Educational Assessment, Evaluation and Accountability
https://doi.org/10.1007/s11092-024-09427-8

Signal, error, or bias? exploring the uses of scores from observation systems

Mark White[1] · Kirsti Klette[1]

Received: 13 July 2023 / Accepted: 28 December 2023

Abstract

Scores from observational measures of teaching have recently been put to many uses within school systems, including communicating a standard of practice and providing teacher feedback, identifying teachers for professional development, monitoring system equity, and making employment decisions. In each of these uses, observation scores are interpreted as representing some aspect of the enacted instruction or teachers' capacity to enact instruction, as seen through the observation systems lens for understanding teaching quality. The quality of these interpretations, or the extent to which observation scores are composed of a signal that accurately reflects the interpretation, has important implications for the overall validity of uses of observation systems. Starting from an explicit conceptualization of instruction, this paper combines generalizability theory and hierarchical linear modelling approaches to decompose observation scores to explore the extent to which scores from observation systems are composed of signal, error, and bias across four different uses (i.e., teacher feedback, professional development, monitoring system equity, and employment decisions) of scores. We show that the quality of observation scores may depend more on what scores are interpreted as representing (i.e., the proposed use) than on the specific observation rubric being used. Further, we show that rater errors and biases are a major threat to any attempt to interpret observation scores as capturing the observation system's understanding of teaching quality. We discuss implications for using scores from observation systems.

Keywords Teaching quality · Data use · Observation system · Teacher quality

An earlier version of this paper was presented at the AERA conference in 2022.

✉ Mark White
mark.white@ils.uio.no

[1] Department of Teacher Education and School Research, University of Oslo, Postboks 1099, Blindern, 0317 Oslo, Norway

Published online: 30 January 2024

1 Introduction

Efforts to directly measure teaching quality have a long history, arising from the process–product research of the 1970–1980s (e.g., Brophy, 1973; Brophy & Good, 1984; Emmer et al., 1979). Modern approaches arising from this tradition have come to be called observation systems (ObsSys). ObsSys are used in many ways, including comparing enacted instruction across settings (e.g., Martinez et al., 2016; OECD, 2020; Praetorius et al., 2019), providing feedback to teachers (e.g., Cohen et al., 2016; Kraft & Hill, 2020), and teacher evaluation (e.g., Kraft & Gilmour, 2017; Steinberg & Donaldson, 2016). Recent work on ObsSys highlights several challenges that call into question the appropriateness of these uses, including critiques of the lack of conceptualization of teaching (e.g., Charalambous & Praetorius 2020), mismatches between measurement approaches and underlying conceptualizations of teaching quality (e.g., White et al., 2022) and high rater error (e.g., Bell et al., 2014). These challenges raise the need to carefully consider how well scores from ObsSys support their intended uses.

This paper addresses the following research question: To what extent are scores from ObsSys composed of signal, error, and/or bias across four common uses of observation scores? In addressing this question, the paper empirically explores the validity of common interpretations of scores from ObsSys that are tied to specific score uses, highlighting how an explicit conceptualization of teaching can identify multiple distinct factors that influence enacted instruction and factors that could contribute either useful signal, error, and/or bias. We then model, in two ObsSys, the extent to which scores from ObsSys are composed of signal, error, and/or bias when scores are used for different purposes, providing evidence towards the validity of different interpretations of scores from ObsSys. The paper highlights several important points. First, the proposed interpretations of scores have an important impact on the quality of scores, possibly more than the specific ObsSys being used. Second, across all uses, rater error contributes a large amount of potential bias. Only after better understanding the nature of rater error's contribution to scores can one effectively judge the validity of different uses.

2 A conceptualization of teaching and its measurement

We start by presenting a basic conceptualization of instruction and its measurement, which allows us to consider the various contributors to observation scores. Figure 1 shows this paper's conceptualization, which is based on the didactic triangle (Klafki, 2000) (often termed instructional triangle in the US context; Cohen et al., 2003). Teachers' skills and knowledge, students, and the instructional context (e.g., lesson goals, content) each influence teachers' choices in what and how to teach, all of which contribute to determining the enacted instruction. The enacted instruction is observed through the lens of raters' understanding of the observation rubric to generate estimates of observed teaching quality or

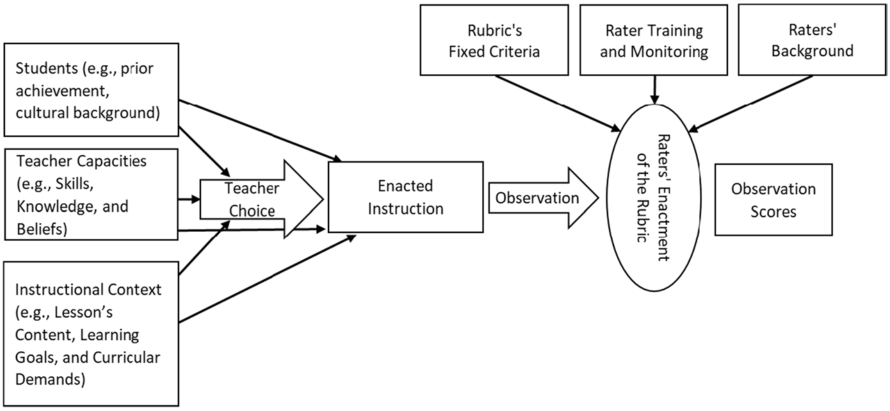
Rubric's Fixed Criteria
Rater Training and Monitoring
Raters' Background
Students (e.g., prior achievement, cultural background)
Teacher Capacities (e.g., Skills, Knowledge, and Beliefs)
Teacher Choice
Enacted Instruction
Observation
Raters' Enactment of the Rubric
Observation Scores
Instructional Context (e.g., Lesson's Content, Learning Goals, and Curricular Demands)

Fig. 1 Conceptualization of teaching and measuring teaching quality

observation scores. The explicit assumption here is that observation scores are not direct or clear reflections of either enacted instruction or teacher capacities but arise in the complex interplay of students, content, and teachers, as filtered through a specific observation system's understanding of teaching quality and the potentially distorted lens of rater perceptions. One cannot, then, simply interpret observation scores as directly representing any particular construct, but observation scores represent a complex interplay of several factors (i.e., students, instructional contexts, teachers, raters).

Each factor represented on the left-hand side of Fig. 1 has the potential to impact observation scores. The instructional context includes all features associated with lessons that might impact enacted instruction, including time of year, day of the week, the lesson content and learning goals, aspects of the curriculum, and many other factors (e.g., Bohn et al, 2004; Casabianca et al., 2015). Organizational and curricular structures, as well as the nature of different subject areas or learning goals, may lead instruction to be systematically different across features of the instructional context (Kelcey & Carlisle, 2013). Students are conceptualized as active contributors to the quality of enacted instruction through their participation in instructional routines and their cognitive engagement with the content of instruction (or lack thereof; Cohen et al., 2003). This contribution of students is reciprocal, as teachers may also choose instructional approaches based on their perception of student capacities (e.g., The New Teacher Project [TNTP], 2018). Last, teacher capacities, which include all stable characteristics of teachers, most notably their skills, knowledge, and repertoires, are assumed to directly impact enacted instruction both through the choice of what to teach and the skill with which instruction is enacted.

