A peer reviewed, open-access electronic journal: ISSN 1531-7714

What's in a Score? A decomposition of observation scores across various sources to better understand what contributes to scores

Mark White, *University of Oslo*

Abstract: Systematized, observational approaches to measuring teaching quality are an important tool in research and practice. Termed observation systems, these approaches include a rubric that operationalizes a set of teaching quality constructs and structures to support rater training and monitoring. Scores from observation systems, through their interpretation as capturing the intended teaching quality constructs, are used to develop theoretical understandings of teaching quality. This paper explores what factors contribute to observation scores in a secondary analysis of the Understanding Teaching Quality project. Leveraging calibration data, I combine mixed-effects regression analyses of calibration data that examine rater accuracy (i.e., deviations from master scores) with analyses of operational data to explore the extent to which raters, students, teachers, and the teaching context contribute to scores. These analyses highlight that (1) some rater error may be invisible in typical analyses examining rater agreement; (2) rater error is largely systematic; and (3) differences in student composition across schools largely explain between-school differences in scores. These results highlight potential biases in estimates of score reliability and validity coefficients that might exist in studies that model rater agreement rather than rater accuracy and/or that fail to consider differences in between-teacher and between-school variation in observation scores.

Keywords: Observation System; Validity; Educational Measurement; Teaching Quality; Teacher Quality

Introduction

Systematized, observational approaches for measuring teaching quality have become a common and important tool in studying teaching (e.g., Bacher-Hicks et al., 2019; Blazar et al., 2017). These approaches, commonly called observation systems (Hill et al., 2012), include a formalized observation rubric, that

operationalizes specific, discrete teaching quality constructs, and a set of routines and procedures for conducting measurement, such as rater training and monitoring, observation procedures (Bell et al., 2019)[1]. Scores from observation systems are typically interpreted as capturing the teaching quality construct that was operationalized through the rubric. Through this interpretation, scores are then used to build an empirical picture of teaching practice and develop theoretical understandings of teaching quality.

This use of scores leads this paper to pose two research questions. First (RQ1), to what extent do scores represent the intended understanding of teaching quality, as opposed to construct irrelevant sources of variation? Using master scores as a criterion measure for the intended teaching quality constructs, I break from previous studies by exploring the accuracy of observation scores rather than examining rater agreement. This allows a more careful examination of the construct validity of scores. Secondly (RQ2), to what extent are scores capturing variation in the teaching context, characteristics of students, and/or the knowledge and skills of teachers? Here, I use mixed effects modelling to explore the relative contributions of these three factors to observation scores. Together, these research questions get at the fundamental question of this paper, what is in a score?

Prior Efforts Examining Scores from Observation Systems

Previous examinations of observation scores have taken several forms. Generalizability theory studies examine the extent to which scores vary across measurement facets (e.g., class periods or lessons, teachers, raters), providing information on how much these facets contribute to scores (e.g., Jentsch et al., 2022; Kane et al., 2012; Patrick et al., 2019). This research often finds large variation of scores across raters and lessons, which we define as a class period. Nuancing the general high variation of scores across lessons, one study found that scores of teaching quality constructs that are theoretically stable across lessons, such as classroom management, may not vary much across lessons (Praetorius et al., 2014). A reasonable amount of variation in scores is also typically observed between teachers, especially after aggregating scores across multiple lessons and raters (e.g., Kane et al., 2012; OECD, 2020).

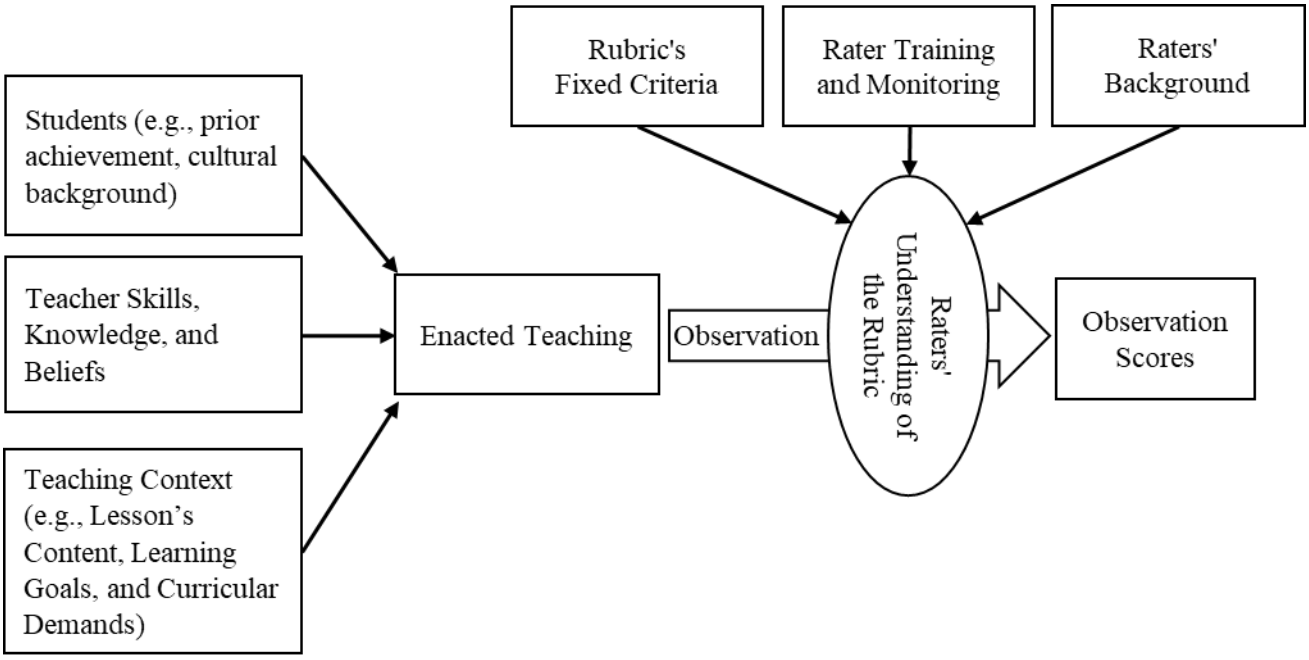
A second focus of past work has looked at the association of scores and other characteristics of instruction, the characteristics of students in the classroom, other measures of teaching and teacher quality, and student outcomes (e.g., Kane et al., 2012; Patrick et al., 2019). Here, student characteristics could include any stable attribute of students, but it is typically operationalized by student ethnicity/race, prior test scores, and/or school-based classifications, such as having an IEP. Observation scores are generally found to be strongly associated with the characteristics of students in classrooms (e.g., Cherng et al., 2022; Cowan et al., 2022) and characteristics of the lesson and classroom context (e.g., Grossman et al., 2014; Plank & Condliffe, 2013). On the other hand, the relationship between scores and other measures of teaching and teacher quality and student learning tend to be quite modest (Kelly et al., 2020). However, simulations find the size of the relationship is very sensitive to study design (van der Lans, 2018), which should be unsurprising since study design heavily impacts score reliability (Kane et al., 2012).

Conceptual Understanding of Instruction

This section presents a basic conceptual model of teaching and its measurement with observation systems to guide the analyses (see Figure 1). This model builds on the instructional triangle (Cohen et al., 2003), viewing enacted teaching as co-determined by students, teachers, and the teaching context. Here, enacted teaching is conceptualized as the interactions between teachers, students, and content within a teaching context (Hamre et al., 2013). Observation systems do not attend to this complexity directly, but simply seek to transform enacted instruction into a series of codes that are meant to represent specific

teaching quality constructs (Klette & Blikstad-Balas, 2018). However, this coding process is filtered through raters' understanding of the observation rubric. Then, while observation scores are meant to represent the teaching quality constructs embedded in the rubric, it is rather raters' understanding of the rubric's lens that is encoded in observation scores. The first research question addresses the extent to which the filter of rater's understanding of rubrics introduces bias and error, using master scores as an operationalization of the rubric-embedded understanding of teaching quality and treating deviations of scores from master scores as error. The second research question explores the relative contribution of the teaching context, students, and teacher knowledge, skills, and beliefs to observation scores. I turn now to discussing these three sets of factors and their impact on enacted teaching and observation scores.

Figure 1. Conceptual Framework for Instruction and its Measurement



Teaching Context

The teaching context includes all features associated with lessons that create a structure within which teaching interactions might occur, such as content, learning goals, curricular demands, institutional school demands (Casabianca et al., 2015; Grossman et al., 2014; Kelcey & Carlisle, 2013). Teaching is generally expected to vary across features of the teaching context, as the nature and type of interactions is targeted towards specific content and learning goals and constrained by other teaching context features. Most past work has not directly studied variation in observation scores across the teaching context but has rather examined variation in scores across lessons (e.g., Kane et al., 2012). Work that has directly studied the teaching context has found that several features of the teaching context are systematically related to scores. For example, the grade level of a classroom and content being taught seem to be systematically related to observation scores (Curby et al., 2011; Grossman et al., 2014).

Understanding the relationship between observation scores and the teaching context can contribute to both our understanding of teaching quality and the measurement of teaching quality. Consider, for example, the finding that writing instruction receives lower observation scores than reading instruction (Grossman et al., 2014). This could reflect a true difference in teaching quality that would be interesting to further explore. Additionally, this difference suggests possible improvements to uses of scores. For example, it implies that

the sampling error in teacher-level estimates would be reduced by controlling for the content of instruction (either through sampling or statistically). On the other hand, this could indicate a problem with the observation rubric not capturing aspects of teaching quality important to writing instruction, prompting the revision of the observation rubric or the restriction of the observation rubric to only reading instruction.

