evaluation decisions. Variation in class lessons accounted for up to 39% of total variability for some measures. Adequate reliabilities for four lessons were only achieved when multiple raters were used, in most cases not a practical condition for actual classroom evaluation. In a similar study of classroom teachers, Newton (2010) found that 17% of total variability was due to class-to-class variability; adequate reliability could only be reached with 4 raters and 6 visits to classrooms. Van der Lans et al. (2016), studied primary and secondary teachers in the Netherlands, and reached a reliability standard of .90 with ten classroom observations of lessons with one observer. The authors admit that this is (at best) impractical in public school settings and would take four years to complete using their current observational schedule.

Other classroom studies have produced similar results. In a comprehensive review of teacher evaluation for the Gates Foundation conducted with over 1000 teachers in grades 4-8, Kane et al. (2012) found that four visits to classrooms only produced a reliability of .65, and that adequate reliability could only be achieved with multiple raters. In other teacher accountability studies (Halpin & Kieffer, 2015), observations of classes taken at different occasions likewise varied substantially; these researchers recommended eight or more observations of teachers to create reliable teaching profiles. Mashburn et al. (2013) looked at the quality of teacher-student interactions in 5th and 6th grade and saw that these interactions not only varied between days, but also at different times within days with 17% to 22% occasion variance depending upon which observational measure was used. With one rater, reliabilities remained in the .70 range with four classroom visits. In a G-study of high school history teachers, Huijgen et al. (2017) had much lower lesson variability with 2% variation. In their study, adequate reliabilities could be reached with only four observations.

G-studies have also been conducted in other areas such as psychology and child development. In an observational study of infant behaviors, Lei et al. (2007) achieved adequate reliability with 10 observations and two raters. These authors did not report variance percentages, but variance components for occasions were over ten times larger than those for observers. Hintze et al. (2000) used ten observations of reading fluency and saw 8% of total variability due to observations.

Most of the G-studies cited above are found in K-12 contexts and are thus oriented toward the practical realities of evaluating teachers in public schools. Due to resource limitations, in many cases it is impossible to visit any given classroom more than a few times per year (Huijgen et al., 2017). Having more than one observer visit a classroom is another way to improve reliability but is also difficult to implement. In some cases, video can be substituted for an in-person visit making multiple raters more feasible, but still requires equipment and logistical work to record class sessions (Lee et al., 2017). For research and evaluation purposes in higher education, it may be easier to make more frequent visits than is possible in K-12 teacher evaluation (Smith et al., 2014). However, it is important to know how many visits are necessary for a reliable measure, given the still limited resources available for evaluation and research on teaching, especially in higher educational contexts.

Research Questions

This study was part of a larger effort carried out to compare classroom observations of teaching with instructor self-characterizations of their teaching on a survey. We sought to characterize and quantify the extent of particular classroom activities, such as lecture and group work, in a sample of college mathematics classes. In order to confidently make course-level

estimates of teaching from class observations, we sought to answer the following research questions:

1. How did rater agreement and bias differ for different activity codes? How much rater error was present when codes were combined?
2. How many observations are needed to make a reliable measure over a semester?

The study occurred in two stages. First, we wanted to know if our raters had been trained to a standard high enough so that evaluation could be conducted with one rater. At the same time, we also wanted to learn if rater agreement varied substantially given which activity code or teaching behavior was observed. The primary emphasis of our study (research question two) was to examine class-to-class variability with a generalizability study. We wanted to know how many observations were needed in our data to provide a reliable estimate of instructors' teaching over a semester.

Research Methods

Instruments

We gathered data for multiple math instructors involved in a validity study comparing teacher survey responses to observed teaching. Our observational protocol is part of a broader study matching survey responses to observational data. After reviewing various observation protocols, we started with the COPUS (Smith et al., 2013), which draws heavily from the TDOP protocol (Hora et al., 2013). We modified these protocols to reflect teaching practices common in undergraduate mathematics classrooms, but kept the TDOP's segmented, descriptive approach. The resulting protocol is the Toolkit for Assessing Mathematics Instruction (TAMI-OP) (Hayward et al., 2018). At two-minute intervals during the class, observers coded for the

presence (yes/no) of 11 student behaviors and 9 instructor behaviors. We called these categories *activity codes* or more generally, observation *items*. In addition, observers counted the frequencies of three types of student questions and answers, and three types of instructor questions and answers. We finally decided upon eight activity codes in the analysis for the rater study, and ten codes for the occasions study; the remaining codes were absent or seen very infrequently in the data. We were able to add items in the occasions study by separating types of questions (informational and reasoning) asked by teachers and answered by students, but needed to combine question type in the rater study due to the smaller amount of data available for this study. Observers also completed 12 information questions that identified and described the class being observed, such as the date, the class name, and the day of the week. Table 1 lists the activity codes and their description for our study.

Samples and Research Design

Our wider sample included 177 in-person class observations from 16 courses and 15 teachers. This included 4789 two-minute observations, or nearly 160 hours of observations. Observations were carried out over two terms at three public universities. Courses included College Algebra, Calculus, Geometry, Statistics, and Advanced Mathematical Modeling. All courses were on semester schedules. The work here reflects two phases of the wider study. In our rater study, we used the observational protocol with two raters to establish interrater reliability. Then, in the occasions study, one rater observed teachers to gather data to learn how many observations were needed for a reliable measure.

For our generalizability studies, we randomly sampled within the wider data set to create two balanced datasets. The first dataset (rater study) included 2 raters (r), 4 teachers (t), 8 items (I) and 4 classes nested within each teacher's course (c:t) with 25 two-minute observations

within each class (d). The second data set (the occasions study) examined teachers (13), classes (9), items (10), and observations (23). Table 2 presents our G-study design.

Generalizability and Decision Studies

Generalizability theory is a complex method of determining the reliability of measures and the source of variability in a measure (Brennen, 1992; Marcoulides, 1989). Similar to Analysis of Variance (ANOVA), G-studies calculate variance components, or the amount of variance attributable to different sources. G-studies have specific nomenclature and notation that is used in our study. The elements of generalizability and decision studies are described below with our methods.

Facets

Similar to factors in ANOVA, facets are the main sources of variability in a score or summative observation. In our study, we used facets for *raters, teachers,* and *class sessions.* The objects of measurement are the actual two-minute observations coded 0 and 1; these are denoted "d" for data in our analyses, although technically they are not a facet.

True and Error Score Variance

Some sources of variance (such as teachers) are considered "true" variance, with naturally occurring differences between teachers and items. In contrast, "error" variance is considered spurious or "noise", such as the differences in judgments between two independent raters scoring the same test. In our study, rater variation is considered error because raters may disagree if an activity is present or not during a two-minute period, and one rater may tend to see the activity during a class consistently more than another rater (sometimes called rater bias). We also considered class-to-class variation as error in our analysis. While nominally class-to-class

variation is to be expected, here it is considered error given that we are attempting to generalize what we see in a few class sessions to a whole course term. As noted, this is a common practice in other generalizability studies that examine the number of occasions classes have been observed (Hill et al., 2012; Briggs & Alzen, 2019). Some researchers also call these "differentiation" and "instrumentation" sources of variance corresponding to true and error variance (Cardinet et al., 2011).