Beyond these factors that together determine the quality of enacted instruction lie factors that determine the lens that encodes enacted instruction into observation scores (i.e., the raters' enactment of the rubric, see Fig. 1). This includes, most importantly, the rubric's fixed criteria for assessing teaching quality and the systems that are in place to train raters on applying the observation rubric and monitor raters' faithful application of the rubric. Last, raters' background (including knowledge and beliefs) also impacts how raters enact the rubric (e.g., Cash et al., 2012). Ideally, raters accurately and consistently score the rubric so that scores capture the intended understanding of teaching quality. However, evidence suggests that raters often introduce both random and systematic sources of error to observation scores (e.g., Bell et al., 2014). Table 1 defines four types of rater error that may arise, as well as highlighting the complications for interpreting observation scores that can arise from each type of rater error.

2.1 Different uses of scores involve different interpretations of what scores represent

We assume here that each use of observation scores rests on the interpretation of scores as representing a feature of classroom instruction and/or teacher capacities (i.e., "interpretation for use"; AERA, APA, & NCME, 2014). For example, when used to identify teachers for professional development, observation scores are interpreted as capturing teachers' capacity to provide a certain type of instruction, such that teachers who struggle to enact a given type of instruction can receive support to build their capacity. On the other hand, when communicating a standard of practice and providing feedback to teachers, observation scores are interpreted as representing the quality of enacted instruction (rather than teachers' capacity to enact instruction), such that teachers receive feedback on the strengths and weaknesses of the enacted instruction. Since each use of observation scores rests on different interpretations of what scores represent, observation scores are differentially useful depending on the extent to which the scores represent the intended construct (i.e., the size of the signal, error, and bias).

2.2 Observation scores as representations of specific understandings of teaching quality

Users of ObsSys have many choices in which ObsSys to adopt. They could, further, simply rely on the professional judgement of observers rather than using a formal ObsSys. Considering the cost in time and resources of implementing an ObsSys (e.g., Bell et al., 2015), we assume that the given ObsSys was chosen because the way it defines and operationalizes teaching quality is thought to be particularly meaningful. As such, we acknowledge multiple (potentially contradictory) understandings of teaching quality and emphasize the importance of interpreting observation scores as representing the particular understanding of teaching quality operationalized by the chosen ObsSys.

2.3 Definitions

In line with this understanding, we define some terms, which come from signal detection theory, a field that emphasizes extracting useful information from noisy environments (Abdi, 2007). The signal is the variation in observation scores that is associated with the intended construct, as viewed by the understanding of teaching

Table 1 Explanations and complications for different types of rater error

Type of rater error	Definition	Complications arising from this rater error
Shared bias	All raters systematically make the same error when scoring. For example, confusions in training may lead all raters to misunderstand some rubric dimensions. Alternatively, shared prior experiences in classrooms may lead to scores that are based on local understandings of teaching quality rather than the understanding of teaching quality embedded in the rubric	• This error could systematically bias estimates of the reliability and concurrent validity of observation scores • Error is invisible when only looking at inter-rater agreement
Rater bias	One rater systematically makes the same error when scoring specific types of instruction. For example, a misunderstanding of the rubric may lead a rater to systematically inflate scores for instruction containing independent seatwork	• This error could systematically bias estimates of the reliability and concurrent validity of observation scores • Absent efforts to specifically test for rater bias across features of instruction that are related to rater bias, rater bias is likely to be misinterpreted as random error
Rater random error	One rater systematically makes the same error when scoring all instances of instruction (i.e., rater leniency/severity). For example, a misunderstanding of the thresholds between score categories may lead a rater to systematically score all instruction higher than deserved	• Assuming the random assignment of raters, this error introduces random error to observation scores
Random error	One rater non-systematically makes errors when scoring. For example, a rater may mishear or not hear something said by a student. Here, non-systematic implies that raters may not make the same mistake if they were to score the lesson again while systematic implies that raters are likely to make the same mistake in every viewing of the lesson	• Reduces power to study teaching quality

embedded in the ObsSys. The specific nature of the signal varies for each use of scores and is discussed in more detail in the next section. The variation in observation scores that is not signal is further divided into bias (i.e., systematic deviations of scores from signal) and error (i.e., random deviations of scores from signal). The specific nature of bias and error varies by the use of observation scores and is discussed in the following section.

3 Multiple uses of observation scores

This section defines the nature of signal, error, and bias for each of the explored uses of observation scores: (i) communicating a standard of practice and providing feedback, (ii) identifying teachers for professional development, (iii) ensuring equitable access, and (iv) making employment decisions.

3.1 Communicating a standard of practice and providing feedback relative to this standard

This use of scores considers the observation rubric as defining instructional standards or describing the instruction that is expected of teachers (e.g., Danielson, 2000; Kraft & Hill, 2020). In this use, the observed instruction is compared to the observation rubric to judge the quality of enacted instruction relative to the rubric's standards. Then, observation scores are interpreted as representing the quality of enacted instruction as operationalized by the ObsSys. All factors impacting the quality of enacted instruction (see Fig. 1) are relevant considerations (i.e., signal) when making sense of observation scores. Errors and biases, then, are introduced only through the scoring process (i.e., rater errors). For example, errors could arise if raters fail to observe an event that leads them to assign the wrong scores while biases could arise if raters systematically misinterpret the observation rubric categories (see Table 1). Table 2 summarizes the definitions of signal, error, and bias across uses.

3.2 Identifying teachers for professional development (pd)

This use of scores identifies teachers who lack the skills and/or knowledge that will be addressed by a given PD offering to ensure that the offering is given to teachers who would benefit from it. Here, observation scores are interpreted as providing information about teachers' skills and/or knowledge (see Table 2). For example, if little student discussion is repeatedly observed in a teacher's instruction, that teacher may be recommended to a PD session on leading discussions to build their capacity to lead discussions. Since many factors other than teacher skills and knowledge affect observation scores (see Fig. 1), there are many potential sources of error or bias when using scores for this purpose. For example, variation in scores across lessons or classrooms would be a source of error since this use of scores tries to estimate teacher's capacities to provide

Table 2 Sources of signal, error, and bias across the different uses of observation scores