Students

Students actively contribute to the quality of enacted teaching (Hamre et al., 2013). For example, some students contribute to the quality of instruction by actively participating and asking questions while others may hinder quality instruction by interrupting teachers. Here, I use the term student characteristics to include all characteristics of students, including ethnicity/race, content knowledge, personality, and culture among others. Student characteristics are typically operationalized through measures of prior knowledge, gender, and race/ethnicity. Past research has generally found a strong relationship between student characteristics and observation scores, but this relationship generally is strongest at the between-teacher and between-school levels, making it difficult to interpret (e.g., Drake et al., 2019; Steinberg & Sartain, 2021). This difficulty arises because there are several pathways which could lead to an empirical relationship between students and observation scores (Milanowski, 2017; White, in press). Students could actively and directly contribute to the quality of enacted teaching (e.g., through actively participating or refusing to participate), teachers could choose different teaching approaches based on their understanding of their students, or the systematic sorting of teachers and students across and within schools could lead teachers who teach in certain ways to happen to teach specific types of students (i.e., the relationship is exogenous; Donaldson et al., 2017; Goldhaber et al., 2015). The distinction between these possibilities has important implications for both our understanding of teaching and many interpretations of observation scores (e.g., as a measure of teacher quality; c.f., Milanowski, 2017).

Teacher's Knowledge, Skills, and Beliefs

Teachers also contribute to observation scores. Different teachers have different characteristics, such as knowledge, skills, and beliefs, enabling some teachers to enact higher quality teaching than others, given the same students and instructional context (Bell et al., 2012). Much research on observation systems has focused on understanding the contribution of teachers to observation scores, especially research that seeks to interpret observation scores as a measure of teacher quality (e.g., Kane et al., 2012). Research exploring the contribution of teachers to observation scores typically models a teacher or classroom facet within a generalizability theory framework (e.g., Praetorius et al., 2014). This facet, though, does not necessarily capture the teacher's contribution to observation scores but only variation in scores across teachers. In the case of an unmodeled school level (White, 2017) or student sorting (Milanowski, 2017), the teacher facet may include effects related to schools or students.

Rater Error

While raters do not contribute directly to the quality of enacted instruction, they contribute to observation scores. As I have argued, observation systems seek to measure a specific understanding of teaching quality and scores are interpreted in light of that specific understanding. Rater error is a major threat to this interpretation because it adds construct irrelevant variation to scores that make scores poorer representations of the intended conceptualization of teaching quality. Past work exploring rater error has only looked at rater agreement (i.e., if raters agree on a score), rather than rater accuracy (i.e., if raters assign the correct score; Myford & Wolfe, 2003). Recent work has shown that raters frequently agree on the wrong score (White & Ronfeldt, 2024). In this case, raters' full contribution to observation scores can be hidden when not examining rater accuracy, which may have led past studies to underestimate raters' contribution to observation scores.

Method

The analyses presented in this paper seek to identify the extent to which observation scores capture the intended construct (i.e., construct validity; RQ1) and the extent to which the factors of students, teachers, and teaching context contribute to scores (RQ2). Deviations of scores from master scores, which are interpreted as capturing the intended construct, are used to address RQ1. RQ2 is addressed by examining the extent to which variation in scores is related to variables associated with each of the factors. The study that provided data is first described and then specific analyses are discussed.

Understanding Teacher Quality (UTQ)

This paper is a secondary analysis of anonymized data from the Understanding Teaching Quality project (UTQ; Casabianca et al., 2015), a research project that tested the quality of observation systems for evaluation purposes. The UTQ project conducted live and video observations of mathematics and English language arts teachers in grades 6-8 in three large, southeastern US school systems from 2009-2011. The project had a sample of 458 volunteer teachers. This paper uses data from the 228 teachers who taught English. Each teacher was observed and videotaped teaching one lesson on four separate days across 2 classrooms (908 lessons).[2] Each video was scored using the Classroom Assessment Scoring System (CLASS; Pianta et al., 2010), the Framework for Teaching (FFT; Danielson, 2000), and one other protocol not examined here due to space limitations.

CLASS and FFT are widespread observation systems that include an established rubric and scoring procedures. The CLASS rubric included 11 items and the FFT rubric included 11 items (only domains 2 and 3 were used, along with the Demonstrating Knowledge of Content and Pedagogy element). Due to space restraints and for simplicity, I present only results for the average score and refer readers to prior work to understand the items (Danielson, 2000; Pianta et al., 2010). Appendix A replicates results for each item, showing largely similar results across items, justifying this decision.

The UTQ study recruited 12 former teachers as raters. These raters engaged in standard training and certification for both CLASS and FFT (and one other protocol). Raters underwent calibration exercises every week (once every three weeks for each protocol), which involved scoring a common video and discussing scores on that video, in order to maintain scoring reliability across time. Calibration was low-stakes and provided raters with consistent and clear feedback on their scoring across the scoring process. Raters were randomly assigned to videos to score. Double scoring was completed for one of the four videos submitted by each teacher (25% rate) by a randomly assigned rater. CLASS videos were scored in 15-minute segments (usually 3 per lesson) while FFT videos were scored as 30-minute segments (usually 1 segment per lesson). Data is aggregated to the lesson-level (across segments), both to align data across CLASS and FFT and under the premise that the equal-interval segmentation is a convenience to reliably estimate average teaching quality in a lesson (c.f., White, Luoto, et al., 2022).

The teaching context proxy measures were the observed lessons' semester (Fall or Spring); the classrooms' grade level; the time of day, the month, and the day of the week of the observed lesson; and indicators for whether the lesson focused on grammar, literature, reading comprehension, and writing. Proxy measures for students were (all measured as a percentage at the classroom level): English language learners, students in special education, students with gifted status, students on free-reduced price lunch, Asian students, African American students, White students, and multi-racial students. The classroom-averaged prior test score on the district standardized test also served as a proxy measure. To avoid over-fitting, these variables were reduced using principal components analysis (Greco et al., 2019), keeping the first two components, which explained 39% and 15% of the variation, respectively.

The teacher proxy measures were the teacher's highest educational degree, teacher certification status for English and middle school, the teachers' years of experience, a measure of teacher content knowledge

from the Teacher Knowledge Assessment System (TKAS; Phelps et al., 2014), teacher value-added scores from the current year, and teacher value-added scores from the previous year (see Lockwood & McCaffrey, 2012 for specifications on value-added models). Again, to avoid over-fitting, the teacher variables were reduced to two dimensions using principal components analysis (the two components explained 24% and 18% of the variation, respectively).

UTQ Calibration Data

To explore rater accuracy, I incorporate UTQ calibration data into the analyses. The UTQ calibration data includes scores from every rater scoring each of the 18 lessons used in weekly calibration. Otherwise, the calibration data is structured like the main data set. The UTQ calibration data includes master scores. Master scores are scores made through consensus by teams of master raters, who are carefully trained and highly experienced raters. Rubric developers certify master raters to be able to work in teams to identify the score that would have been assigned if the rubric were applied correctly. This makes master scores a useful criterion measure for representing the rubric-intended teaching quality construct. In treating master scores as a criterion score, I make no assumption that they are "true" in any abstract sense, but simply that they are useful estimates of the intended score given that they are assigned by teams of highly trained and experienced raters that take extensive amounts of time to ensure that the scores are correct.

A limitation of the calibration data is that raters knew they were scoring the calibration videos and that they would discuss these videos as a group, which raises the threat that they scored these videos differently than they scored typical videos, though the low stakes nature of UTQ calibration makes this less likely. Further, should raters have scored differently, it is likely that raters more carefully scored in line with the rubric guidelines, such that the calibration data should contain less rater errors than operational data.

Analyses

The analyses are based on mixed-effects regression models. Separate models are run for the full and calibration data sets. The calibration models use the rater-assigned score minus the master score as an outcome measure (White & Ronfeldt, 2024). This removes all construct-related variation in rater-assigned scores (under the assumption that the master scores capture all construct-related variation). Thus, these regressions model only rater error. The regression models for the main data set use the rater-assigned score as the outcome measure, as no master scores exist. The calibration data models, then, provide a detailed analysis of rater error while the main data models provide a decomposition of how scores vary across the measurement facets, teaching context, students, and teacher while modelling rater disagreement. To understand both the variation of scores and the role of rater error, the two models must be combined by replacing estimates of rater disagreement from the main data models with estimates of rater error from the calibration data models, which requires assuming that raters score both calibration and main data in similar ways. Given that raters knew they were scoring calibration videos, this assumption may not be accurate, but deviations from the assumption should lead to biases that reduce the perceived impact of rater error, as raters would likely score calibration videos especially carefully since they knew that they would be judged.

Across both sets of models, the random effects are based on the data's structure. In the calibration data, which includes one lesson per teacher, the random effects include a crossed rater and lesson effect. In the CLASS calibration models, the segment level codes allowed me to further include the rater-by-lesson and segment (nested in lesson) facets. The segment level was only included in this calibration data because it allowed the estimation of the rater-by-lesson facet, which provides information about biases specific raters might have for specific lessons, something that highly informs interpretations of scores. The calibration models do not include fixed effects since only rater error is being modelled. In the main data set, the random effects included school, teacher (nested in schools), classroom (nested in teachers), lesson (nested in classrooms), and rater (crossed) effects. The fixed effects include the proxy measures described above. All

facets were modeled as normally distributed random effects using Restricted Maximum Likelihood (REML) with the package lme4 (Bates et al., 2015) in R (R Core Team, 2020).

Results from the calibration and full data set models were combined for interpretation. The calibration models model rater accuracy while the main models include variance components that capture rater disagreement (namely the rater and residual facets3). Since the calibration models better model rater accuracy (the more relevant construct for our purposes), the rater agreement variance in the full data models were replaced by the rater accuracy variance components in the calibration model for interpretation. This is valid under the assumption that the quality of rater scoring is equivalent in the full data and the 16 calibration videos. As noted above, the fact that calibration scores were publicly discussed would suggest that, if this assumption is false, the quality of scoring should be higher in the calibration data, introducing conservative biases that would reduce the apparent size of rater error.