Fixed and Random Variables

Variables used in G-studies can be fixed, random, or finite random depending upon the "universe of generalization" posited by a researcher. *Fixed variables* have a finite number of levels such as gender, race/ethnicity, or grade level. The elements of a *random variable* are (in theory) interchangeable and the observed units sampled are meant to generalize to an infinite universe of similar units. Raters and teachers were considered random in our study. *Finite random variables* have a universe of a known size. In our case, we knew that semesters have 45 class sessions, so we considered classes sampled from this universe as a finite random variable. We considered activity codes or items as fixed variables given that the observations did not obviously come from a wider generalizable universe of similar activities, and the activities and behaviors in any given class are finite in number. In the notation, random and random finite variables use lower case letters and fixed variables are capitalized.

Crossed and Nested Variables

As in ANOVA, facets can be crossed with each other or nested. For instance, for crossed variables, all levels of one facet are seen in another variable. In our study, all teachers (t) in our sample are observed for each activity code (I), signified by the notation tI, "teachers crossed with items." The number of crossed terms can be numerous with three- and four-facet combinations

(e.g. rIt). Nested variables are contained within the levels of other variables. In our study, class sessions (c) are nested within teachers (t), written c:t, or "classes nested within teachers." This means that each teacher instructs only their own students in multiple classes during a semester.

Reliability Coefficient

The reliability coefficient summarizes (on a scale of 0 to 1) the proportion of true score variation to total variation.

We use coefficient_G (G), found by the formula:

$$G = 1 - (\sigma^2_e/(\sigma^2_e + \sigma^2_t)$$

where σ^2_e is error variability (in our case all error due to (and interacting with) raters and class sessions and residual error, and σ^2_t is true score variability from teachers and items:

$$\sigma^2_e = \sigma^2_{residual} + \sigma^2_{rater} + \sigma^2_{c:t} + \sigma^2_{rater\ x\ class:teacher} + \sigma^2_{rater\ x\ item}$$

$$\sigma^2_t = \sigma^2_{teacher} + \sigma^2_{item} + \sigma^2_{item\ x\ teacher}$$

The G-coefficient is derived from variance components for each facet and facet interaction given the empirical units in each facet. This means we report the G-coefficient for the G-study as performed, in our case with 2 raters or 9 class observations.

Decision Studies

Decision, or D-studies are extrapolations of empirical results adjusting for differing numbers of raters or class sessions. D-studies are possible given the mathematical structure of G, which segments each source of variance and then divides this variance by the number of units used to compute the variance. Extrapolations are made by substituting the existing numbers for the empirical estimate with an extrapolated value in the denominator of the formula. We made

estimates of reliability for one to four raters, and two to 14 class sessions to learn how many raters or classes are needed to achieve a reliable measure. One aspect of D-studies that can be counterintuitive is the concept of "reliability of one rater." In practice it is impossible to gauge agreement or rater bias with one observer, so the estimate of one rater's reliability is necessarily a purely mathematical concept. Also, in research studies like ours with one rater visiting multiple classrooms, the extent of the rater's bias in practice is unknowable without sending different raters to the same classrooms. The elements of our G-study are listed in Table 3.

Other Methodological Considerations

The standard of reliability we used is a G equal to or above .80. This is more of a rule of thumb than a hard-and-fast standard, but is common in many research studies gauging rater reliability, reliability of survey composite variables, and tests involving rater agreement, especially in group versus individual contexts (Kottner et al., 2011). As mentioned above (Van der Lans et al., 2016), some researchers use a higher standard of .90 for summative studies, especially in high stakes contexts where consequential decisions are made about individuals.

All analyses were conducted in EduG 6.0 (Cardinet et al., 2011), a software package designed for conducting G- and D- studies.

Results

We examined two models. The *rater study* examined only rater variability across activity codes and teachers depending upon what is observed. The *occasions study* used all available data to examine variability due to class-to-class variation. Results of both G-studies (the reliability of the empirical data), and D-studies, extrapolations of reliability for different numbers of units for each facet are presented.

Rater Study

The first G-study found reliabilities for the *rater* by *teacher* design for each activity code. We wanted to learn how reliability varied by each activity code given that the ability for raters to agree varies substantially given what is observed. The D-study examined how reliability changed for one to four raters for each individual activity code. Figures 1 and 2 show that rater reliability increased substantially for all items from one to two raters, then leveled off with three, four and five raters. All eight activity codes meet the G = .80 standard for two raters; the four activity codes for lecturing, moving and guiding, working in groups and real time writing by instructor met the .80 standard with one rater.

Reliabilities for one rater ranged from G = .67 for *Student Questions* to G = .94 for *Moving & guiding*. Differences in rater variance are driven mainly by differences in the error component of the residual term (rtd), reflecting differing rates of agreement by raters, rater bias, and differences in rater agreement across teachers. Figures 1 and 2 present the reliability of raters for each activity code and Table 4 lists all variance percentages for each activity code separately.

We then examined the reliability of all codes combined using the "items" facet which represents all items combined. In this study we examined overall reliability due to raters and any interactions with items or teachers. The G-coefficient for this study was .91 for two raters, with reliability for one rater .84, meeting the .80 standard we set for our evaluation practice. Table 5 lists the variance decomposition for each facet for all combined activity codes in the rater study, and Figure 3 shows generalizability for number of class sessions for the combined activity codes.

Occasions Study

The occasions study examined the reliability of observations over items, teachers and classes. We used the larger dataset for this study with facets for occasions, teachers, classes, and activity codes. These data are meant to reflect real life research and evaluation on teaching with one rater visiting numerous class sessions and courses. We assume with this study that an acceptable level of rater reliability has been reached through training and piloting, as was the case for our study. The occasions study also has the advantage of having a larger sample of classes, teachers, and items than was possible with the rater study.

The percentage of variance for each facet of the study is shown in Table 6, and reliability across observations in Figure 4. Researchers would need 11 observations needed to reach the .80 standard. Again, most error in this model (19.6%) is due to class-to-class variability across items (Ic:t). The large error component (76.9%) reflects the "worst case" scenario of only having one classroom observation.

Discussion

Summary

We examined how many raters and class observations are needed to reach an adequate standard of reliability for direct observation of undergraduate mathematics teaching faculty. We first established that observations could be made reliably with one rater (G = .84). In the occasions study we then found that one rater needed 11 classes to reliably observe all eight activity codes. We believe the latter model provides a good estimate of conditions experienced in

many research and evaluation contexts where observations are used and one rater is sent out to collect observations.