Specified use	Sources of signal	Sources of error	Sources of bias	Notes on use
Communicating a standard of practice and providing feedback relative to this standard	The variation associated with the quality of enacted instruction as interpreted through the ObsSys's lens	The variation associated with non-systematic rater error	The variation associated with systematic rater errors	We assume that the goal is to provide feedback on the ObsSys's operationalized understanding of teaching quality
Identifying teachers for PD	The variation associated with teachers' knowledge or skills as interpreted through the ObsSys's lens	The variation associated with the instructional context and non-systematic rater error	The variation associated with teachers' choice in instructional practice and systematic rater errors	We assume PD is focused on building teachers' capacity to enact instruction
Ensuring equitable access to teaching quality	The variation associated with average enacted instruction at the student sub-group level as interpreted through the ObsSys' lens, including variation driven by differences in the enacted curriculum across sub-groups	The variation associated with (1) biases in sampling the instructional context; (2) sampling of lessons, classrooms, and teachers in general; and (3) non-systematic rater error	The variation associated with systematic rater errors	
Making employment decisions	The variation associated with teachers' capacities as interpreted through the ObsSys's lens	The variation associated with (1) the instructional context and (2) non-systematic rater error	The variation determined by (1) the students being taught and (2) systematic rater errors	We assume a system-level effort to identify the teachers with the highest overall level of capacity (e.g., skills and knowledge)

instruction to their students. This is generally true for any variation in scores that is associated with the instructional context. An example of a bias for this use is if teachers choose to avoid certain instructional practices (e.g., discussions) that they can successfully enact (due to, for example, the lesson's learning goals or implicit assumptions about their students' capacities/needs). Since, in this scenario, observation scores would then (incorrectly) suggest that teachers lack the capacity to engage in certain types of instruction (e.g., discussions).[1]

3.3 Ensuring equitable access

This use of scores seeks to measure the enacted instruction experienced by students within student sub-groups and then compare the measured level of enacted instruction across sub-groups. The desire here is that different sub-groups of students will have experienced equal levels of enacted instruction. Here, observation scores are interpreted as representing the average quality of the enacted instruction for a sub-group of students, as interpreted through the ObsSys's lens (see Table 2).[2] For example, a school system may compare the average observation scores from all schools with over 10% of immigrant students to all schools with less than 10% of immigrant students to determine whether immigrant students receive equally effective instruction as non-immigrant students.

The aggregation here creates challenges that are not present in the feedback use, despite both uses interpreting scores as representing enacted instruction. For example, imagine that writing instruction was observed more often in schools with many immigrant students and that writing lessons typically received lower observation scores. If this difference in the frequency of writing instruction reflected a difference in schools' enacted curriculum (Polikoff & Porter, 2014), the difference in observation scores would reflect actual differences in the enacted instruction received by students (i.e., be a part of the signal). However, if the difference in the frequency of writing instruction were driven by some schools observing proportionally more writing lessons (e.g., because those schools were trying to improve the quality of writing instruction), this would reflect a sampling bias and the difference in scores associated with differences in writing instruction would reflect error (if non-systematic) or bias (if systematic). Without knowing whether there are differences in the enacted curriculum across sub-groups, interpreting the variation in scores associated with the instructional context, then, can be challenging in this use of ObsSys.

3.4 Employment decisions: hiring, bonuses, tenure, etc.

This use of scores seeks to identify teachers with specific skills and/or knowledge in order to provide an evidence base with which to make employment decisions (e.g.,

[1] Note that this might not reflect a bias if the PD was oriented to changing teachers' beliefs about when, where, and why to use specific instructional practices (e.g., discussion), as opposed to the skill-focused PD that we emphasize.

[2] We assume throughout that the chosen observation system characterizes teaching quality equally well for all sub-groups of students, but this assumption should be questioned (c.f. Milanowski, 2017).

Dee & Wyckoff, 2015). Note that employment-based decisions often combine data from multiple sources. This paper focuses solely on assessing the contribution to overall scores that come from ObsSys under the assumption that each part of a composite score should be independently examined. In this use, observation scores are interpreted as representing teachers' capacity to enact rubric-aligned instruction (see Table 2).

Employment decisions at different levels of a school system may interpret scores differently (implying that signals vary across different uses within this category). For example, a school-level system will likely interpret scores as capturing teachers' capacity to enact high-quality instruction within the given school while a statewide system will likely interpret scores as capturing teachers' average capacity to enact high-quality instruction in schools across the state, a far broader interpretation of scores that is more challenging to defend. For example, research suggests that most of the race gap in observation scores occurs between schools (Steinberg & Sartain, 2021). This between-school race gap would not affect interpretations of scores at the within-school level (since the race gap is negligible within-schools) but points to a potential bias in scores when scores are interpreted as capturing teachers' capacity to enact rubric-aligned instruction across all schools in the state. Similarly, research suggests that between-school differences in scores may be driven mostly by rater error (Cowan et al., 2022), further complicating between-school interpretations of observation scores.

Since observation scores are interpreted as providing information about teacher capacities and beliefs while enacted instruction is determined by many factors, there are many potential sources of error for this use of scores. For example, the non-random sampling of lessons may lead teachers to be observed in different instructional contexts (e.g., reading vs. writing lessons) so variation in the instructional context is likely to reflect error (i.e., be unrelated to teacher capacities; though note the discussion of the enacted curriculum above). Other sources of variation in scores are less clear-cut. Consider the variation in scores associated with student characteristics (e.g., Campbell & Ronfeldt, 2018; Steinberg & Sartain, 2021). If this variation is caused by how easy it is to teach some students,[3] that variation would be bias since it is driven by student factors and not teachers' knowledge and skills. However, that variation could be the result of teacher sorting (e.g., Goldhaber et al., 2015), such that teachers with higher levels of skill and knowledge happen to teach specific groups of students (i.e., the relationship between student characteristics and observation scores is exogenous; White, 2023), in which case the variation in scores associated with student characteristics would be part of the signal (i.e., accurately reflect teacher capacities).

[3] One might, for example, propose that students with more prior knowledge are more receptive to instruction than other students and/or that students have a pre-existing tendency to actively engage with content. Either of these or other causes might make it decidedly easier to teach some students than others.

3.5 Section summary

Different uses of scores are supported by different interpretations of observation scores. Different interpretations, in turn, imply that different aspects of observation scores are signal, error, or bias (see Table 2). The first takeaway of this paper is the need to be clear and explicit about how one is interpreting observation scores (i.e., defining the signal) and how this interpretation supports a proposed use of those scores. Only after doing so can one appropriately turn to consider the potential sources of error and bias that might impact the extent to which an observation score supports a given use.

Recall again that this paper explores the extent to which observation scores accurately reflect the construct that scores are interpreted as representing. Since observation scores are interpreted in order to support each proposed score use (AERA, APA, & NCME, 2014), the interpretation of observation scores is an important aspect of a full validity argument supporting a given use of scores, but it is only one aspect. A full consideration of the validity of any given use of observation scores would need to examine potential impacts of the proposed use. Further, some validity arguments might not propose interpretations of scores but focus solely on the consequences of the proposed use. In this case, the current paper has little relevance. For example, in teacher evaluation systems, it is widely accepted that principals assign scores to balance a wide range of different (unknown) goals (Halverson et al., 2004), obscuring any clear interpretation of scores. However, the validity of providing feedback to teachers with such scores is still advocated based on the face validity of the process and the positive consequences of the given use (Halverson et al., 2004). Despite this limitation, the time and cost of implementing an ObsSys (e.g., Bell et al., 2015) mean that most uses of observation scores will (at least implicitly) seek to interpret scores as representing a specific construct.