The results focus on two areas. Consistent with typical generalizability theory analyses (Brennan, 2001), I interpret the relative sizes of the random effects, which is the variation in scores associated with the data structure. Second, consistent with hierarchical linear modelling traditions (Raudenbush & Bryk, 2001), I interpret the amount of variance explained in each random effect by the proxy measures. Variance explained by the proxy measures is assumed to be variance in scores associated with the given factor.

In order to facilitate interpretations of the variance explained by the proxy measures (RQ2), the full data models were built up across five sets of models. The base model includes no fixed effects. The teaching context model adds the teaching context proxy variables to the base model. The student characteristics model adds the student proxy variables to the teaching context model. The teacher model adds the teacher quality proxy variables to the teaching context model. The full model includes all proxy variables. Building up the models in this way allows us to distinguish between variation in observation scores associated with only the student proxy measures, variation associated with only the teacher proxy measures, and variation associated with both the teacher and study proxy measures (e.g., the variance associated with only the teacher proxy measures is the variance explained in the full model minus the variance explained in the student characteristics model). This is necessary for interpretation because the student and teacher proxy measures are often highly correlated.

Results

Variance Decomposition of Scores

This section compares the rater-related facets in the calibration data model with the base model in the full data set. This allows for a comparison of the two models and consideration of the impact of combining models. Table 1 shows the results of all models. The columns "variance associated with each facet" show the amount of variance associated with the indicated facet and observation system for the full and calibration models. These are shown in amount of variance to facilitate comparison across the full data and calibration data models since percentages are non-comparable across these models. Table 2 shows the percentage of each facet associated with the proxy measures (and so attributable to each of the three factors).

Both models estimate the rater facet, also called rater leniency. In the calibration models, this captures average rater deviations from the master score while, in the full models, this captures average deviations from the scores assigned by other raters (after accounting for other facets). Rater facet variances in the full data, then, are likely to be inflated since raters do not precisely score lessons with the same average teaching quality as each other. That is, a rater who was randomly assigned to score lessons with lower scores on average teaching quality will be viewed as harsher than they are in the full models. This could explain the higher variance attributable to the rater facet in the full versus the calibration models, though this could also

arise from other sources, such as raters scoring calibration lessons differently than regular lessons. I use the calibration data estimate for the rater facet in later analyses since this captures rater accuracy and is likely more precisely estimated. Note that this choice potentially introduces conservative biases, making rater error seem less problematic than it is.

Table 1. Variance decomposition of scores

Protocol	Facet	Variance associated with each facet		
		Full data	Calibration data	Final "Model"
CLASS	School	0.032		0.032
CLASS	Teacher	0.045		0.045
CLASS	Classroom	0.007		0.007
CLASS	Lesson	0.041		0.041
CLASS	Rater	0.061	0.015	0.015
CLASS	Residual	0.158	0.022	0.022
CLASS	Rater-by-Lesson		0.081	0.081
CLASS	Lesson and Segment Rater Error Facets		0.042	0.042
FFT	School	0.011		0.011
FFT	Teacher	0.018		0.018
FFT	Classroom	0.002		0.002
FFT	Lesson	0.008		0.008
FFT	Rater	0.012	0.006	0.006
FFT	Residual	0.061	0.040	0.040
FFT	Lesson Rater Error Facets		0.062	0.062

Note: FFT data does not have Rater-by-Lesson (Rater Bias) since this facet was not separable from the residual due to FFT being scored at the lesson level. Blank cells indicate where specific facets could not be estimated within the given modelling framework. Full data models use the dependent variable of the rater-assigned score while calibration data models use the dependent variable of the rater-assigned score minus the master score. Rater-associated facets in the calibration model capture rater error and these facets capture rater disagreement in the full data model.

The residual facet of the full data set captures more systematic and complex forms of rater disagreement since the other facets capture all variability associated with the lesson, classroom, teacher, and school. That is, all non-rater sources of variation. We can compare this residual facet to the rater-by-lesson facet4 and rater residual error facet from the calibration data. For CLASS, the full model estimates higher levels of these complex rater errors than the calibration data while the opposite is true for FFT. I take the calibration data as the better estimate, as it uses the master scores to precisely examine rater error rather than examining rater disagreement.

Note also from Table 1 that the lesson and segment rater error facets are relatively large in size in the calibration models. These facets capture errors that are made by all study raters, which makes them invisible when examining only rater agreement, underlying the importance of studying rater accuracy when examining observation scores. The relatively large size of this facet implies that raters likely agreed on the incorrect score, which is a score other than the score master raters would assign. This would lead to inflated estimates

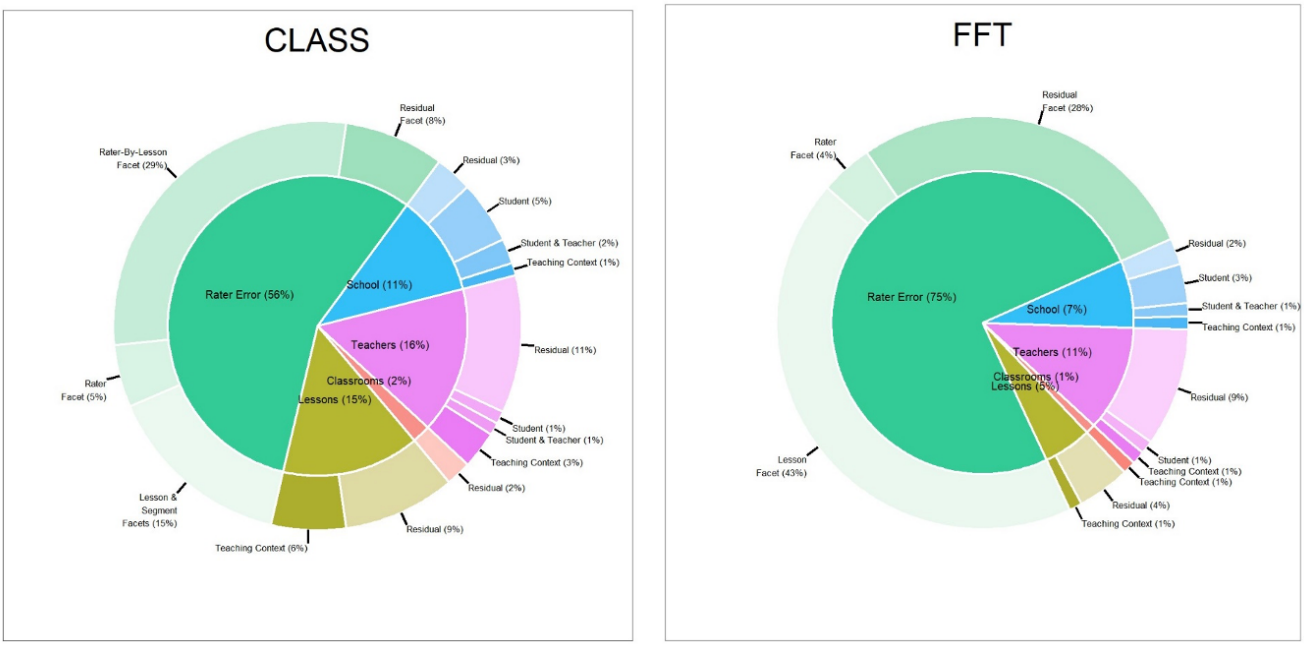
of the variation in scores across lessons, classrooms, teachers, and potentially schools (with lower levels more likely to be inflated). This would again lead our results to under-estimate the threat of rater error while potentially over-estimating the variation in scores attributable to lessons, classroom, teachers, and schools.

What's in a Score?

The models from the full data and the calibration data are combined to form the final model (see Table 1). The final model is not a regression model per se but combines results from the other two regression models to characterize the observed variation in scores across lessons, classrooms, teachers, and schools while also capturing rater accuracy (i.e., facets associated with observed instruction are taken from the full data model and facets associated with rater error are taken from the calibration data model). Figure 2 represents the final model in Tables 1 and 2 as a pie chart, providing a clear visual representation of the relative size of each facet and factor (see the inner circle). For each component, the variance explained by each set of proxy measures is also represented in Figure 2 (see the outer circle). Since it is common to average observation scores across lessons and raters, I present also Figure 3 that shows results after aggregating scores across four lessons scored by two raters (each rater scoring different lessons).

The large rater error stands out in Figure 2. Further, much of this rater error is systematic (i.e., consistent across more than one scoring occasion). That is, the rater facet captures rater leniency, the lesson and segment facets capture common errors across all raters, and the rater-by-lesson facet captures consistent errors made by individual raters across all lesson segments. Only the residual facet under rater error is non-systematic (i.e., random error according to the model). Each of these systematic errors could lead scores to be biased against specific classrooms, teachers, or schools. For example, past research has suggested that raters may give scores that are biased against male teachers and classrooms that have high proportions of minoritized students (e.g., Campbell & Ronfeldt, 2018). These sorts of rater biases would create variance in the rater-by-lesson facet (or the lesson facet of the calibration model if they are made by all raters).

Figure 2. Final model results for a single observation scored by a single rater



Figures 2 and 3 also show that minimal variation in scores is associated with only the teacher proxy measures (see also the "Only Teacher" column of Table 2; i.e., can be clearly attributable to teachers). There is, though, a reasonable amount of variation in scores associated with both teachers and students, most of

which is at the between-school level. This shows the difficulty of empirically distinguishing between teachers and students impacts on scores.

