Bringing the conceptualization and measurement of teaching into alignment

Mark White *, Jennifer Luoto, Kirsti Klette, Marte Blikstad-Balas
Department of Teacher Education and School Research, University of Oslo, Norway

ARTICLE INFO

Keywords:
Observation System
Classroom Observation
Measurement
Theory

ABSTRACT

There is a growing use of standardized observation systems to directly measure teaching quality in classrooms. These systems are based on conceptual understandings of teaching that lead to carefully operationalized rubrics that decompose teaching quality into a number of distinct dimensions. In this paper, we argue that measurement and analysis choices used by observation systems may not fully align with the conceptual understandings of teaching upon which observation systems are based. We discuss three key assumptions that undergird many views of teaching quality and highlight how common analytical approaches violate these assumptions, proposing alternative analytical approaches that would better conform to the conceptual understandings of teaching quality. We end with a discussion of the importance of carefully aligning conceptual understandings with measurement approaches.

1. Introduction

Teaching is a highly complex activity (Bell et al., 2012). This makes it quite challenging to measure in meaningful ways, but education research must take on this task of directly studying teaching to support its improvement (Ball & Forzani, 2007). An important tool for measuring the quality of teaching is standardized classroom observation systems (ObsSys; Bell et al., 2019). ObsSys take a community of practice’s view of teaching quality and package that understanding within a measurement structure that provides observation scores aligned to that view. The foundation of an ObsSys is the rubric, which operationalizes a community of practice’s view of teaching quality into a set of observable behaviors spread across a number of explicitly defined dimensions. The ObsSys also contains a set of procedures (e.g., rater training and monitoring, scoring guidelines, guidelines for sampling instructional practice) that facilitate the broad adoption and use of the ObsSys. Due to their systematic way of capturing facets of teaching and relative ease of use, ObsSys have been identified as a useful tool across a number of purposes, including examining the relationship between teaching and learning (e.g., White, 2018; Bacher-Hicks et al., 2019; Blazar et al., 2017; Lynch et al., 2017; OECD, 2020; Reddy et al., 2019); comparing enacted instruction across settings (e.g., Martinez et al., 2016; Maulana et al., 2019, 2020; OECD, 2020; Praetorius et al., 2019); and providing feedback to teachers (e.g., Cohen et al., 2016; Kimball, 2002; Kraft & Hill, 2020; Muijs et al., 2018; Steinberg & Sartain, 2015; van der Lans et al., 2018).

ObsSys exist at the boundary of conceptualizations of teaching, enacted teaching practice, and measurement, with the potential to contribute to each. Fig. 1 shows a representation of how we think of this. Observation systems are, as argued, operationalizations of a community’s understanding of teaching quality. The measurement process (including sampling and conducting observations, scoring, and analyzing data) is the meeting of the community’s understanding of teaching, as captured by the observation system, and enacted practice. The measurement process results in rubric-aligned empirical evidence that supports refinements to conceptual understandings of teaching and rubric-aligned feedback that supports changes to teachers’ understandings of teaching. This paper argues for the need to better align conceptual understandings of teaching with measurement approaches so that the data generated by measuring enacted practice supports refinements to conceptual understandings of teaching and teacher feedback aligned to this conceptual understandings of teaching (Charalambous et al., 2021; Hoelscher, 2017; Kelcey, 2015; Shavelson & Towne, 2002; Stein et al., 2017).

Abbreviations: ObsSys, Observation System.

* Correspondence to: Postboks 1099, Blindern, 0317 Oslo.
E-mail address: mark.white@ils.uio.no (M. White).
OrcidID: https://orcid.org/0000-0003-2394-3151.

https://doi.org/10.1016/j.stueduc.2022.101204
Received 8 February 2022; Received in revised form 13 May 2022; Accepted 18 August 2022
Available online 25 August 2022
0191-491X/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
To explore the alignment of conceptualizations of teaching quality and measurement, this paper discusses whether common measurement procedures align with three assumptions that commonly characterize teaching: (1) Teaching is a contingent activity; (2) Teaching takes place within a broader institutional and cultural context; and (3) Teaching requires attending to multiple levels. We hope to push ObsSys developers to more clearly lay out the conceptual underpinnings of their ObsSys and to provide explicit guidance on data collection, scoring, and analysis approaches consistent with that vision. This alignment will move the field beyond tests of empirical relationships (e.g., correlations) and towards developing empirically supported conceptualizations of teaching quality (Charalambous & Praetorius, 2020; Stein et al., 2017). We believe that stronger, coherent connections between conceptualizations of teaching quality and its measurement can lead to (a) joint refinement of both conceptions of instruction and methods for measuring instruction, (b) stronger connections between observation scores and measures of student learning and thus (c) clearer instructional guidance for teachers (Hoelscher, 2017; Kelcey, 2015; Stein et al., 2017).

2. Uses of observation scores

In the following discussion, we try to emphasize the practical implications of our rather technical points. To facilitate this, we focus on two common uses of ObsSys, examining the implications of each use of ObsSys: (1) understanding the relationship between teaching and learning, and (2) providing feedback to teachers. These two uses capture both the research side (use 1) and the practice side (use 2) of our framework highlighting the importance of aligned measurement. Note that we do not consider using ObsSys for teacher evaluation and do not believe sufficient knowledge of observation scores exists to support this use (c.f., Rowan & Raudenbush, 2016).