We also found that rater reliability varied depending upon what activity is being observed. For single activity codes, rater reliability (without the class-to-class error component) varied; four codes did not reach the .80 reliability standard for one rater. Activity codes with lower reliabilities primarily concerned those requiring agreement on what counted as a teacher or student question. It was more difficult to agree on what constituted a valid observation of a question during any given two-minute period for these activities, perhaps because of the short duration of most teacher or student questions, teachers' use of rhetorical questions, and discrepancies in coding when instructors rephrase the same question in multiple ways. Observers may also see discrepancies in coding when instructors rephrase the same question in multiple ways. Higher reliability activities included more continuous activities such as lecture or group work. It should be noted that in field conditions, the reliability of a score would be evaluated by all observations bundled together. The activity codes with higher reliability would (in effect) pull up the overall quality of the measure. Being aware of these differences in reliability across activity codes can inform practice such as directions in technical manuals (Hora et al., 2013); observers using related observational protocols may want to be aware that some activity codes are more difficult to reliably observe than others.

Comparison with other generalizability studies provides some context to our findings. The number of observations needed to reach a reliable measure is a function of the variability of class-to-class averages for activity codes as well as the overall design of the G-study. For our study, the amount class-to-class variance was 19% for the occasions study. In Table 7 we compared our variance components with those from other published studies. The percentage of

occasion variance in many of the studies was similar to our current research. For instance, the amount of variability due to occasions was between 6% and 39% in the Hill et al. (2012), Masburn et al. (2014), Newton (2010) and Kane et al. (2012) studies. The Huijgen et al. (2017) study stood out as showing very little variance over occasions. Overall, the preponderance of studies pointed to the fact that teachers vary what they do in classrooms from day-to-day, and that our current study is similar to others in the amount of variability we observed.

Information about the number of classes needed for a reliable measure was less straightforward to compare. Many studies juggle the number of raters and occasions when conducting decision studies. In separate studies by Hill, Newton, and Kane referenced below, four or six occasions were needed to reach adequate reliability, but this level of reliability was only possible with four raters. In studies with one rater, the number of occasions were similar to ours, with eight or 10 observations needed; and in some cases adequate reliability was not reached even with 10 observations (Mashburn et al., 2014).

Implications for Research and Evaluation Practice

These results show that the numbers of observations needed still remain much higher than is seen in general research and evaluation practice (Cohen & Goldhaber, 2016). Anyone contemplating conducting consequential teacher evaluation or research based on observational studies should consider that a substantial number of observations may be needed to reliably estimate the frequency of specific teaching activities over a semester.

Calling for more observations has serious implications for research and evaluation designs used to assess teaching methods when the intent is to generalize to a semester, or even to generalize to overall teaching style. Our findings are most relevant to those using observational data for comparative research or evaluation studies in mathematics higher education because the

protocol we used was designed for that purpose, but could also apply to faculty evaluation by department heads or others visiting classrooms. Using observations as an outcome measure (or even as an independent variable) raises the standard for the rigor and statistical power needed to make confident claims. Observational data can be used in most of the common comparative research designs to evaluate the effect of an intervention (such as professional development), "before and after" comparisons, or compare outcomes of participating and comparison groups.

Consider a (fictional) study where researchers are evaluating the effects of a summer workshop on inquiry-based learning. The researchers have made baseline observations of instructors' teaching style during the spring semester using a segmented observational protocol. After the teachers participate in the workshop, observers visit classrooms during the fall and spring semesters to learn if instructors did in fact change their teaching, and if so, what changes they made. While the ability to detect a pre-to-post difference in overall teaching is a function of the number of teachers participating in the study (statistical power), the actual comparison is only as good as the reliability of the measure used. As the present analysis demonstrates, if the researcher only uses two or three observations to calculate an average, reliabilities are so low as to be essentially meaningless as an estimator for a semester-long course. Another way of framing this is by creating confidence intervals based on class-to-class variability as characterized by the standard deviation of class-to-class variation in an activity code. Even with a conservative standard deviation of 15% among classes over a semester, if an instructor was estimated to lecture 50% of the time, the real percentage over a semester could be between 33% and 67% if only three observations are used.[1] This would be a wide range to work within, and low and high estimates would give qualitatively different pictures of how much an instructor lectured. When

unreliable measures are used in comparative research designs, compounding measurement error over multiple individuals and groups does nothing to improve a study.

Also affected by these findings are validity comparisons between teachers' self-report about their practices, and observations of their actual teaching (Ebert-May et al., 2011; Lewin et al., 2014), the original impetus for our study. For example, in our ongoing study we wanted to know if teachers accurately report their practices on a survey, and if a survey could be used as a proxy for more costly observations when evaluating professional development programs. Because the comparison rested on the assumption that we had a trustable criterion measure in the observations, that criterion had to be highly reliable. The primary danger in using a measure with poor reliability in comparing survey and observation data is the possibility of inaccurate characterizations of teaching if the observer only sees a teacher on a day (or days) when they do an atypical activity. If the criterion measure is not trustable and reliable, it is essentially impossible to know if an instructor's survey report of their activity does or does not represent their practice over a course term. To validate survey items characterizing course-level teaching against observations, we must ensure that the observations are themselves trustworthy reports of teaching practice by observing a sufficient number of classes to achieve a representative and reliable measure.

Another common use of observations, as stated, is to describe or assess an individual instructor's teaching methods for personnel evaluation. Increasingly for undergraduate faculty, observations are part of a teacher evaluation with supervisors (or peers) rating an instructor's overall teaching style, their interaction with students, or their use of technology in the classroom (Smith et al., 2017). For teacher or faculty evaluations with consequences such as promotion, pay

raises, or retention, sufficient rigor in reliability is required to make fair assessments; although in many cases frequent observations may not be logistically, financially or politically feasible.

We believe that in some cases triangulation with other methods can be used to leverage a reliable measure with fewer observations. For instance, we can ask instructors to provide a syllabus or schedule of activities for a semester that provides information about scheduled activities such as group work or presentations. We have found that this can provide a better source of information about the relative frequency of these broadly defined activities than observations made throughout the semester. Instructors can also be interviewed about their teaching methods and the amount of class-to-class variability that they believe is present. As shown above (Kane et al., 2012), increasing the number of raters will also increase reliability and lower the number of observations needed; this is made easier in practice with video coding than in-person classroom visits. However, adding raters only works to increase reliability if there is no ceiling for improvement. For our study, adding third or fourth raters did not make the observation substantially more reliable.

For many purposes, statistical considerations for reliability and sampling may be relaxed. For example, department heads, faculty peers, or pedagogical experts may observe teachers and provide feedback from coding or rating schemes augmented with expert assessments and qualitative descriptions of how effectively instructors are implementing their instruction, while suggesting ways to improve practices. Fewer observations raise less concern for these purposes because the quantitative data are used to support qualitative descriptions and spark discussion, not to make high-stakes decisions (Kane et al., 2012). Additionally, both instructors and observers can place the observation in context during the discussion, so ensuring that the sample of observations is representative of the full course is less of a concern. Still, expanding the

number of observations for classroom observations may be a good idea even in these cases, not only to provide more stable and reliable assessments of the frequency and nature of teaching activities, but also by overcoming possible reactive effects with teachers who may experience difficulty performing in front of an observer, or who put their best foot forward for a one-time observer.