Figure 3. Final model results for scores averaged across 4 lessons and 2 raters

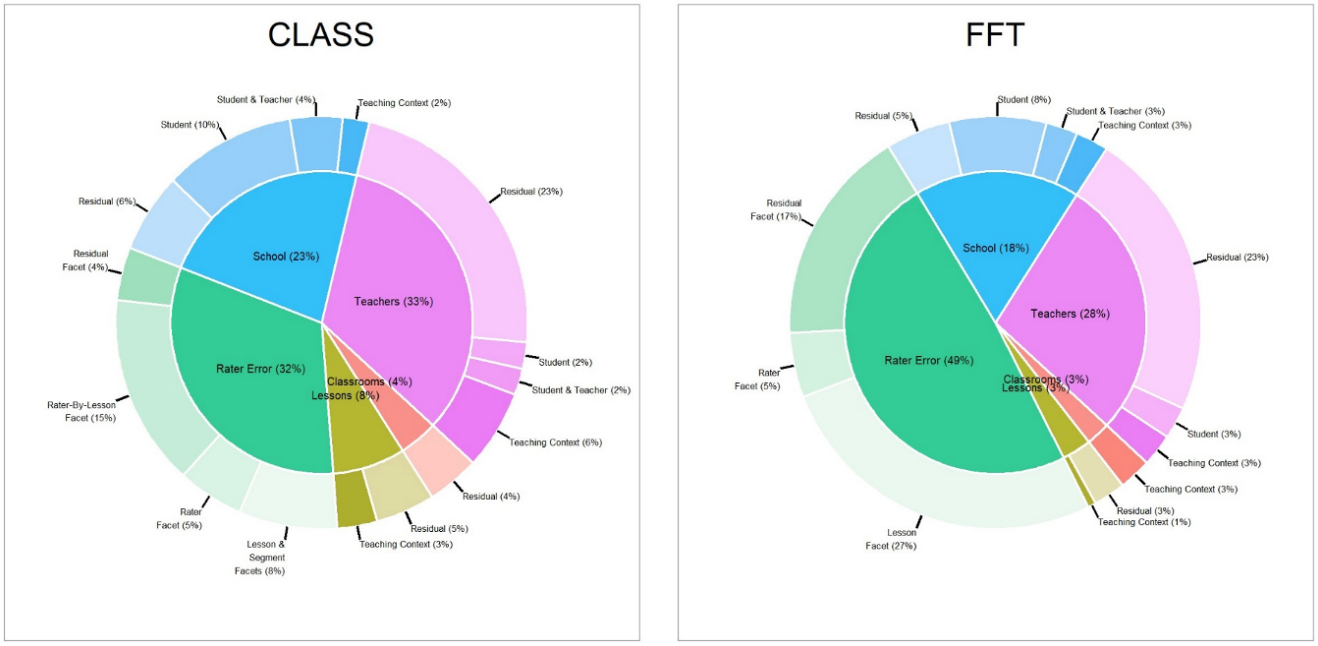


Table 2. Variance attributable to each factor and facet

		Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error				
Protocol	**Facet**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
CLASS	School	13 %	41 %	0 %	19 %	28 %
CLASS	Teacher	18 %	5 %	2 %	2 %	73 %
CLASS	Classroom	0 %	0 %	14 %	0 %	86 %
CLASS	Lesson	39 %	2 %	0 %	0 %	59 %
CLASS	Rater	11 %	0 %	0 %	0 %	89 %
FFT	School	10 %	40 %	0 %	20 %	30 %
FFT	Teacher	11 %	11 %	0 %	5 %	74 %
FFT	Classroom	100 %	0 %	0 %	0 %	0 %
FFT	Lesson	25 %	0 %	0 %	0 %	75 %
FFT	Rater	0 %	0 %	0 %	0 %	100 %

Note: FFT data does not have Rater-by-Lesson (Rater Bias) since this facet was not separable from the residual due to FFT being scored at the lesson level. Full data models use the dependent variable of the rater-assigned score while calibration data models use the dependent variable of the rater-assigned score minus the master score. Rater-associated facets in the calibration model capture rater error and these facets capture rater disagreement in the full data model.

Figures 2 and 3 show that most between-school variability in scores is associated with student characteristics, suggesting the conflation of between-school differences in scores and schools' student composition. Note that relatively little of the between-teacher variation is associated with student composition, suggesting that there is a fundamental difference between the variation observed within-schools and across teachers and between-schools. This suggests that researchers should separately consider these two sources of variation.

Last, Figure 2 shows that the teaching context explains some variation in observed scores across classrooms, teachers, and schools. This could indicate that variance decompositions are attributing data to the wrong level, since lesson level context variables are explaining variation in scores between classrooms, teachers, and schools. This would occur if an inadequate number of lessons within classrooms were sampled such that variation in scores cannot be attributed to the correct level. Given the sampling of only two lessons per classroom and evidence that 15-30 logs are needed for reliable estimates of context features (Rowan et al., 2004), this would be unsurprising and variation in classrooms, teachers, and schools associated with the teaching context should be lesson-level variation. However, this might also reflect systematic differences in the enacted curriculum (e.g., some classrooms provide more writing instruction than others).

Discussion

Uses of observation systems rest on observation scores being interpreted as capturing important aspects of instructional practice. This paper sought to decompose observation scores across both typical measurement facets and the factors of the teaching context, students, teachers, and raters in order to better understand the quality of scores. The analyses presented here seek to consider the extent to which scores from observation systems capture the intended construct (RQ1) and the factors that contribute to scores (RQ2). These analyses are novel due to the incorporation of calibration data, which allows for a consideration of rater accuracy rather than rater disagreement.

RQ 1

In the observation systems, rubric developers intentionally build rubrics to operationalize specific teaching quality constructs and ensure raters can accurately and consistently score these intended constructs. That is, raters should be assigning scores that are equivalent to master scores. Then, the examination of rater accuracy, operationalized as the difference from master scores, is important when considering what contributes to observation scores. This focus on rater accuracy (RQ1) has highlighted that (1) much rater error remains invisible when examining rater agreement and (2) most rater error is systematic (i.e., unlikely to be purely random).

The rater error lesson and segment facets capture variation associated with all raters assigning scores to a lesson/segment that deviate in the same way from master scores. They were about as large as the teacher facet in CLASS and over three times as large as the teacher facet in FFT. These facets are effectively invisible when modelling rater agreement, making them invisible in nearly all prior research. This "invisible" rater error leads to positive biases in estimates of score reliability. Since nearly all prior studies have model rater agreement, one implication of this study is that prior estimates of score reliability could be positively biased. Replication of the analyses conducted here in other data sets and with other observation systems is important to consider the extent of possible bias.

The systematic nature of rater errors identified in analyses has important implications for potential biases in validity coefficients. The rater error facets of lesson and rater-by-lesson are potentially correlated with classroom or teacher characteristics. Using the fact that a correlation is the square root of the shared variance between two measures, we can estimate the maximum possible size of bias that would be introduced if these

rater error facets were correlated with variables used to estimate validity coefficients (e.g., value-added test scores). These facets capture 43% of the variation in both CLASS and FFT scores, suggesting a maximum possible bias of 0.66 (i.e., $\sqrt{0.43} \approx 0.66$). Aggregating across multiple lessons and raters would reduce this potential bias, but it remains as high as 0.48 in CLASS and 0.52 in FFT. This implies that a true validity coefficient of 0.5 for CLASS scores could be estimated to be as low as 0.02 or as high as 0.98 solely due to biases related to rater error. While this is the maximum level of bias and actual bias is likely to be substantially lower, the finding still casts doubt on claims of the concurrent or predictive validity of observation scores using the UTQ data. Consider how such biases could manifest in practice. For example, a rater could notice and be influenced by the socio-demographic features of classrooms when scoring. At the same time, value-added scores, a common measure of concurrent/predictive validity, are associated socio-demographic features of classrooms (Lockwood & McCaffrey, 2012), creating a positive association between rater-introduced error in scores and value-added measures. Unfortunately, information on teacher quality was not available for the calibration data so I cannot empirically test for such biases.

The analyses presented related to RQ1, then, provide quite a negative view in the UTQ data. Rater error is easily large enough to cause severe problems in interpreting CLASS and FFT scores as representing the intended construct in the UTQ data. Note that Appendix A shows that this is also the case at the individual item level, across all items. It is important to note that CLASS and FFT are among the most studied and, arguably, most highly developed observation systems. Since this finding replicates quite strongly across these two rubrics, it should raise serious concerns about other, similar observation systems. This is confirmed by the high levels of rater disagreement in most large-scale studies (Gitomer, 2024; e.g., Jentsch et al., 2022; Liu et al., 2019; OECD, 2020). However, further research is needed to explore the generalizability of findings to other rubrics.

If these results replicate to other data sets and observation systems, it has serious implications for using observation systems. Observation scores are often interpreted as representing the intended understanding of teaching quality and, through this interpretation, are used to make judgements about teaching and teachers and used to develop theoretical understandings of teaching quality. The high levels of rater error in the UTQ data call into question whether scores can be interpreted as reflecting the intended teaching quality constructs. Rather, UTQ scores are more appropriately interpreted as capturing raters' idiosyncratic understandings of teaching quality as filtered through the application of specific rubrics. Conceptual replication of this result across data and observation systems is needed to understand if this, more limited, interpretation of observation scores is applicable to other observation systems and data sets. Given that a key benefit of observation systems is their purported capacity to provide consistent scores across settings and studies (Klette, 2020), this replication work is highly important.