4 Methods

This section introduces the data and analysis approaches. These approaches were designed to decompose observation scores into signal, bias, and error when using observation scores across a range of purposes.

4.1 Understanding Teacher Quality (UTQ)

Data comes from secondary analyses of a research project meant to evaluate ObsSys called the Understanding Teaching Quality project (UTQ; Casabianca et al., 2015). In the UTQ project, 458 volunteer mathematics and English language arts teachers in grades 6–8 in three large, southeastern US school systems were observed in 2009–2011. All 228 teachers who taught English are used in analyses presented in this paper. Teachers were observed teaching up to four lessons (two lessons for each of the two classrooms). Each video was scored using the Classroom Assessment Scoring System (CLASS; Pianta et al., 2010) and the Framework for Teaching (FFT; Danielson, 2000).

Twelve former teachers were recruited as raters and underwent standard training and certification for both CLASS and FFT (and one other protocol). Rater calibration exercises were held weekly (once every 3 weeks for each protocol). During these low-stakes exercises, raters watched and discussed scores in a common video, receiving feedback on scoring and documenting rater scoring accuracy across time. Rater agreement was tested by randomly selecting 25% of videos (one per teacher) to double-score. Following standard protocols, CLASS was scored using 15-min segments (usually three per lesson) and FFT was scored using 30-min segments (usually 1 segment per lesson). All data in this paper 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 way of estimating average teaching quality in a lesson (c.f., White et al., 2022).

Two different data sets from the UTQ project were analyzed: first, the full data set, which includes the scores assigned by raters during normal scoring, and second, the calibration data set, which includes the scores assigned independently by raters prior to participating in calibration meeting discussions. The calibration data is unique because it contains master scores, which we interpret as estimates of the score that would have been assigned had the rater perfectly applied the observation rubric. Further, raters scored the same video multiple times in the calibration data, allowing us to distinguish between systematic rater errors (i.e., errors made repeatedly across scoring occasions) and random rater errors (i.e., errors not made repeatedly across scoring occasions).

One cannot empirically identify the complete contribution of students, the instructional context, and teacher capacities to the quality of enacted instruction, as measured by the ObsSys. Rather, there is a need to use proxy measures to empirically explore the relative contribution of each factor. The assumption here is that the variation in observation scores that is systematically related to the proxy measures represents variation that is driven by the factor those proxy measures represent. The instructional context factor used the following proxy measures: the semester of the observed lesson (Fall or Spring); the classroom's grade level; the time of day, the month, and the day of the week of the lesson; and curriculum indicators capturing whether the lesson focused on grammar, literature, reading comprehension, and writing. The student factor used the following proxy measures (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 was also used. Over-fitting was avoided by using the first two principal components in regression models (Greco et al., 2019), which accounted for 39% and 15% of the variance, respectively. The teacher capacity factor used the following proxy measures: highest educational degree, whether the teacher was certified in their subject area, whether the teacher was certified to teach middle school, years of experience, a content knowledge measure called Teacher Knowledge Assessment System (TKAS;

Phelps et al., 2014), value-added scores from the current year, and teacher value-added scores from the previous year (see Lockwood & McCaffrey, 2012 for value-added specifications). Over-fitting was again avoided by using the first two principal components, which accounted for 24% and 18% of the variance, respectively.

4.2 Augmented generalizability theory (GT) models

Our main analytic approach is augmented GT models (Casabianca et al., 2015), which blend typical GT models (Brennan, 2001) with the hierarchical linear modelling (HLM; Raudenbush & Bryk, 2001) approach of examining the decrease in the size of variance components after adding fixed effects in the model. Analyses were done using package *lme4* (Bates et al., 2015) in R (R Core Team, 2020). This lets us examine the extent to which each measurement facet is associated with the three focal factors: the instructional contexts, students, and teacher capacities. The variation associated with each measurement facet is then recast as a source of signal, error, and/or bias based on the discussion in the section "Multiple uses of observation scores" and summarized in Table 2 (for further details on this approach, see White, 2022).

Models run on the full data set use the assigned score as the outcome measures while calibration data models use the assigned score minus the master score as the outcome measure. Then, the calibration models purely decompose rater error and are not like other GT models. Their goal is to distinguish between the four types of rater error that are described in Table 1, especially the systematic versus non-systematic types of rater error.

The augmented GT models decompose scores into components that capture signal, error, or bias for a given use of scores. However, several uses of scores involve aggregating scores across many observations to estimate classroom-level, teacher-level, or school-level teaching quality. Under the assumption that the modelled facets are normally distributed and random (i.e., teachers are observed in a randomly selected instructional context in each observation), aggregating across a facet will decrease that facet's contribution to the variance of scores. Note that some systematic errors cannot be averaged across to reduce their impact on the resulting scores. For example, the lesson and segment facets in the calibration model capture bias made by all raters, so the influence of this systematic error cannot be reduced by aggregating across more raters.

For the use of communicating a standard of practice, the feedback was assumed to be focused on a single lesson observed by one observer. For identifying teachers for PD and making employment decisions, scores were assumed to be aggregated across four lessons observed by a total of 2 raters. For ensuring equitable access, scores were assumed to be aggregated to the school level with ten teachers in each school and each teacher observed across two lessons with a different observer in each lesson. Note that the number of lessons and raters that we aggregate across is based on formal recommendations (e.g., Kane et al., 2012) but involves more lesson observations and raters than are typically available in practice-based settings (e.g., Steinberg & Donaldson, 2016), which again should, if anything, introduce conservative biases, making results overly optimistic about score quality.

5 Results

5.1 Variance decomposition of scores

Table 3 presents the main results of the augmented GT models. The columns under “variance associated with each facet” show the variance in observation scores associated with each facet for each data set. Note that the calibration data models only estimated facets associated with rater error to provide a more detailed look into rater accuracy while the full data models examined variation in scores across schools, teachers, classrooms, and lessons, as well as rater agreement. The last set of columns looks at the percentage of variance within each facet that is explained by the three factors impacting enacted instruction (see Fig. 1). Appendix A replicates Table 3 for each dimension of FFT and CLASS.

When examining Table 3, note that the variance decomposition estimates for the main models fall within the expected range given past work (e.g., Kane et al., 2012; Mantzicopoulos et al., 2018), though differences in model specification and analytic choices (e.g., whether to aggregate to the lesson level before analyses) make precise comparisons with past work difficult. There may, though, be somewhat less variation in scores across teachers/schools and more residual variation than in past studies (Bell et al., 2014). We know of no prior studies that have tried to systematically explore the extent to which proxy measures explain school facets or that decompose the accuracy of rater scoring in data with master scores so we cannot state whether these analyses align with past work.