Limitations of our study

Our study has three primary limitations. First, these are the results of a single research study; others conducting similar studies may encounter different conditions that lead to higher or lower reliabilities and number of observations needed. However, as shown above, other decision studies using class occasions as a facet have found somewhat similar results to ours, suggesting that our findings are not atypical. Secondly, our study is conducted with a small group of teachers which may not have the power to generalize to undergraduate STEM mathematics instructors in the US. Third, our rater study was, in all likelihood, too small to fully and accurately represent the class-to-class variation we later saw during our occasions study. Our presentation of class variation in the one-rater studies is a better estimate of class-to-class error variance, but did not include the rater facet.

We believe that overcoming rater disagreement and bias is possible through training and piloting. Findings in the first result section show that we reduced this amount of error through training, and in fact, rater error is less of a concern than class-to-class variability in making reliable estimates of a teacher's activities over the course of a semester.

Acknowledgments

We sincerely thank the instructors and students who welcomed us into their classrooms, and the site coordinators who assisted with our recruitment efforts. The statistics presented here obscure all of your individual contributions, but we value them highly and wish we could thank each of you publicly. We thank Myles Boylan from NSF for his sustained interest and encouragement, and Tim Archie from the University of Colorado who also helped with analysis of some observation and survey data. We also thank Holly Devaul for her help in editing.

This study was supported by the National Science Foundation under awards DUE-1245436 and DUE-1821704. Additional data collection was supported by the Spencer Foundation. All findings and opinions are those of the authors and not the funder.

References

Adamson, S. L., Banks, D., Burtch, M., Cox, F., Judson, E., Turley, J. B., Benford, R., & Lawson, A. E. (2003). Reformed undergraduate instruction and its subsequent impact on secondary school teaching practice and student achievement. *Journal of Research in Science Teaching*, *40*(10), 939-957. https://doi.org/10.1002/tea.10117

American Association for the Advancement of Science (AAAS) (2013). *Describing & measuring undergraduate STEM teaching practices: A report from a national meeting on the measurement of undergraduate science, technology, engineering and mathematics (STEM) teaching, December 17-19, 2012*. Washington, DC: AAAS. Accessed 11/14/14 from http://ccliconference.org/files/2013/11/Measuring-STEM-Teaching-Practices.pdf

Amrein-Beardsley, A., & Osborn Popp, S. E. (2011). Peer observations among faculty in a College of Education: Investigating the summative and formative uses of the reformed teaching observation protocol (RTOP). *Educational Assessment, Evaluation and Accountability*, *24*(1), 5-24. https://doi.org/10.1007/s11092-011-9135-1

Auerbach, A. J., & Schussler, E. (2016). Research and teaching: Instructor use of group active learning in an introductory biology sequence. *Journal of College Science Teaching*, *045*(05). https://doi.org/10.2505/4/jcst16_045_05_67

Beach, A. L., Henderson, C., & Famiano, M. (2008). 13: Co-teaching as a faculty development model. *To Improve the Academy*, *26*(1), 199-216. https://doi.org/10.1002/j.2334-4822.2008.tb00509.x

Blanchard, M. R., Southerland, S. A., Osborne, J. W., Sampson, V. D., Annetta, L. A., & Granger, E. M. (2010). Is inquiry possible in light of accountability?: A quantitative

comparison of the relative effectiveness of guided inquiry and verification laboratory instruction. *Science Education*, *94*(4), 577-616. https://doi.org/10.1002/sce.20390

Brewer, C., & Smith, D. (2011). Vision and change in undergraduate biology education: A call to action. Washington, DC: American Association for the Advancement of Science.

Briggs, D. C., & Alzen, J. L. (2019). Making Inferences About Teacher Observation Scores Over Time. *Educational and Psychological Measurement*, *79*(4), 636-664.

Budd, D. A., van der Hoeven Kraft, K. J., McConnell, D. A., & Vislova, T. (2013). Characterizing teaching in introductory geology courses: Measuring classroom practices. *Journal of Geoscience Education*, *61*(4), 461-475.

Bowling, B. V., Acra, E. E., Wang, L., Myers, M. F., Dean, G. E., Markle, G. C., Moskalik, C. L., & Huether, C. A. (2008). Development and evaluation of a genetics literacy assessment instrument for undergraduates. *Genetics*, *178*(1), 15-22. https://doi.org/10.1534/genetics.107.079533

Brennan, R. L. (2013). *Generalizability theory*. Springer Science & Business Media.

Cardinet, J., Johnson, S., & Pini, G. (2011). *Applying generalizability theory using EduG*. Taylor & Francis.

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly*, *27*(3), 529-542. https://doi.org/10.1016/j.ecresq.2011.12.006

Campbell, T., Der, J. P., Wolf, P. G., Packenham, E., & Abd-Hamid, N. H. (2012). Scientific Inquiry in the genetics laboratory: Biologists and university science teacher educators collaborating to increase engagement in science processes. *Journal of College Science Teaching*, *41*(3), 74-81.

Carins, J. (2016). Visual observation techniques. *Formative Research in Social Marketing*, 107-123. https://doi.org/10.1007/978-981-10-1829-9_7

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378-387. https://doi.org/10.3102/0013189x16659442

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Wadsworth Publishing Company.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8-15. https://doi.org/10.1177/003172171209300603

Darling-Hammond, L. (2015). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. Teachers College Press.

Ebert-May, D., Derting, T. L., Hodder, J., Momsen, J. L., Long, T. M., & Jardeleza, S. E. (2011). What we say is not what we do: Effective evaluation of faculty professional development programs. *BioScience*, *61*(7), 550-558. https://doi.org/10.1525/bio.2011.61.7.9

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410-8415. https://doi.org/10.1073/pnas.1319030111

Gardner, F. (2000). Methodological issues in the direct observation of parent–child interaction: Do observational findings reflect the natural behavior of participants? *Clinical child and family psychology review*, *3*(3), 185-198.

Grossman, P., Cohen, J., & Brown, L. (2015). Understanding instructional quality in English Language Arts: Variations in PLATO scores by content and context. *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project*, 303-331.