It is important to consider the limitations of the analyses related to RQ1. Namely, analyses showed that existing studies would have positively biased estimates of score reliability if the sorts of rater error found in UTQ existed more broadly and that the rater error existing in the UTQ data could lead to large biases in estimates of score validity. The analyses did not show that prior studies have biased estimates of reliability or that biases in validity correlations existed in UTQ. At best, then, the analyses related to RQ1 raise a number of cautions that should be heeded by other researchers and users of observation systems, especially when one seeks to interpret observation scores as representing the rubric-defined teaching quality constructs. For other uses of scores that are not dependent on this interpretation, rater "error" may, in fact, be positive. For example, errors introduced by administrators make scores more stable across time, which may be positive in a teacher evaluation context (Liu et al., 2019) even if it makes scores less representative of the intended teaching quality constructs. Then, while the results suggest a need for carefully considering the nature and amount of rater error and developing approaches to reduce rater error, work on rater error should

be very conscious of different interpretations and uses of scores and the extent to which different types of rater error are more or less problematic for a given interpretation/use.

These points lead to a set of practical takeaways. First, observation scores should probably not be interpreted as capturing an objective level of teaching quality but should be understood as representing raters' subjective interpretations of instruction, as provided through the observation rubric. Importantly, this takeaway is consistent with other summaries of the field (Gitomer, 2024). This severely limits the usefulness of observation scores. For example, uses of observation scores related to teacher observation are questionable if scores are interpreted as representing a principal's idiosyncratic view of teachers. However, uses of observation scores related to providing teachers feedback may still be defensible, especially if the observer has expertise that is deemed relevant to providing feedback.

In order to overcome this limitation, observation systems would likely need to (1) make rubrics easier to score, (2) improve rater training, and (3) create other systems to reduce rater error. Even if these changes could be made, users of observation systems would need to show that raters can accurately and consistently assign scores equivalent to master scores to defend the interpretation of scores as capturing the intended teaching quality construct. This could be shown by having raters blindly score a set of videos that have master scores and demonstrating that rater error is a minor contributor to scores using analyses similar to those demonstrated here. Such analyses should systematically test for biases that might undermine intended conclusions. For example, an intervention study should compare rater scores to master scores for both intervention and control cases to ensure no biases would impact conclusions about intervention effectiveness. The weak standards of rater certification imply that rater certification cannot serve this function (White, 2018).

RQ 2

The analyses revealed important information about what factors contribute to observation scores, beyond rater error. The teaching context explained 20-40% of the variation in scores across the lesson facet, suggesting that a significant component of day-to-day variation in scores is systematically related to the teaching context. The teaching context proxy measures (i.e., grade level, time of day, day of the week, the month of the lesson, content domain) were arguably quite weak in this study. That is, collecting more fine-grained information about the teaching context (e.g., lesson content; learning goals; aspects of the curriculum such as location in the unit) would likely lead models to explain more of the lesson-level variation in scores. This could have important implications for the sampling for lessons and other aspects of conducting observations. This points to a second takeaway for users of observation systems. Systematically stratifying the sample of lessons based on important characteristics of the lesson context, such as learning goals (e.g., introducing new content, practicing skills) and/or content areas, could improve the quality of score estimates for individual classrooms or teachers. This is because there simply are not enough lessons sampled for each teacher to rely on random sampling to generate equivalent lesson samples for each teacher. This is less important when not trying to generate estimates for individual classrooms or teachers but could still lead to somewhat better score reliability.

The student (i.e., student demographic characteristics and prior test scores) and teacher proxy measures (i.e., typical teacher quality measures) were arguably much stronger. These proxy measures explained almost all the between-school variation in scores. While both the student and teacher proxy measures explained between-school variation, the student proxy measures explained more of this variation. In fact, the teacher proxy measures did not uniquely explain between-school differences in scores. This raises important questions for how to interpret between-school differences in observation scores since these are largely collinear with differences in demographic characteristics of schools. This could imply that differences in student characteristics are driving differences in average teaching quality at a school, though many other explanations, such as rater bias, could also explain this finding. In this case, it would be worth reflecting on

whether the observation system is equally appropriate in different school contexts. Alternatively, it could be that the sorting of students and teachers leads to a high correlation between student characteristics and school quality (e.g., Donaldson et al., 2017; Goldhaber et al., 2015). It would seem, though, that between-school differences in observation scores are not highly driven by differences in teacher quality, at least as indexed by typical teacher quality metrics.

The teacher facet, on the other hand, consists largely of residual variation, though the student and teaching context proxy measures do explain some of this variation. The different patterns of associations with proxy measures across the teacher and school facet would imply that the two facets are capturing very different dynamics. This residual variation in the teacher facet could result from inadequate proxy measures. For example, if observation scores captured aspects of teaching quality that were unique from other variables (and not related to student characteristics), we would expect to find the high levels of residual variance at the teacher level that we observed. This residual variation, then, could be interpreted as justifying claims that observation systems capture unique aspects of teaching quality (e.g., Kane et al., 2012), though further research should consider this. These points imply another important takeaway for users of observation systems: Between-school differences in observation scores should be separately estimated from between-teacher and within-school differences. These between-school differences should likely be removed from estimates of teacher quality, at least until further research can provide assurances that such effects are driven by differences in teacher quality.

The reader will note a lack of comparison of results between CLASS and FFT. This is because while the details of some results were different across the instruments (e.g., the lesson facet of the calibration data in FFT was twice as large as that for CLASS), the overall picture and key results were the same across the rubrics. Further, the results provide no basis for determining what might have caused the differences. It could be differences in the rubric scaling (i.e., CLASS has a 7-point scale and FFT a 4-point scale); differences in scoring procedures (i.e., CLASS scores 15-minute segments and FFT scored 30-minute segments); differences in the focus of rubrics (i.e., CLASS focuses more on the socio-emotional climate than FFT); or any of a number of other differences. While it is tempting to make cross-rubric comparisons, the joint analysis should be considered from a replication perspective.

Considering Observation Systems More Broadly

If the results found here generalize to other observation systems and data sources, it would raise serious questions about the extent to which observation scores capture the intended teaching quality constructs. As noted above, though, not all uses of observation systems rely on this interpretation of scores. Conceptually, observation systems may be most useful when used to provide a common framework and language for supporting teachers in studying and discussing their practice (Charalambous & Praetorius, 2020; Gitomer, 2024; White & Maher, 2024). Empirically, observation rubric's focus on observable aspects of teaching and the common language has been shown to support higher quality feedback and instructional improvement that focuses on improving teaching (e.g., Koedel et al, 2019; Kraft & Gilmour, 2017; Taylor & Tyler, 2012). These uses are potentially independent of whether observation scores represent the intended teaching quality constructs. This paper should not be understood as a critique of observation systems, then, but as an exploration of what factors contribute to observation scores and caution about interpretations of scores. This leads to another takeaway for users of observation systems: Uses of observation systems that do not emphasize the accuracy of scores are more justifiable, such as using observation systems to provide feedback and instructional improvement within coaching regimes (see also Gitomer, 2024). Observation scores are simply not reliable enough measures of the rubric-intended understanding of teaching quality.

Generalizability of Data Source

A word on the nature of the UTQ data and the generalizability of the findings is important here. Analyses of calibration data, similar to those presented here, have been conducted by the author (White & Ronfeldt, 2024) using both data from the measures of effective teaching project (MET; Kane et al., 2012) and UTQ. In these analyses, rater error was a significant problem in both data sets, but it was more problematic in the UTQ data. Then, the results presented here may over-estimate the threat of rater error, as the UTQ data contains more rater error than the MET study. Given this caveat, though, it is important to note that rater error was still the largest source of variation in scores in the MET data and there was a significant amount of shared systematic rater error (across the 100+ MET raters). So, while results may not have been as extreme in the MET data, the overall patterns of findings would likely have been similar. Additionally, many decisions in the conduct of the analyses presented here made choices that could have under-estimated rater error. The results of this study, like any study, need to be replicated with data from other research studies, with data from practice settings, and with additional observation systems.

A second caveat is the age of the data. The data was collected between 2009 and 2011, making it quite dated at this point. Research suggests that shifts in instructional practice are quite small from 2011 to 2018 (White, Maher, & Rowan, 2022). The largest challenge with the age of the data set is that the observation systems may have changed. While CLASS has released a version 2.0 for pre-kindergarten to 3rd grade, the same version as used in this paper is still in active use for middle school. The FFT was updated in 2022 (see https://danielsongroup.org/the-framework-for-teaching), but this update seems to be somewhat modest as the same, slightly renamed, dimensions exist in the new version. The largest change seems to be an expansion of the rubric to include more components within each dimension. This would, all else being equal, increase the complexity of scoring, which should increase levels of rater error. Then, the age of the data does not seem to preclude drawing conclusions, though these findings must be replicated across other data sets to check their generalizability.

Conclusion

Scores from observation systems have become a common data source in exploring instruction, evaluating educational interventions and policies, and have been incorporated into many educational policies. The analysis shown here suggests that much of what is being measured by these scores, even after aggregation, may be systematic and random rater errors. These errors could be large enough to call into question interpretations of scores as capturing the intended teaching quality constructs, especially as the socio-demographic features of students or teachers could easily influence raters, leading errors to be correlated with such features of classrooms. Further, these errors are poorly understood. There is a need for studies to more carefully consider how rater error might impact conclusions and more research to explore the nature of these errors.

Declarations

Notes

1. Not all observational approaches to studying teaching would fit this definition. This paper is restricted to those that follow this definition, which reflects typical usage of the term observation system in many settings (Bell et al., 2019; Hill et al., 2012).
2. Four classrooms were observed only once due to scheduling problems.
3. Simple simulations that fix rater error to zero find approximately zero residual variance, providing strong empirical evidence regarding the claim that the residual facet captures rater error-related variation in scores. Results available from author upon request.

4. Since the rater-by-lesson facet could not be estimated in the full data models, the variation associated with this facet is absorbed by the residual in the full data models.

Funding
Project was funded by Nordforsk, grant number: 87663 and Norges forskingsråd, grant number: 300791.

Acknowledgements
The author would like to thank the anonymous reviewers for their useful feedback.