Table 3 shows that a large percentage of the variation in the school facet is explained by the proxy measures for student, teacher capacity, and the instructional context. This implies that school-level average observation scores are highly related to both school demographics and the observed instructional contexts within the school (see also Steinberg & Sartain, 2021). A fair amount of the teacher facet is related to proxy measures for the instructional context and students, suggesting that within-school differences in scores across teachers are associated with the types of students each teacher teaches and the instructional context in which teachers were observed.

When looking at the rater and residual facets, which were estimated in both the full data and calibration models, we see that these facets are both smaller in the calibration data. Since the calibration data captures rater accuracy (i.e., distance from the master score) while the full data captures only rater agreement, we use the rater error estimates from the calibration data models in the next section. Since these estimates are lower, this should again only potentially introduce conservative biases, making rater error seem less problematic than it is.

Table 3 Variance decomposition of scores and variance attributable to each factor

Obs. system	Facet (type of rater error)	Variance associated with each facet		Percentage of the facet in the full data model associated with the proxy measures				
		Full data	Calibration data	Instruct context	Only student	Only teacher	Both student and teacher	Residual (i.e., not associated with any factor)
CLASS	School	0.032		13%	41%	0%	19%	28%
CLASS	Teacher	0.045		18%	5%	2%	2%	73%
CLASS	Classroom	0.007		0%	0%	14%	0%	86%
CLASS	Lesson	0.041		39%	2%	0%	0%	59%
CLASS	Rater (rater random error, or leniency)	0.061	0.015	11%	0%	0%	0%	89%
CLASS	Residual (random error)	0.158	0.022	0%	0%	0%	0%	100%
CLASS	Rater shared bias (lesson and segment facets)		0.042					
CLASS	Rater-by-lesson (rater bias)		0.081					
FFT	School	0.011		10%	40%	0%	20%	30%
FFT	Teacher	0.018		11%	11%	0%	5%	74%
FFT	Classroom	0.002		100%	0%	0%	0%	0%
FFT	Lesson	0.008		25%	0%	0%	0%	75%
FFT	Rater (rater random error, or leniency)	0.012	0.006	0%	0%	0%	0%	100%
FFT	Residual (random error)	0.061	0.040	0%	0%	0%	0%	100%
FFT	Rater shared bias (lesson and segment facets)		0.062					

Teacher is the teacher capacity; student is the student characteristics; instruct context is the instructional context. 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

5.2 Signal, systematic error, and random error across uses of scores

This section combines the conceptual view of teaching and discussion of different uses of observation scores with the augmented GT models just discussed. Using the discussion of the various uses of scores from ObsSys, we break down the extent to which scores, after aggregation, are composed of signal, systematic error, or random error across the four discussed uses. A note on the interpretation of findings is important first. The percentage of the score attributable to signal can also be interpreted as the reliability of the scores (under a given interpretation) for the given purpose (i.e., since reliability is the percentage of non-error variation in scores). Then, traditional standards would expect signal to be about 70% of scores for low-stake uses and 80% of scores for high-stake uses (i.e., reliability is 0.7 or 0.8; Kane et al., 2012). However, most uses of reliability do not consider bias and even low levels of bias can make scores untenable for specific uses, even if reliability is high. For example, when using scores to examine if minoritized and non-minoritized students receive equal levels of teaching quality, a small bias that leads classrooms with more non-minoritized students to score higher than classrooms with fewer non-minoritized students can lead scores to be problematic, even if score reliability is high (e.g., Campbell & Ronfeldt, 2018). Then, interpretations of scores are most defensibly used for a given purpose when bias is near zero and signal is over 70–80%. That said, this paper focuses only on the interpretation of scores as capturing the intended construct, so one cannot make clear conclusions about the validity of specific uses of scores because if specific interpretations of scores are not justifiable, then new interpretations of scores can be built to support those uses.[4]

Note that uncertainties in the causal explanation of the observed association between the three factors and scores lead, on occasion, to some portion of the variance in scores to be an unknown combination of signal and error. Figure 2 shows the results of this decomposition across the four discussed uses of scores. The variation in scores labelled “bias or signal” in identifying teachers for PD is the variability in scores associated with students. When this variability is driven by teachers’ choice in what to teach, this variability may be bias since teachers who are simply choosing to not enact specific practices based on their belief in what is best for students may not benefit from skill-focused PD offerings. Otherwise, this variability is signal. The variation in scores labelled “random error or signal” for ensuring equitable access is driven by factors related to the instructional context. When this variation is associated with sampling errors, it is random error and when it is associated with true differences in the enacted curriculum across student sub-groups, it is signal.

[4] An example of a new interpretation includes arguing that it is not important to measure the rubric-intended understanding of teaching quality, but that measuring raters’ idiosyncratic understandings of teaching quality is acceptable. In this case, the rater errors and biases might be understood as signal that represents raters’ idiosyncratic understandings of teaching quality (note that under this interpretation, scores are not necessarily comparable across or even within raters). A second example of a new interpretation includes arguing that using scores has led to a specific positive impact, such that the further use of scores is justified regardless of what scores represent. Comparisons of scores are also not necessarily valid under this interpretation.

For the employment decision use, variation in scores labelled “signal or bias” is the variability associated with between-school differences in scores and the between-teacher variation in scores associated with student characteristics. If this is due to the systematic sorting of teachers within and across schools, it is signal; otherwise, it is bias (as this would imply teachers’ scores are dependent on the school within which they teach). For each of these cases where it is unclear whether scores are signal or bias/error, it would be, in principle, possible to adjust scores (using our proxy measures) to remove this variation. Such adjustments would remove this source of variation from scores completely. If this variation is signal, such a removal would introduce bias. If this variation is bias/error, such a removal would improve the quality of scores.[5]

The clearest pattern across the uses shown in Fig. 2 is that rater error, especially systematic types of rater error, is often quite large. The ensuring equitable access use assumes scores are aggregated across ten teachers, each of whom is observed in two lessons with two separate raters scoring (but each rater observing separate lessons). It is commonly accepted that aggregating scores across this many lessons and raters will lead to reliable scores (e.g., Kane et al., 2012). However, the analyses presented here show that when accounting for rater accuracy rather than just rater agreement and when distinguishing systematic versus non-systematic rater errors, rater biases can still make up more than 10% of the variation in scores and random error driven by raters can still make up more than 15% of the variation in scores. To our knowledge, this is the first article to explore score quality in terms of rater accuracy rather than rater agreement. The nature of this rater error, then, has important implications for the validity of all interpretations of observation scores, a point that we take up in the discussion.

A second important result shown in Fig. 2 is that a modest portion of scores, in many, but not all uses, are composed of variability that could be either useful signal or some form of error. For example, in the statewide program to identify and retain teachers with high skills, the variability in scores from ObsSys that is associated with student characteristics and between-school differences could be either signal or bias. It is bias if teachers’ scores are dependent upon the school that they teach in or the students that they teach. It is signal if the associations are driven by teacher sorting (i.e., if there is no direct causal connection between students/schools and scores). This finding suggests that there is a real need to examine and study scores carefully to understand the degree to which each explanation fits, as this would be vital to determining the accuracy of interpretations of scores and help address questions of when to adjust scores to try to remove these sources of error/bias.