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6), 1-32.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64-74. https://doi.org/10.1119/1.18809

Halpin, P. F., & Kieffer, M. J. (2015). Describing profiles of instructional practice. *Educational Researcher*, *44*(5), 263-277. https://doi.org/10.3102/0013189x15590804

Hayward, C. N., Weston, T. J., & Laursen, S. L. (2018). First results from a validation study of TAMI: Toolkit for Assessing Mathematics Instruction. In A. Weinberg, C. Rasmussen, J. Rabin, M. Wawro, & S. Brown (Eds.), Proceedings of the 21st Annual Conference on the Research in Undergraduate Mathematics Education (pp. 727-735). San Diego, CA: Mathematical Association of America, SIGMAA on RUME.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough. *Educational Researcher*, *41*(2), 56-64. https://doi.org/10.3102/0013189x12437203

Hora, M., Olson A., & Ferrare J. J. (2013). *Teaching Dimensions Observation Protocol (TDOP) User's Manual*. Madison: Wisconsin Center for Education Research, University of Wisconsin–Madison; 2013. http://tdop.wceruw.org/Document/TDOP-Users-Guide.pdf

Hora, M., & Ferrare, J. (2014). Remeasuring postsecondary teaching: How singular categories of instruction obscure the multiple dimensions of classroom practice. *Journal of College Science Teaching*, *043*(03). https://doi.org/10.2505/4/jcst14_043_03_36

Huijgen, T., Van de Grift, W., Van Boxtel, C., & Holthuis, P. (2016). Teaching historical contextualization: The construction of a reliable observation instrument. *European Journal of Psychology of Education*, *32*(2), 159-181. https://doi.org/10.1007/s10212-016-0295-8

Iobst, W. F., Sherbino, J., Cate, O. T., Richardson, D. L., Dath, D., Swing, S. R., Harris, P., Mungroo, R., Holmboe, E. S., & Frank, J. R. (2010). Competency-based medical education in postgraduate medical education. *Medical Teacher*, *32*(8), 651-656. https://doi.org/10.3109/0142159x.2010.500709

Kane, T. J., Staiger, D. O., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., & Parker, D. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Bill & Melinda Gates Foundation.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International Journal of Nursing Studies*, *48*(6), 661-671. https://doi.org/10.1016/j.ijnurstu.2011.01.016

Lane, E., & Harris, S. (2015). Research and teaching: A new tool for measuring student behavioral engagement in large university classes. *Journal of College Science Teaching*, *044*(06). https://doi.org/10.2505/4/jcst15_044_06_83

Laursen, S. L., Hassi, M.-L., Kogan, M., & Weston, T. J. (2014). Benefits for women and men of inquiry-based learning in college mathematics: A multi-institution study. *Journal for Research in Mathematics Education*, *45*(4), 406-418.

Lee, D., Arthur, I. T., & Morrone, A. S. (2015). Using video surveillance footage to support validity of self-reported classroom data. *International Journal of Research & Method in Education*, *40*(2), 154-180. https://doi.org/10.1080/1743727x.2015.1075496

Lei, P., Smith, M., & Suen, H. K. (2007). The use of generalizability theory to estimate data reliability in single-subject observational research. *Psychology in the Schools*, *44*(5), 433-439. https://doi.org/10.1002/pits.20235

Lewin, J. D., Vinson, E. L., Stetzer, M. R., & Smith, M. K. (2016). A campus-wide investigation of clicker implementation: The status of peer discussion in STEM classes. *CBE—Life Sciences Education*, *15*(1), ar6. https://doi.org/10.1187/cbe.15-10-0224

Lund, T. J., Pilarz, M., Velasco, J. B., Chakraverty, D., Rosploch, K., Undersander, M., & Stains, M. (2015). The best of both worlds: Building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice. *CBE—Life Sciences Education*, *14*(2), ar18. https://doi.org/10.1187/cbe.14-10-0168

Marcoulides, G. A. (1989). The application of generalizability analysis to observational studies. *Quality and Quantity*, *23*(2), 115-127. https://doi.org/10.1007/bf00151898

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*, *15*(2), 146-155. https://doi.org/10.1007/s11121-012-0357-3

Mertler, D. C. (2016). *Classroom assessment: A practical guide for educators*. Routledge.

Michael, J. (2006). Where's the evidence that active learning works? *Advances in Physiology Education*, *30*(4), 159-167. https://doi.org/10.1152/advan.00053.2006

Murray, H. G. (1991). Effective teaching behaviors in the college classroom. In J. C. Smart (ed.), *Higher education: Handbook of theory and research*. (Vol. 7). New York: Agathon Press.

Murray, H. G. (1997). Effective teaching behavior in the college classroom. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 171-204). New York, NY: Agathon Press.

Nadelson, L. S., Shadle, S. E., & Hettinger, J. K. (2013). A journey toward mastery teaching: STEM faculty engagement in a year-long faculty learning community. *Learning Communities Journal*, *5*, 97-122.

Naeem, N. (2013). Validity, reliability, feasibility, acceptability and educational impact of direct observation of procedural skills (DOPS). *J Coll Physicians Surg Pak*, *23*(1), 77-82.

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*(1-2), 1-13. https://doi.org/10.1016/j.stueduc.2010.10.002

Nolan, B. A., Mathews, R. M., Truesdell-Todd, G., & VanDorp, A. (2002). Evaluation of the effect of orientation cues on wayfinding in persons with dementia. *Alzheimer's Care Today*, *3*(1), 46-49.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109-119. https://doi.org/10.3102/0013189x09332374

Pilburn, M., Sawada, K. Falconer, J. Turley, R. Benford, I. (2000) *Reformed Teaching Observation Protocol (RTOP).* Tempe, AZ: Arizona Collaborative for Excellence in the Preparation of Teachers.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics*, *102*(6), 245-253.

Smith, M., Wood, W., Krauter, K., & Knight, J. (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. *CBE—Life Sciences Education*, *10*(1), 55-63. https://doi.org/10.1187/cbe.10-08-0101

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The classroom observation protocol for undergraduate stem (Copus): A new instrument to characterize university stem classroom practices. *CBE—Life Sciences Education*, *12*(4), 618-627. https://doi.org/10.1187/cbe.13-08-0154

Smith, M. K., Vinson, E. L., Smith, J. A., Lewin, J. D., & Stetzer, M. R. (2014). A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions. *CBE—Life Sciences Education*, *13*(4), 624-635. https://doi.org/10.1187/cbe.14-06-0108

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan, M. K., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M., McDonnell, L. M., McKay, T. A., Michelotti, N., Musgrove, A., Palmer, M. S., Plank, K. M., … Young, A. M. (2018). Anatomy of STEM teaching in North American universities. *Science*, *359*(6383), 1468-1470. https://doi.org/10.1126/science.aap8892

Stains, M., Pilarz, M., & Chakraverty, D. (2015). Short and long-term impacts of the Cottrell scholars collaborative new faculty workshop. *Journal of Chemical Education*, *92*(9), 1466-1476. https://doi.org/10.1021/acs.jchemed.5b00324

Tashakkori, A., & Teddlie, C. (2010). SAGE handbook of mixed methods in social & behavioral research. https://doi.org/10.4135/9781506335193

Turpen, C., & Finkelstein, N. D. (2009). Not all interactive engagement is the same: Variations in physics professors' implementation of Peer instruction. *Physical Review Special Topics - Physics Education Research*, *5*(2). https://doi.org/10.1103/physrevstper.5.020101

Van der Lans, R. M., Van de Grift, W. J., Van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, *50*, 88-95. https://doi.org/10.1016/j.stueduc.2016.08.001

Waxman, H. C., & Padron, Y. N. (2004). The uses of the Classroom Observation Schedule to improve classroom instruction. In H. C. Waxman, R. G. Tharp, & R. Soleste Hilberg (Eds.), *Observational research in U.S. classrooms: New approaches for understanding cultural and linguistic diversity* (pp. 72-96). Cambridge, England: Cambridge University Press.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. *Handbook of statistics*, *26*, 81-124.