Received: 5/24/2024. **Accepted:** 4/21/2025. **Published:** 4/29/2025.

Citation: White, M. (2025). What is in a Score? Exploring the contribution of raters, students, teachers, and the teaching context to scores from observation systems. *Practical Assessment, Research, & Evaluation*, 30(3). Available online: https://doi.org/10.7275/pare.2106

Corresponding Author: Mark White, Postboks 1099, Blindern, 0317 Oslo, Norway. Email: mark.white@ils.uio.no.

References

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, 73. https://doi.org/10.1016/j.econedurev.2019.101919

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1). https://doi.org/10.18637/jss.v067.i01

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, 30(1), 1–27. DOI: ggf5gq

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(2–3), 62–87. https://doi.org/10.1080/10627197.2012.715014

Blazar, D. (2017). Validating Teacher Effects on Students' Attitudes and Behaviors: Evidence from Random Assignment of Teachers to Students. *Education Finance and Policy*, 13(3), 281–309. https://doi.org/10.1162/edfp_a_00251

Brennan, R. L. (2001). *Generalizability Theory*. Springer New York. https://doi.org/gwqz

Campbell, S. L., & Ronfeldt, M. (2018). Observational Evaluation of Teachers: Measuring More Than We Bargained for? American Educational Research Journal. https://doi.org/gd32fh

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, 75(2), 311–337.

Charalambous, C. Y., & Praetorius, A.-K. (2020). Creating a forum for researching teaching and its quality more synergistically. *Studies in Educational Evaluation*, 67, 8. https://doi.org/10/gwsf

Cherng, H.-Y. S., Halpin, P. F., & Rodriguez, L. A. (2022). Teaching Bias? Relations between Teaching Quality and Classroom Demographic Composition. American Journal of Education, 128(2), 171–201. https://doi.org/10.1086/717676

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, Instruction, and Research. *Educational Evaluation and Policy Analysis*, 25(2), 119–142. https://doi.org/b88jtw

Cowan, J., Goldhaber, D., & Theobald, R. (2022). Performance Evaluations as a Measure of Teacher Effectiveness When Implementation Differs: Accounting for Variation across Classrooms, Schools, and

Districts. Journal of Research on Educational Effectiveness, 15(3), 510–531. https://doi.org/10.1080/19345747.2021.2018747

Curby, T. W., Stuhlman, M. W., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., Hamre, B. K., & Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade. *The Elementary School Journal*, 112(1), 16–37.

Danielson, C. (2000). *Teacher Evaluation to Enhance Professional Practice*. Association for Supervision & Curriculum Development.

Donaldson, M. L., LeChasseur, K., & Mayer, A. (2017). Tracking instructional quality across secondary mathematics and English Language Arts classes. *Journal of Educational Change*, 18(2), 183–207. https://doi.org/10.1007/s10833-015-9269-x

Drake, S., Auletto, A., & Cowen, J. M. (2019). Grading Teachers: Race and Gender Differences in Low Evaluation Ratings and Teacher Employment Outcomes. *American Educational Research Journal*, 56(5), 1800–1833. https://doi.org/10.3102/0002831219835776

Gitomer, D. (2024, August 23-25). *The Productive Use of Disagreement in Observational Ratings* [Conference presentation]. EARLI Sig 18 conference, Nicosia, Cyprus.

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven Playing Field? Assessing the Teacher Quality Gap Between Advantaged and Disadvantaged Students. *Educational Researcher*, 44(5), 293–307. https://doi.org/10.3102/0013189X15592622

Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2019). On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. *Social Indicators Research*, 141(1), 61–94. https://doi.org/ghw7hb

Grossman, P., Cohen, J. J., & Brown, L. (2014). Understanding Instructional Quality in English Language Arts: Variations in PLATO Scores by Content and Context. In *Designing teacher evaluation systems: New guidance from the measures of effecting project* (pp. 303–331). Jossey-Bass.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M., & Hamagami, A. (2013). Teaching through Interactions: Testing a Developmental Framework of Teacher Effectiveness in over 4,000 Classrooms. The Elementary School Journal, 113(4), 461–487. https://doi.org/gf7rj2

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(2), 56–64. https://doi.org/hbm3

Jentsch, A., Heinrichs, H., Schlesinger, L., Kaiser, G., König, J., & Blömeke, S. (2022). 4. Multi-Group Measurement Invariance and Generalizability Analyses for an Instructional Quality Observational Instrument. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Ways of Analyzing Teaching Quality: Potentials and Pitfalls* (pp. 121–139). Scandinavian University Press. https://doi.org/10.18261/9788215045054-2021

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*. Bill & Melinda Gates Foundation, Measures of Effective Teaching Project. http://eric.ed.gov/?id=ED540960

Kelcey, B., & Carlisle, J. (2013). Learning About Teachers' Literacy Instruction From Classroom Observations. *Reading Research Quarterly*, 48(3), 301–317. https://doi.org/f43nts

Kelly, S., Bringe, R., Aucejo, E., & Fruehwirth, J. C. (2020). Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives*, 28(0), 62. https://doi.org/jbj3

Klette, K., & Blikstad-Balas, M. (2018). Observation manuals as lenses to classroom teaching: Pitfalls and possibilities. *European Educational Research Journal*, 17(1), 129–146. https://doi.org/10.1177/1474904117703228

Klette, K. (2020). Towards programmatic research when studying classroom teaching and learning. In L. Ligozat, A. Rakhkochkine, & J Almqvist (Eds.), *Thinking through Didactics in a Changing World. European Perspectives on Learning, Teaching and the Curriculum*. Routledge Education.

Koedel, C., Li, J., Springer, M. G., & Tan, L. (2019). Teacher Performance Ratings and Professional Improvement. *Journal of Research on Educational Effectiveness*, 12(1), 90–115. https://doi.org/10.1080/19345747.2018.1490471

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness. *Educational Researcher*, 46(5), 234–249.

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, 31(1), 61–95. https://doi.org/10.1007/s11092-018-09291-3

Lockwood, J. R., & McCaffrey, D. (2012, Spring). Reducing Bias in Teacher Value-Added Estimates by Accounting for Test Measurement Error. SREE.

Milanowski, A. (2017). Lower Performance Evaluation Practice Ratings for Teachers of Disadvantaged Students: Bias or Reflection of Reality? *AERA Open*, 3(1). https://doi.org/gcgnwn

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386–422.

OECD. (2020). *Global Teaching InSights: A Video Study of Teaching*. OECD Publishing. https://doi.org/10.1787/20d6f36b-en

Patrick, H., Mantzicopoulos, P., & French, B. F. (2019). The Predictive Validity of Classroom Observations: Do Teachers' Framework for Teaching Scores Predict Kindergarteners' Achievement and Motivation? *American Educational Research Journal*, 57(5), 2021–2058. https://doi.org/10.3102/0002831219891409

Phelps, G., Jones, N., Liu, S., & Kisa, Z. (2014). Examining Teacher, School, and Program Moderators in the Context of Teacher Professional Development Studies. In Society for Research on Educational Effectiveness. Society for Research on Educational Effectiveness. https://eric.ed.gov/?id=ED562735

Pianta, R. C., Hamre, B. K., & Mintz, S. L. (2010). *CLASS Upper Elementary Manual*. Teachstone.

Plank, S. B., & Condliffe, B. (2013). Pressures of the Season: An Examination of Classroom Quality and High-Stakes Accountability. American Educational Research Journal, 50(5), 1152–1182. https://doi.org/10.3102/0002831213500691

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd edition). SAGE Publications, Inc.

Rowan, B., Camburn, E., & Correnti, R. (2004). Using Teacher Logs to Measure the Enacted Curriculum: A Study of Literacy Teaching in Third-Grade Classrooms. *The Elementary School Journal*, 105(1), 75–101. https://doi.org/10.1086/428803

Steinberg, M. P., & Sartain, L. (2021). What Explains the Race Gap in Teacher Performance Ratings? Evidence From Chicago Public Schools. *Educational Evaluation and Policy Analysis*, 43(1), 60–82. https://doi.org/10.3102/0162373720970204

Taylor, E. S., & Tyler, J. H. (2012). The Effect of Evaluation on Teacher Performance. *The American Economic Review*, 102(7), 3628–3651.

van der Lans, R. M. (2018). On the "association between two things": The case of student surveys and classroom observations of teaching quality. *Educational Assessment, Evaluation and Accountability*, 30(4), he had an abortion347–366. https://doi.org/10.1007/s11092-018-9285-5

White, M. (2017). *Generalizability of Scores from Classroom Observation Instruments* [University of Michigan]. http://hdl.handle.net/2027.42/138742

White, M. C. (2018). Rater Performance Standards for Classroom Observation Instruments. *Educational Researcher*, 47(8), 492–501. https://doi.org/gd32gn

White, M., Luoto, J., Klette, K., & Blikstad-Balas, M. (2022). Bringing the conceptualization and measurement of teaching into alignment. *Studies in Educational Evaluation*, 75. https://doi.org/10.1016/j.stueduc.2022.101204

White, M., & Ronfeldt, M. (2024). Monitoring Rater Quality in Observational Systems: Issues Due to Unreliable Estimates of Rater Quality. Educational Assessment, 29(2), 124–146. https://doi.org/10.1080/10627197.2024.2354311

White, M., & Maher, B. L. (2024). How might rubric-based observations better support teacher learning and development? *Educational Research*, 66(1), 1–16. https://doi.org/mndm

White, M., Maher, B., & Rowan, B. (2022). Common Core–Related Shifts in English Language Arts Teaching From 2010 to 2018. *The Elementary School Journal*, 123(1). https://doi.org/10.1086/720732

White, M. (in press). The Impact of the Association between Student Characteristics and Classroom Observation Scores on Different Interpretations of Scores. *Journal of Research in Educational Effectiveness*