Last, the amount of signal and error in scores varies considerably across the different uses and across the two rubrics. The variation in scores that is signal ranges

[5] One could calculate the improvement in quality by readjusting percentages based on the amount of variation in scores removed. For example, 19% of the variation in scores is either Bias or Signal when CLASS scores are used for identifying teachers for professional development. Adjusting scores to remove this variation would remove this 19%, leaving 81% of the total variation in scores. Dividing the remaining percentages by 0.81 to bring the total variation back up to 100% would show the results for the adjusted scores under the assumption that the removed variation in scores was, in fact, error/bias.

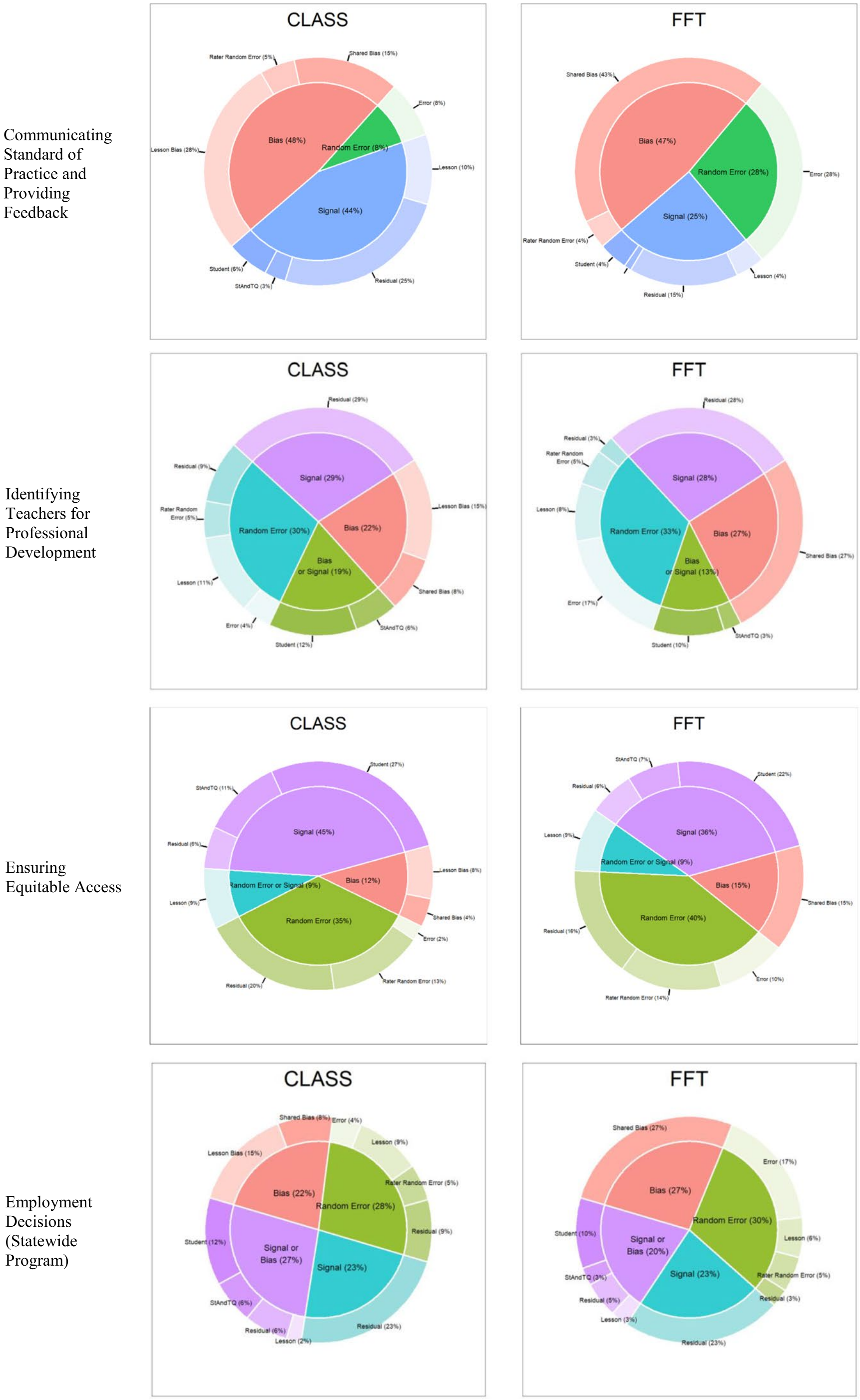

Fig. 2 Augmented GT theory results for each use of scores

from 23 to 54% for CLASS scores and 23 to 45% for FFT scores. There is more variation in the level of signal across the different interpretations of scores than across CLASS and FFT. Scores from ObsSys seem to be differentially valid based on different interpretations. The validity of any specific interpretation of scores must be carefully studied and explored if that interpretation is used to justify a score use. Preferably, this would be done with the data from the ObsSys being used through replicating the types of analyses conducted here using locally agreed-upon conceptualizations of teaching.

6 Discussion

This paper argued that uses of observation scores are often premised on interpreting observation scores as representing a specific construct (typically teaching or teacher quality), as seen through the rubric's lens. We argue for the need for users of ObsSys to (even more) explicitly define the intended interpretation of scores. This is because the validity of a use of observation scores is often justified (in part) through the interpretation of scores as representing a specific construct (i.e., interpretation for use; AERA, APA, & NCME, 2014) and different uses imply different interpretations of scores. This paper's results show that the quality of observation scores (i.e., the extent to which scores represent the intended construct) can vary widely across interpretations. In fact, we find more variation in the quality of scores across interpretations than across ObsSys. This does not downplay the importance of selecting an appropriate ObsSys but rather highlights the importance of interpretations.

The intended interpretation of scores determines whether the various contributors to scores are either signal, error, or bias, which in turn determines the quality of scores. This is an important point for both research and practice, as it remains common practice to calculate overall summaries of the reliability and concurrent/predictive validity of observation scores, evaluating the quality of scores based on these standard summaries (e.g., Kane et al., 2012; Panayioutou et al., 2021). The fallacy of relying on overall summaries of reliability and concurrent/predictive validity can be seen by examining differences in the quality of scores across this paper's proposed interpretations, through research highlighting the impact of specific errors and biases on estimates of score reliability/validity (e.g., van der Lans, 2018; White & Ronfeldt, 2022), and through modern validity theory's emphasis on considering specific uses of scores (e.g., AERA, APA, & NCME, 2014; Brennan, 2001).