West, E. A., Paul, C. A., Webb, D., & Potter, W. H. (2013). Variation of instructor-student interactions in an introductory interactive physics course. *Physical Review Special Topics - Physics Education Research*, *9*(1). https://doi.org/10.1103/physrevstper.9.010109

Wragg, E. C. (2012). *An introduction to classroom observation*. Routledge.

[1] This is based on the standard error for the mean adjusted for a finite sample of 45 class sessions in a semester. The standard deviation is for the class-to-class variability of observations. We found that the standard deviation for lecture was 22%. Making a more conservative estimate of a 15% standard deviation with 3 observations would give a standard error of 15/1.73 = 5.7. The 95%CI is 1.96 * 5.7 = 16.97. Adjusted for a finite sample this is reduced slightly to 16.58.

Table 1

Activity Codes and Their Descriptions Used in Study

Activity Code	Description
Instructor question (combined for rater study)	Instructor asks question
Student answers question (combined for rater study)	Students answers teacher question
Student question	Students ask question of teacher
Reviewing content	Instructor reviews students' previous work (e.g. homework, group activity)
Realtime writing by instructor	Instructor writes on board, overhead or whiteboard
Moving & guiding	Instructor works with students in groups
(For occasions study, replaces combined instructor question and student answer)	
Instructor asks informational question	Instructor question asking for specific information or answer
Instructor asks for reasoning	Instructor question asks for students to explain answer to problem
Student answers with information	Student answers with specific information or answer
Student answers with reasoning	Student answers with explanation of problem or concept

Table 2

Design and Numbers in Each Facet of G-studies

	Facets					
Study	Raters	Teachers	Items	Classes within teacher	Observations within class	Total data elements
Rater	2	4	(8)	4	25	800
Occasions	1	13	10	9	23	26910

Note. Observations within class are the objects of measurement.

Table 3

Elements of the G-study

Facet	Notation	Type	True/Error	Crossed or nested
Rater	*r*	Random	Error	Interacts with all other facets
Activity code ("item")	*I*	Fixed (I)	True	Crossed with all other facets
Teacher	*T*	Random	True	Crossed with all other facets
Class	*c, c:t*	Finite random	Error	Nested within teacher
Data/ Observations (object of measurement)	*d*	Finite random		Object of measurement, observations are nested within each class.

Note. We avoided using the letter "o" for the object of measurement given to differentiate between classroom observation (how many classes observed) and how many two-minute observations within each class.

Table 5

Variance Decomposition for Each Facet All Combined Activity Codes in Rater Study

	Facets						
Variance percentage	Teacher (t)	Item (I)	Teacher by item (tI)	Item by data (Id)	Teacher by item by data (tId)	Data (d)	Teacher by data
True (83.8%)	0.1%	7.4%	6.8%	1.2%	61.5%	1.1%	5.7%
	Rater (r)	Rater by item (rI)	Rater by teacher by data (rtd)	Rater by data (rd)	Item by rater by teacher (Irt)	Item by rater by data within teacher (Irt:d)	
Error (16.2%)	0.2%	.2%	1.9%	.1%	.2%	13.7%	
G = .91 (Two raters)							
G = .87 (One rater)							

Note. Facets are raters = 2, teachers = 4, Items = 8, data =100.

Table 6

Variance Decomposition for Each Facet for Occasions D-study.

	Facets					
Variance percentage	Teacher (t)	Item (I)	Data (d)	Teacher by item (tI)	Item by data (Id)	Teacher by item by data (tId)
True (23.1%)	0.1%	5.5%	0.2%	12.4%	1.1%	3.8%
	Class within teacher (c:t)	Item by class within teacher (Ic:t)	Class by data within teacher (cd:t)	Item by class by data within teacher (Icd:t)		
Error (76.9%)	0%	19.6%	3.1%	54.2%		

G = 0.77 (raters = 1, classes within teachers = 9)

Note. Percentage of total variance due to each source. Teacher considered random variable, class within teacher random finite (n = 45 for whole semester). Some facet interactions with zero percentages removed from table. Facets are raters = 1, teachers = 13, classes within teachers = 9, items = 10, data = 23.

Table 7

Comparison of G-studies for Variance Due to Occasions and Raters

Study author(s)	Sample	Variance % due to occasions	Number of observations needed for reliable measure at G = .80
Current study	Undergraduate teachers in mathematics	19% class within teacher variance over items	11
Van der Lans et al. (2016)	K-12 teachers in the Netherlands across subjects	Not provided	10 (.90 criteria for reliability)
Lei et al. (2007)	Infant behavior	8% observation variance	10 observations used
Halpin & Kiefer (2015)	Middle school Language Arts (ELA) teachers	Not given	8 or more
Mashburn et al. (2014)	Quality of teacher-student interactions in 5th or 6th grade classrooms	17% – 22% day- to-day occasion variance, 3% - 7% within day occasion variance for three observational measures	8 observations with one rater only reached reliabilities .70 to 0.74
Newton (2010)	Elementary through high school teachers across subjects	23% due to occasions	6 visits with 4 raters – no estimate for single rater available