Appendix. Dimension Specific Results

			Variance associated with each facet			Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
CLASS	Positive Climate	School	0.068		0.068	13 %	57 %	1 %	18 %	11 %
CLASS	Positive Climate	Teacher	0.092		0.092	14 %	0 %	1 %	0 %	85 %
CLASS	Positive Climate	Classroom	0.022		0.022	0 %	30 %	6 %	0 %	64 %
CLASS	Positive Climate	Lesson	0.058		0.058	42 %	2 %	0 %	1 %	55 %
CLASS	Positive Climate	Rater	0.393	0.32	0.32	1 %	0 %	0 %	0 %	99 %
CLASS	Positive Climate	Residual	0.429	0.09	0.09	0 %	0 %	0 %	0 %	100 %
CLASS	Positive Climate	Rater-By-Lesson		0.2	0.2					
CLASS	Positive Climate	Lesson and Segment Rater Error Facets		0.154	0.154					
CLASS	Negative Climate	School	0.001		0.001	0 %	33 %	0 %	49 %	17 %
CLASS	Negative Climate	Teacher	0.024		0.024	0 %	2 %	6 %	3 %	90 %
CLASS	Negative Climate	Classroom	0		0	50 %	50 %	0 %	0 %	0 %
CLASS	Negative Climate	Lesson	0.016		0.016	18 %	1 %	0 %	0 %	81 %
CLASS	Negative Climate	Rater	0.059	0.061	0.061	1 %	0 %	0 %	0 %	99 %
CLASS	Negative Climate	Residual	0.124	0.04	0.04	0 %	0 %	0 %	0 %	100 %

			Variance associated with each facet			**Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error**				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
CLASS	Negative Climate	Rater-By-Lesson		0.049	0.049					
CLASS	Negative Climate	Lesson and Segment Rater Error Facets		0.034	0.034					
CLASS	Teacher Sensitivity	School	0.073		0.073	4 %	36 %	2 %	15 %	43 %
CLASS	Teacher Sensitivity	Teacher	0.089		0.089	10 %	4 %	3 %	2 %	82 %
CLASS	Teacher Sensitivity	Classroom	0.001		0.001	0 %	0 %	0 %	100 %	0 %
CLASS	Teacher Sensitivity	Lesson	0.077		0.077	24 %	4 %	0 %	0 %	72 %
CLASS	Teacher Sensitivity	Rater	0.422	0.282	0.282	1 %	0 %	0 %	0 %	98 %
CLASS	Teacher Sensitivity	Residual	0.481	0.119	0.119	0 %	0 %	0 %	0 %	100 %
CLASS	Teacher Sensitivity	Rater-By-Lesson		0.222	0.222					
CLASS	Teacher Sensitivity	Lesson and Segment Rater Error Facets		0.18	0.18					
CLASS	Regard for Student Perspectives	School	0.088		0.088	15 %	29 %	0 %	9 %	46 %
CLASS	Regard for Student Perspectives	Teacher	0.099		0.099	28 %	6 %	0 %	1 %	64 %

			Variance associated with each facet			**Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error**				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
CLASS	Regard for Student Perspectives	Classroom	0		0	50 %	0 %	50 %	0 %	0 %
CLASS	Regard for Student Perspectives	Lesson	0.136		0.136	11 %	0 %	0 %	0 %	89 %
CLASS	Regard for Student Perspectives	Rater	0.3	0.156	0.156	5 %	0 %	0 %	0 %	95 %
CLASS	Regard for Student Perspectives	Residual	0.577	0.093	0.093	0 %	1 %	0 %	0 %	99 %
CLASS	Regard for Student Perspectives	Rater-By-Lesson		0.249	0.249					
CLASS	Regard for Student Perspectives	Lesson and Segment Rater Error Facets		0.194	0.194					
CLASS	Behavior Management	School	0.011		0.011	0 %	41 %	2 %	26 %	31 %
CLASS	Behavior Management	Teacher	0.034		0.034	1 %	0 %	1 %	0 %	98 %
CLASS	Behavior Management	Classroom	0.022		0.022	3 %	8 %	3 %	1 %	85 %
CLASS	Behavior Management	Lesson	0.057		0.057	23 %	0 %	0 %	0 %	77 %
CLASS	Behavior Management	Rater	0.078	0.101	0.101	1 %	0 %	0 %	0 %	98 %
CLASS	Behavior Management	Residual	0.223	0.03	0.03	0 %	0 %	0 %	0 %	100 %
CLASS	Behavior Management	Rater-By-Lesson		0.066	0.066					

			Variance associated with each facet			**Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error**				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
CLASS	Behavior Management	Lesson and Segment Rater Error Facets		0.037	0.037					
CLASS	Productivity	School	0.014		0.014	2 %	21 %	7 %	17 %	52 %
CLASS	Productivity	Teacher	0.029		0.029	17 %	2 %	5 %	2 %	74 %
CLASS	Productivity	Classroom	0.007		0.007	26 %	0 %	0 %	0 %	74 %
CLASS	Productivity	Lesson	0.068		0.068	14 %	3 %	0 %	0 %	84 %
CLASS	Productivity	Rater	0.143	0.079	0.079	0 %	0 %	0 %	0 %	99 %
CLASS	Productivity	Residual	0.296	0.067	0.067	0 %	0 %	0 %	0 %	100 %
CLASS	Productivity	Rater-By-Lesson		0.068	0.068					
CLASS	Productivity	Lesson and Segment Rater Error Facets		0.15	0.15					
CLASS	Analysis and Problem Solving	School	0.02		0.02	17 %	57 %	0 %	10 %	16 %
CLASS	Analysis and Problem Solving	Teacher	0.029		0.029	46 %	17 %	5 %	8 %	23 %
CLASS	Analysis and Problem Solving	Classroom	0.053		0.053	17 %	0 %	0 %	0 %	83 %

			Variance associated with each facet			Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
CLASS	Analysis and Problem Solving	Lesson	0.051		0.051	67 %	5 %	0 %	0 %	28 %
CLASS	Analysis and Problem Solving	Rater	0.352	0.173	0.173	5 %	1 %	0 %	0 %	94 %
CLASS	Analysis and Problem Solving	Residual	0.438	0.084	0.084	0 %	0 %	0 %	0 %	100 %
CLASS	Analysis and Problem Solving	Rater-By-Lesson		0.255	0.255					
CLASS	Analysis and Problem Solving	Lesson and Segment Rater Error Facets		0.315	0.315					
CLASS	Engagement	School	0.034		0.034	17 %	50 %	2 %	17 %	14 %
CLASS	Engagement	Teacher	0.048		0.048	10 %	7 %	0 %	1 %	82 %
CLASS	Engagement	Classroom	0		0	50 %	0 %	0 %	0 %	50 %
CLASS	Engagement	Lesson	0.053		0.053	21 %	0 %	1 %	0 %	79 %
CLASS	Engagement	Rater	0.239	0.121	0.121	3 %	0 %	0 %	0 %	97 %
CLASS	Engagement	Residual	0.345	0.074	0.074	0 %	0 %	0 %	0 %	100 %
CLASS	Engagement	Rater-By-Lesson		0.124	0.124					
CLASS	Engagement	Lesson and Segment		0.097	0.097					

			Variance associated with each facet			**Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error**				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
		Rater Error Facets								
CLASS	Content Understanding	School	0.058		0.058	13 %	29 %	0 %	7 %	50 %
CLASS	Content Understanding	Teacher	0.055		0.055	4 %	4 %	3 %	2 %	88 %
CLASS	Content Understanding	Classroom	0.022		0.022	23 %	0 %	6 %	0 %	71 %
CLASS	Content Understanding	Lesson	0.149		0.149	27 %	2 %	0 %	0 %	72 %
CLASS	Content Understanding	Rater	0.239	0.088	0.088	3 %	0 %	0 %	0 %	97 %
CLASS	Content Understanding	Residual	0.51	0.109	0.109	0 %	0 %	0 %	0 %	100 %
CLASS	Content Understanding	Rater-By-Lesson		0.367	0.367					
CLASS	Content Understanding	Lesson and Segment Rater Error Facets		0.224	0.224					
CLASS	Instructional Learning Formats	School	0.065		0.065	9 %	36 %	0 %	11 %	44 %

			Variance associated with each facet			**Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error**				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
CLASS	Instructional Learning Formats	Teacher	0.079		0.079	22 %	6 %	1 %	3 %	69 %
CLASS	Instructional Learning Formats	Classroom	0.057		0.057	12 %	0 %	4 %	0 %	84 %
CLASS	Instructional Learning Formats	Lesson	0.022		0.022	78 %	11 %	1 %	1 %	8 %
CLASS	Instructional Learning Formats	Rater	0.244	0.116	0.116	4 %	1 %	0 %	0 %	96 %
CLASS	Instructional Learning Formats	Residual	0.564	0.104	0.104	0 %	0 %	0 %	0 %	100 %
CLASS	Instructional Learning Formats	Rater-By-Lesson		0.249	0.249					
CLASS	Instructional Learning Formats	Lesson and Segment Rater Error Facets		0.101	0.101					
CLASS	Quality of Feedback	School	0.061		0.061	18 %	22 %	0 %	12 %	48 %
CLASS	Quality of Feedback	Teacher	0.049		0.049	24 %	3 %	4 %	3 %	66 %
CLASS	Quality of Feedback	Classroom	0		0	0 %	0 %	0 %	100 %	0 %
CLASS	Quality of Feedback	Lesson	0.11		0.11	35 %	8 %	0 %	0 %	57 %
CLASS	Quality of Feedback	Rater	0.395	0.245	0.245	5 %	0 %	0 %	0 %	95 %
CLASS	Quality of Feedback	Residual	0.581	0.121	0.121	1 %	0 %	0 %	0 %	99 %