6.1 Using observation scores based on the proposed interpretations

None of the proposed interpretations of scores resulted in high-quality representations of the intended constructs (i.e., signal/reliability over 70%/0.7 and bias near 0%). This is a more negative overall assessment of scores compared to past studies (e.g., Kane et al., 2012; Mantzicopoulos et al., 2018), but see Kelly et al., (2020).

Our more negative assessment is driven by the fact (1) that we explored rater accuracy rather than rater agreement (which unearthed previously hidden types of rater error) and (2) that we identified additional sources of error and bias in scores by considering the relative contributions of students, teachers, and the instructional context in light of a specific intended interpretation of scores. This study, then, arguably provides a more detailed and nuanced understanding of the quality of observation scores (for the proposed uses) than past studies.

While the generalizability of these empirical results should be examined (see discussion below), a strict interpretation of these findings would suggest that none of the suggested interpretations of scores is defensible. This does not, though, rule out using observation scores for the discussed (or other) uses but suggests the need to develop alternative interpretations of scores to support the proposed uses. Given the high levels of rater error, the most tempting interpretation would be to consider raters as providing their own, idiosyncratic judgement (i.e., reinterpreting scores as capturing raters' understanding of teaching quality rather than the ObsSys's understanding of teaching quality). This reinterprets rater error as capturing useful signal, transforming regions of rater error in Fig. 2 from error/bias to signal. This shift in the intended construct, though, would remove the comparability of scores across raters since such comparability is based on raters scoring the same construct. For example, for the use of "communicating a standard of practice and providing feedback to teachers," interpreting scores as capturing raters' unique perspectives implies that all of the variation in scores is signal but raises important questions about whether a clear standard of practice is being communicated since each rater has different understandings of this standard. In this way, through considering alternative interpretations of scores, readers can use Fig. 2 to reason about how the quality of scores shifts across interpretations. That said, readers are better off replicating analyses seen here in the local context of specific score uses, treating the results of this paper as a simple demonstration that observation score quality for specific interpretations of scores could be quite low.

Another way of shifting the interpretation of scores would be through more careful measurement. For example, when identifying teachers for a professional development on leading discussions, observers can focus on observing lessons where teachers intend to lead discussions, removing the variation in scores labelled "bias or signal" and improving the overall quality of score interpretations by controlling for the factor of teachers' choice in what to teach. The overall takeaway, then, is the importance of carefully considering the different sources of error and bias under a given interpretation of scores, which allows one to either plan for ways of controlling those sources of error and bias (e.g., through sampling lessons) or adjusting intended interpretations so that scores better represent the intended construct. Additionally, there is a need to replicate results in local settings.

6.2 Implications and meaning of rater error

This paper examined the extent to which observation scores can be interpreted as capturing the rubric-intended understanding of teaching quality. As noted previously, under this interpretation, deviations between raters' scores and the master scores are either error or bias. This accounts for most of the variation in observation scores. In some ways, this is surprising, as our focus on rater accuracy rather (than rater agreement) identified important new sources of rater error that were hidden in past studies. That is, all raters score calibration data wrong in the same way. The high level of shared bias could be due to challenges in rater training, complexities in scoring calibration videos, ambiguities in the rubrics, etc. Of particular interest is whether this same error is present in other data sources or whether this is a strange quirk of the UTQ data. We have run similar analyses on the calibration data for the Measures of Effective Teaching (MET; Kane et al., 2012) data set (White & Ronfeldt, 2022), which found the same shared bias in calibration data across the 100 + MET raters, though the size of this shared rater effect was smaller than in the UTQ data. Rater error is likely worse when ObsSys are used by school systems (e.g., teacher evaluation systems) because of a lack of training and monitoring of raters (Steinberg & Donaldson, 2016) and the multiple goals that administrators balance when assigning scores (Halverson et al., 2004), which likely leads scores to be less reflective of the intended vision of teaching quality (White & Ronfeldt, 2022).

For users of observation systems, the takeaway from the high observed level of rater error in this study and the high levels of rater error in past studies (e.g., Bell et al., 2014; Halverson et al., 2004) is to interpret scores as capturing raters' understanding of teaching quality (as provided through the observation rubric) rather than as capturing the rubric-intended understanding of teaching quality. This limits how observation scores can be used, as scores are no longer comparable across raters under this interpretation (i.e., this interpretation assumes that raters apply idiosyncratic standards when scoring), but it likely represents the only defensible interpretation of scores, at least unless/until one can provide strong evidence that raters score accurately. Justifications for using scores, under this new interpretation, would have to rely on the observer's ability to provide meaningful scores rather than how scores represent the rubric-intended understanding of teaching quality.

The apparent need to interpret observation scores as capturing only raters' judgements rather than the rubric-intended understanding of teaching quality raises important questions about the need for ObsSys, as ObsSys are thought to be useful based on their ability to provide a common framework and language for understanding teaching (i.e., measuring the intended construct is a key argument for the usefulness of ObsSys; e.g., Bell et al., 2019; Klette, 2023). Further, as we have argued, we see little justification for the high cost of adopting and implementing an ObsSys, except insofar as that ObsSys provide a useful lens for understanding teaching quality.

6.3 Statement on generalizability

As with every study, this study needs replication in other data sets, with other ObsSys, and in data from practice settings. The UTQ data comes from a research study that specifically hired and trained raters to provide scores (Casabianca et al, 2015). Most uses of scores from ObsSys come from practice-based applications of ObsSys. We can make some predictions as to the differences that would be expected when using scores from practice-based systems, namely rater error may be more problematic due to the weaker controls over rater error and scores may be less representative of the ObsSys's intended vision of teaching quality due to the multiple goals administrators balance when assigning scores (Halverson et al., 2004; Steinberg & Donaldson, 2016). This would suggest that our results present an overly optimistic picture of the support the proposed interpretations have.

Problematically, ObsSys in school settings often do not capture the sort of data necessary to replicate the analyses conducted here. Even the Network for Educator Effectiveness (Wind et al., 2018), which does more to explore the quality of scores than most practice-based settings, collects comparable scores from principals only outside of the school context, where principals are not balancing multiple goals when scoring. Such collection of data on score quality from non-typical scoring situations raises the risk that the conclusions about the quality of scores do not apply to the scores used in practice. Without a greater emphasis on collecting better data to explore the comparability of scores across settings and raters, the field is stuck extrapolating from research data like UTQ with little empirical evidence for the reasonableness of such extrapolation. Where possible results here should be replicated in local settings.