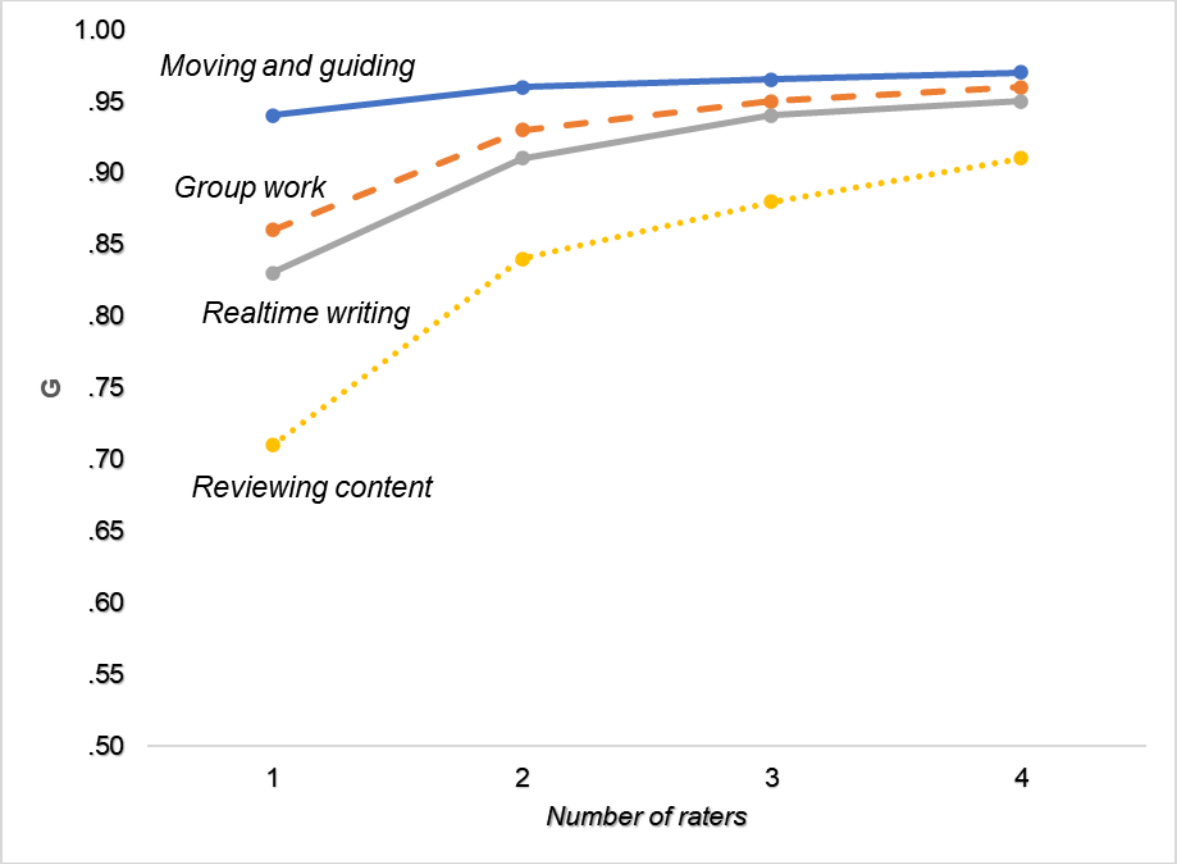
(Table 7 cont.)

Study author(s)	Sample	Variance % due to occasions	Number of observations needed for reliable measure at G = .80
Hill et al. (2012)	Middle school math teachers	Ranges from 6% to 39% for three observational dimensions	4 observations with 4 raters. Extrapolation of D-study estimates number of observations needed for one rater between 5 and 12 depending on which dimension observed.
Kane et al. (2012)	Over 1000 teachers in grades 4 – 8	27% variance for CLASS measure, 15% UTOP measure	4 observations, 4 raters gave reliability of .65
Huijgen et al. (2017)	High school history teachers	2% observation variance	4 observations

Note. Above studies sorted for number of observations needed for a reliable measure.

Figure 1

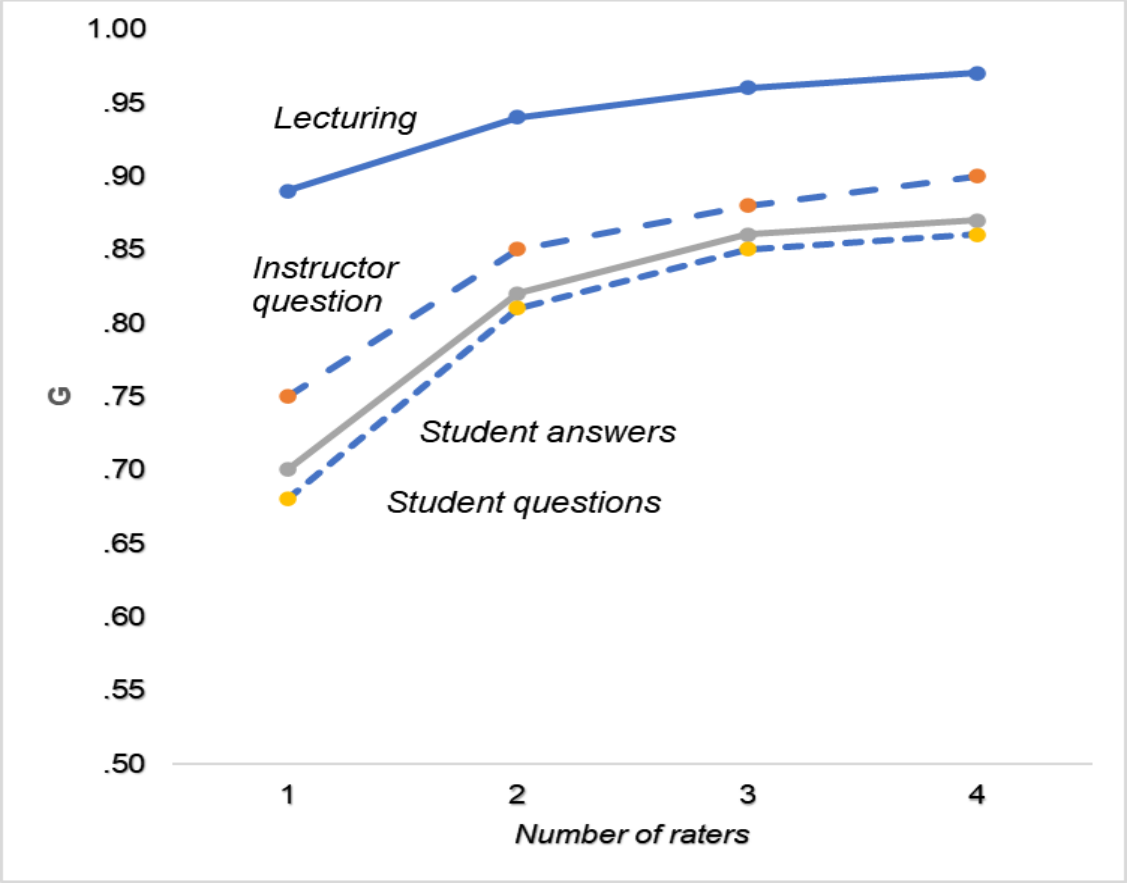
Reliability (G) By Number Of Raters for Each Activity Code, Rater Study (Part 1)



Note. Activity codes : Moving and guiding, group work, realtime writing, and reviewing content.