			Variance associated with each facet			**Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error**				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
CLASS	Quality of Feedback	Rater-By-Lesson		0.258	0.258					
CLASS	Quality of Feedback	Lesson and Segment Rater Error Facets		0.145	0.145					
FFT	Establishing a culture for learning	School	0.021		0.021	21 %	45 %	0 %	12 %	21 %
FFT	CL	Teacher	0.046		0.046	11 %	16 %	2 %	1 %	70 %
FFT	Establishing a culture for learning	Classroom	0		0					
FFT	Establishing a culture for learning	Lesson	0.015		0.015	74 %	8 %	0 %	0 %	18 %
FFT	Establishing a culture for learning	Rater	0.027	0.012	0.012	5 %	0 %	0 %	0 %	95 %
FFT	Establishing a culture for learning	Residual	0.227	0.189	0.189	1 %	0 %	0 %	0 %	99 %
FFT	Establishing a culture for learning	Lesson Rater Error Facets		0.221	0.221					
FFT	Creating an environment of respect and rapport	School	0.005		0.005	8 %	39 %	8 %	38 %	7 %

			Variance associated with each facet			Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
FFT	Creating an environment of respect and rapport	Teacher	0.029		0.029	1 %	6 %	0 %	0 %	93 %
FFT	Creating an environment of respect and rapport	Classroom	0.002		0.002	48 %	0 %	8 %	0 %	44 %
FFT	Creating an environment of respect and rapport	Lesson	0.013		0.013	29 %	4 %	1 %	1 %	66 %
FFT	Creating an environment of respect and rapport	Rater	0.005	0.001	0.001	2 %	0 %	1 %	0 %	97 %
FFT	Creating an environment of respect and rapport	Residual	0.13	0.056	0.056	0 %	0 %	0 %	0 %	100 %
FFT	Creating an environment of respect and rapport	Lesson Rater Error Facets		0.038	0.038					
FFT	Communicating with students	School	0.019		0.019	14 %	23 %	0 %	24 %	39 %
FFT	Communicating with students	Teacher	0.015		0.015	0 %	13 %	8 %	2 %	78 %
FFT	Communicating with students	Classroom	0.009		0.009	13 %	0 %	12 %	0 %	75 %

			Variance associated with each facet			**Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error**				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
FFT	Communicating with students	Lesson	0.018		0.018	23 %	0 %	0 %	0 %	77 %
FFT	Communicating with students	Rater	0.021	0.031	0.031	5 %	0 %	1 %	0 %	93 %
FFT	Communicating with students	Residual	0.189	0.136	0.136	1 %	0 %	0 %	0 %	98 %
FFT	Communicating with students	Lesson Rater Error Facets		0.157	0.157					
FFT	Engaging students in learning	School	0.014		0.014	23 %	27 %	1 %	19 %	30 %
FFT	Engaging students in learning	Teacher	0.035		0.035	27 %	17 %	0 %	0 %	56 %
FFT	Engaging students in learning	Classroom	0		0	0 %	0 %	0 %	100 %	0 %
FFT	Engaging students in learning	Lesson	0.029		0.029	21 %	11 %	0 %	0 %	68 %
FFT	Engaging students in learning	Rater	0.052	0.045	0.045	1 %	0 %	0 %	0 %	99 %
FFT	Engaging students in learning	Residual	0.221	0.203	0.203	1 %	0 %	0 %	0 %	99 %
FFT	Engaging students in learning	Lesson Rater Error Facets		0.134	0.134					

			Variance associated with each facet			**Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error**				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
FFT	Flexibility and Responsiveness	School	0.012		0.012	10 %	49 %	2 %	14 %	26 %
FFT	Flexibility and Responsiveness	Teacher	0.008		0.008	27 %	31 %	0 %	0 %	42 %
FFT	Flexibility and Responsiveness	Classroom	0		0					
FFT	Flexibility and Responsiveness	Lesson	0.009		0.009	30 %	0 %	0 %	0 %	70 %
FFT	Flexibility and Responsiveness	Rater	0.022	0.009	0.009	3 %	0 %	0 %	0 %	96 %
FFT	Flexibility and Responsiveness	Residual	0.207	0.182	0.182	0 %	0 %	0 %	0 %	100 %
FFT	Flexibility and Responsiveness	Lesson Rater Error Facets		0.119	0.119					
FFT	Demonstrating Content Knowledge	School	0.02		0.02	13 %	42 %	1 %	18 %	26 %
FFT	Demonstrating Content Knowledge	Teacher	0.017		0.017	18 %	16 %	0 %	2 %	65 %
FFT	Demonstrating Content Knowledge	Classroom	0.008		0.008	35 %	34 %	10 %	8 %	13 %
FFT	Demonstrating Content Knowledge	Lesson	0.006		0.006	20 %	12 %	0 %	0 %	67 %

			Variance associated with each facet			Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
FFT	Demonstrating Content Knowledge	Rater	0.035	0.011	0.011	1 %	0 %	0 %	0 %	98 %
FFT	Demonstrating Content Knowledge	Residual	0.226	0.191	0.191	0 %	0 %	0 %	0 %	100 %
FFT	Demonstrating Content Knowledge	Lesson Rater Error Facets		0.228	0.228					
FFT	Managing Classroom Procedures	School	0.01		0.01	17 %	25 %	0 %	19 %	39 %
FFT	Managing Classroom Procedures	Teacher	0.037		0.037	7 %	0 %	4 %	1 %	89 %
FFT	Managing Classroom Procedures	Classroom	0.011		0.011	32 %	8 %	0 %	0 %	59 %
FFT	Managing Classroom Procedures	Lesson	0.018		0.018	12 %	3 %	0 %	2 %	83 %
FFT	Managing Classroom Procedures	Rater	0.024	0.005	0.005	2 %	0 %	1 %	0 %	97 %
FFT	Managing Classroom Procedures	Residual	0.208	0.201	0.201	1 %	0 %	0 %	0 %	99 %
FFT	Managing Classroom Procedures	Lesson Rater Error Facets		0.245	0.245					
FFT	Managing Student Behavior	School	0.006		0.006	7 %	21 %	7 %	23 %	43 %

			Variance associated with each facet			Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
FFT	Managing Student Behavior	Teacher	0.036		0.036	2 %	1 %	0 %	0 %	97 %
FFT	Managing Student Behavior	Classroom	0.014		0.014	23 %	6 %	2 %	2 %	68 %
FFT	Managing Student Behavior	Lesson	0.014		0.014	1 %	0 %	1 %	0 %	98 %
FFT	Managing Student Behavior	Rater	0.011	0.002	0.002	7 %	0 %	2 %	0 %	91 %
FFT	Managing Student Behavior	Residual	0.133	0.078	0.078	0 %	0 %	0 %	0 %	99 %
FFT	Managing Student Behavior	Lesson Rater Error Facets		0.107	0.107					
FFT	Organizing Physical Space	School	0.016		0.016	18 %	30 %	0 %	7 %	45 %
FFT	Organizing Physical Space	Teacher	0.015		0.015	9 %	15 %	15 %	8 %	53 %
FFT	Organizing Physical Space	Classroom	0		0	50 %	0 %	50 %	0 %	0 %
FFT	Organizing Physical Space	Lesson	0.028		0.028	17 %	3 %	0 %	0 %	80 %
FFT	Organizing Physical Space	Rater	0.007	0.024	0.024	4 %	0 %	4 %	0 %	92 %

			Variance associated with each facet			**Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error**				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
FFT	Organizing Physical Space	Residual	0.189	0.188	0.188	0 %	0 %	0 %	0 %	100 %
FFT	Organizing Physical Space	Lesson Rater Error Facets		0.179	0.179					
FFT	Use of Questioning and Discussion Techniques	School	0.01		0.01	18 %	38 %	0 %	15 %	30 %
FFT	Use of Questioning and Discussion Techniques	Teacher	0.028		0.028	25 %	11 %	3 %	4 %	57 %
FFT	Use of Questioning and Discussion Techniques	Classroom	0.001		0.001	100 %	0 %	0 %	0 %	0 %
FFT	Use of Questioning and Discussion Techniques	Lesson	0.017		0.017	10 %	0 %	0 %	0 %	90 %
FFT	Use of Questioning and Discussion Techniques	Rater	0.014	0.004	0.004	5 %	0 %	1 %	0 %	94 %
FFT	Use of Questioning and Discussion Techniques	Residual	0.18	0.172	0.172	0 %	1 %	0 %	0 %	99 %

			Variance associated with each facet			**Percentage of facet variance in the final data model associated with the sets of proxy measures for facets not associated with rater error**				
Protocol	**Dimension**	**Facet**	**Full Data**	**Calibration Data**	**Final Model**	**Teaching context proxy measures**	**Only student proxy measures**	**Only teacher proxy measures**	**Both student & teacher proxy measures**	**Residual (i.e., not associated with any proxy measure)**
FFT	Use of Questioning and Discussion Techniques	Lesson Rater Error Facets		0.154	0.154					
FFT	Use of Assessment in Instruction	School	0.006		0.006	0 %	22 %	3 %	14 %	61 %
FFT	Use of Assessment in Instruction	Teacher	0.005		0.005	6 %	19 %	0 %	0 %	76 %
FFT	Use of Assessment in Instruction	Classroom	0.008		0.008	35 %	0 %	2 %	0 %	63 %
FFT	Use of Assessment in Instruction	Lesson	0.016		0.016	30 %	12 %	0 %	0 %	58 %
FFT	Use of Assessment in Instruction	Rater	0.035	0.024	0.024	0 %	0 %	0 %	0 %	99 %
FFT	Use of Assessment in Instruction	Residual	0.195	0.17	0.17	0 %	0 %	0 %	0 %	100 %
FFT	Use of Assessment in Instruction	Lesson Rater Error Facets		0.215	0.215					