7 Conclusion

The results of this paper highlight the importance of being explicit about the intended interpretation of observation scores and how these interpretations support specific uses of scores. It is only after establishing an intended interpretation of scores that the quality of scores can be properly interpreted since many factors contribute to observation scores and each factor could be signal, error, or bias depending on the intended interpretation of scores. In fact, we found that the specific interpretation of scores was more important for evaluating score quality than the ObsSys that were used. Across all interpretations, though, the analyses presented here raise cautions about interpreting observation scores as capturing anything other than raters' idiosyncratic impressions of teaching quality. Comparisons of the research-based data source used in this study and practice-based ObsSys, which often lack rater monitoring mechanisms (Steinberg & Donaldson, 2016), suggest that interpretations of observation scores as capturing the rubric intended understanding of teaching quality are unlikely to be defensible. This implies that the validity argument supporting uses of scores may have to rely on interpretations of scores as reflecting raters' idiosyncratic impressions of teaching quality, though such conclusions should be replicated in practice-based settings. Importantly, these results should not be used to dismiss the use of ObsSys, as ObsSys could be justified through other interpretations of scores or the practical benefits of their use. However, the results here do suggest some healthy skepticism about claims that ObsSys accurately measure their intended understanding of teaching quality and validity arguments that rely (implicitly or explicitly) on such interpretations of scores.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11092-024-09427-8.

Author contribution Conceptualization, formal analysis and investigation, writing—original draft preparation: MW; writing—review and editing: all authors; funding acquisition: KK.

Funding Open access funding provided by University of Oslo (incl Oslo University Hospital) This project was funded by Nordforsk, grant number: 87663, and Norges forskingsråd, grant number: 300791.

Data availability The data used in this study is not currently publicly available but is in the process of being made available. The data can be made available upon reasonable request to the first author.

Declarations

Competing interests The authors declare no competing interests.

References

Abdi, H. (2007). Signal Detection Theory (SDT). *Encyclopedia of measurement and statistics* (pp. 886–889). SAGE Publications, Inc.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association. http://www.apa.org/science/programs/testing/standards.aspx

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 50–97). Jossey-Bass.

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. ggf5gq.

Bell, C. A., Jones, N., Lewis, J., Qi, Y., Kirui, D., Stickler, L., & Liu, S. (2015). *Understanding consequential assessment systems of teaching: Year 2 final report to Los Angeles Unified School District*. ETS. http://www.ets.org/Media/Research/pdf/RM-15-12.pdf.

Bohn, C. M., Roehrig, A. D., & Pressley, M. (2004). The first days of school in the classrooms of two more effective and four less effective primary-grades teachers. *The Elementary School Journal, 104*(4), 269–287.

Brennan, R. L. (2001). *Generalizability theory*. Springer, New York. gwqz.

Brophy, J. (1973). Stability of teacher effectiveness. *American Educational Research Journal, 10*, 245–252. https://doi.org/10.3102/00028312010003245

Brophy, J. E., & Good, T. L. (1984). *Teacher behavior and student achievement*. Michigan State University.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal., 55*(6), 1233–1267. gd32fh.

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

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

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/10.1016/j.stueduc.20210.100894

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119–142. b88jtw.

Cohen, J., Schuldt, L. C., Brown, L., & Grossman, P. (2016). Leveraging observation tools for instructional improvement: Exploring variability in uptake of ambitious instructional practices. *Teachers College Record, 118*(11), 1–36. jbjf.

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

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Association for Supervision & Curriculum Development.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267–297. https://doi.org/10.1002/pam.21818

Emmer, E., Evertson, C., & Brophy, J. (1979). Stability of teacher effects in junior high classrooms. *American Educational Research Journal, 16*, 71–75. https://doi.org/10.3102/00028312016001071

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. ghw7hb.

Halverson, R. R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. In W. K. Hoy & C. Miskel (Eds.), *Theory and Research in Educational Administration* (pp. 153–188). Information Age Publishing Inc.

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. f43nts.

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

Klafki, W. (2000). Didaktik analysis as the core of preparation. In I. Westbury, S. Hopmann, & K. Riquarts (Eds.), *Teaching as a reflective practice: The German Didaktik tradition* (pp. 139–159). Erlbaum.

Klette, K. (2023). Classroom observation as a means of understanding teaching quality: Towards a shared language of teaching? *Journal of Curriculum Studies, 55*(1), 49–62. https://doi.org/10.1080/00220272.2023.2172360

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.

Kraft, M. A., & Hill, H. C. (2020). Developing ambitious mathematics instruction through web-based coaching: A randomized field trial. *American Educational Research Journal, 57*(6), 2378–2414. gjkxzh.

Lockwood, J. R., & McCaffrey, D. (2012). *Reducing bias in teacher value-added estimates by accounting for test measurement error*. SREE.

Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: A generalizability study comparing the Framework For Teaching and the Classroom Assessment Scoring System. *Educational Assessment, 23*(1), 24–46. gqbn8n.

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation, 49*, 15–29. https://doi.org/10.1016/j.stueduc.2016.03.002

Milanowski, A. (2017). Lower performance evaluation practice ratings for teachers of disadvantaged students: Bias or Reflection of Reality? *AERA Open, 3*(1), 2332858416685550. gcgnwn.

OECD. (2020). Global teaching insights: A video study of teaching. *OECD Publishing*. https://doi.org/10.1787/20d6f36b-en

Panayioutou, A., Herbert, B., Sammons, P., & Kyriakides, L. (2021). Conceptualizing and exploring the quality of teaching using generic frameworks: A way forward. *Studies in Educational Evaluation, 70*(3), 101.

Phelps, G., Jones, N., Liu, S., & Kisa, Z. (2014). *Examining teacher, school, and program moderators in the context of teacher professional development studies [Paper Presentation]*. Washington, DC: 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.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*(4), 399–416. f6qvm8.

Praetorius, A.-K., Rogh, W., Bell, C., & Klieme, E. (2019). Methodological Challenges in conducting international research on teaching quality using standardized observations. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE Handbook of Comparative Studies in Education* (pp. 269–288). SAGE Publications. http://sk.sagepub.com/reference/sage-handbook-of-comparative-studies-in-education.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis Methods* (2nd ed.). SAGE Publications Inc.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 1–40. https://doi.org/10.1162/EDFP_a_00186

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

The New Teacher Project. (2018). *The opportunity myth: What students can show us about how school is letting them down—and how to fix it*. The New Teacher Project. https://tntp.org/publications/view/the-opportunity-myth.

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), 347–366. https://doi.org/10.1007/s11092-018-9285-5

White, M. (2023). *Accounting for Student Composition in Estimates of Teacher Quality from Classroom Observation Instruments*. University of Oslo.

White, M. (2022). What's in a score? Augmented decompositions of scores from observation systems. https://doi.org/10.31219/osf.io/f9vgz

White, M., & Ronfeldt, M. (2022). *Monitoring rater quality in observational systems: Issues due to unreliable estimates of rater quality*. University of Michigan.

White, M., Luoto, J., Klette, K., & Blikstad-Balas, M. (2022). Bringing the conceptualization and measurement of teaching into alignment. *Studies in Educational Evaluation, 75*, 101204. https://doi.org/10.1016/j.stueduc.2022.101204

Wind, S. A., Tsai, C.-L., Grajeda, S. B., & Bergin, C. (2018). Principals' use of rating scale categories in classroom observations for teacher evaluation. *School Effectiveness and School Improvement, 29*(3), 485–510. https://doi.org/10.1080/09243453.2018.1470989