Figure 2

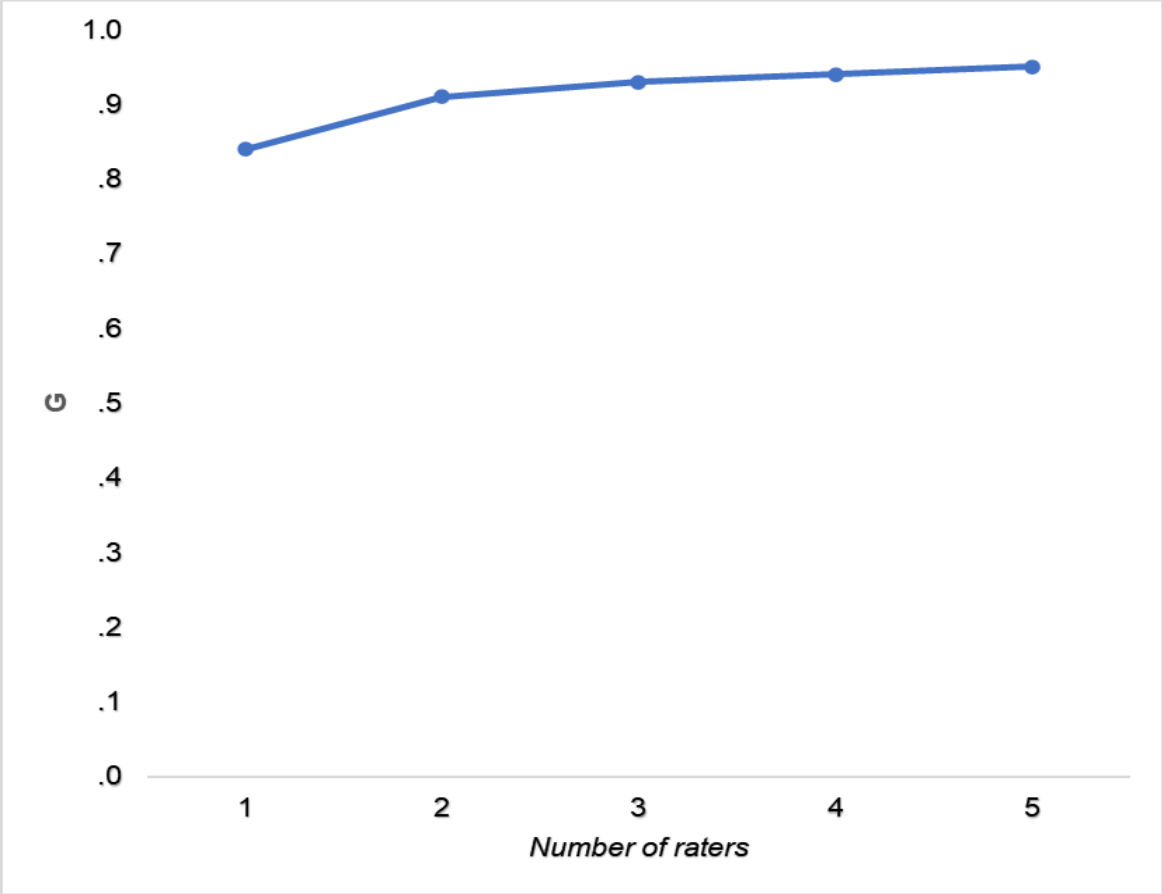
Reliability (G) by Number Of Raters for Each Activity Code, Rater Study (Part 2)



Note. Activity codes: Lecturing, instructor question, student answers, and student questions.

Figure 3

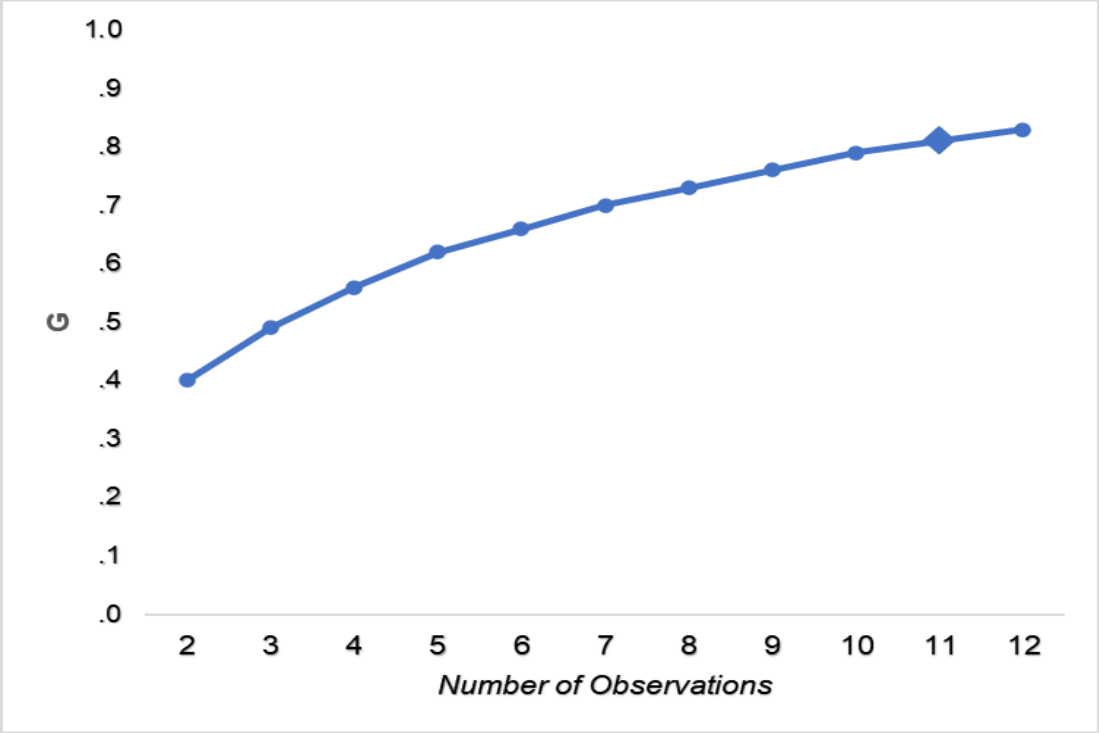
Reliability (G) for Number of Class Sessions for Rater Study



Note. Includes both raters and items as facets.

Figure 4

Reliability (G) for number of class sessions for occasion study



Note. Includes teachers, items and number of class sessions as facets.

Table 4

Variance Decomposition Percentages for Each Facet or Each Activity Code in Rater Study

	Percent of variance for each facet							
Facet	Lecturing	Instructor question	Working in groups	Student question	Student answers	Reviewing content	Real-time writing	Moving & guiding
Rater (r)	0	1.2	0	1.2	1.3	1.3	0.6	0
Teacher (t)	7.6	4.1	13.9	3.3	2.5	2.5	3.3	14.8
Data (d)	0	21.3	2.1	1.4	14.9	14.9	0	0.9
Rater x teacher (rt)	0.2	0.1	0.3	0	0.2	0.2	0	0
Rater by data (rd)	0	0.9	0.9	0	0	0	0.8	0
Teacher by data (td)	81.7	49.5	70.2	63.3	52.9	52.9	79.6	78.1
Rater by teacher by data (rtd)	10.5	23	12.5	30.8	28.2	28.2	15.7	6.1
Rater agreement	94%	92%	95%	92%	88%	97%	92%	98%
G (two raters)	.94	.85	.93	.81	.82	.83	.91	.97
G (one rater)	.89	.75	.86	.68	.70	.70	.83	.94

Note. Values are percentages. Rater and teacher considered random variables. Data is finite random. Facets are raters= 2, teachers = 4, classes within teacher = 4, data = 25