In discussing efforts to understand the relationship between teaching and learning (Use 1), we emphasize the challenges inherent in aggregating observation scores to generate a summary estimate of teaching quality that is on the same time-scale as the measure of student learning. In discussing providing feedback to teachers (Use 2), we emphasize that feedback should be as detailed, specific, and actionable as possible (Hill & Grossman, 2013; Wylie, 2020). Specifically, we focus on feedback about individual lessons1 and three aspects of feedback: (a) helping teachers calibrate their self-perceptions to an external standard (Kraft & Hill, 2020); (b) providing specific and actionable data (Cohen & Grossman, 2016); and (c) helping teachers see the ObsSys as a conceptual tool that supports thinking about teaching rather than rules to follow (Cohen et al., 2016). We do not, though, focus on teachers’ actual use and interpretation of feedback as provided by the ObsSys, which is arguably the most important aspect of using observation scores (O’Leary et al., 2017). Rather, we focus only on intended interpretations and uses of feedback, as actual uptake of feedback is an empirical question and so beyond the scope of this conceptual paper.

3. Assumptions about teaching

In this section, we discuss three assumptions that commonly characterize views of teaching. For each assumption, we highlight its measurement implications, pointing out how common measurement practices implicitly violate the assumption and pointing towards alternative, aligned practices. Our goal is to avoid telling the readers what to believe about the nature of teaching, but to provide a discussion of the implications of different assumptions such that readers can take advice when it aligns with their conceptualizations and ignore advice unaligned to their conceptualizations.

3.1. Assumption 1 Teaching is a contingent activity

The assumption of contingency is the belief that what happens in the classroom and/or the effectiveness of specific instructional actions could be dependent on the instructional context (Clarke, 2013; Kelly et al., 2020). We define the instructional context as the features of a lesson that form a setting within which instruction occurs, including the learning goals, the content taught, student tasks, grouping structures, and other relevant features. These context features often vary within lessons.

An example of the potential contingency of teaching can help clarify this distinction. Consider a sequence of instruction that focuses on students engaging in independent practicing of rote tasks (e.g., practicing adding numbers). Under the assumption that teaching is a contingent activity, dimensions of instruction that are often measured and understood to be important to student learning (e.g., student interactions, cognitive activation) may not meaningfully characterize teaching quality during such practicing. Rather, alternative dimensions of instruction (e.g., silent engagement, the timing of such practicing) might be more useful characterizations of teaching quality during the practicing of rote tasks. The measurement challenge posed by teaching being a contingent activity is that the instructional interactions captured by ObsSys, which are those interactions thought to typically support student learning, are not necessarily the instructional interactions that support student learning in every single observed lesson. That is, ObsSys may not capture how instruction changes across instructional contexts (White, 2018). We can reframe this challenge using statistical language: instructional contexts may both (a) moderate the level of observed teaching quality and (b) moderate the relationship between observation scores and student learning and development.

Rather than thinking about the contingency of teaching as all or nothing, it is better to consider different degrees of contingency. If teaching is thought to be highly contingent, then different instructional contexts will need specifically developed ObsSys to measure teaching quality in that context. For example, the Instructional Quality Assessment (IQA; Matsumura et al., 2008) measures the quality of mathematical discussions about assignments. If teaching were highly contingent, this narrow focus would be necessary to properly measure teaching quality within mathematical discussions. If teaching is thought to be moderately contingent, then the same ObsSys could usefully capture teaching quality across instructional contexts, but rubrics designed to measure narrow contexts would provide useful supplements to broader ObsSys (Kelly et al., 2020). If teaching is thought to be non-contingent, this supplementing of ObsSys would not be necessary. Importantly, the contingency of teaching may depend heavily on the aspect of teaching quality that is measured. Aspects related to culture and climate (i.e., classroom-level constructs) may be less contingent than dimensions more directly related to instructional practices (see the discussion under Assumption 3).

Common measurement practices make exploring the contingency of teaching quality on instructional contexts difficult since the instructional context is often not systematically measured or examined.

---

1 Unlike in some past work that emphasized providing reliable feedback about teachers’ average practice (e.g., Taut & Rakocy, 2016; van der Lans et al., 2018), we focus on feedback provided immediately after a lesson that focuses on the quality of that specific lesson (c.f. White et al., 2021).
Additionally, many ObsSys divide instruction into equal-interval segments (e.g., 15 min) and then score each segment. The segment then becomes the fundamental unit of analysis, but segments do not align with instructional contexts (e.g., the context might shift in the middle of a segment). This misaligns the unit of analysis (i.e., the segment) and the potential moderator (i.e., the context), making exploring moderation difficult. The use of equal-interval segments, rather, helps to simplify estimating mean teaching quality. However, when teaching is contingent, the alignment of teaching and the instructional context could be far more important than the average teaching quality score. Alignment is statistically independent of the average score—making the average score a poor estimate of teaching quality. Then, stronger one’s belief in the contingency of teaching, the less useful is scoring equal-interval segments and/or estimating average teaching quality, unless an ObsSys measures teaching quality in only a narrowly defined instructional context.

When teaching is viewed as contingent, we propose the following measurement approach to align measuring teaching quality with the instructional context, supporting explorations of moderation. First, important aspects of the instructional context should be measured (e.g., instructional goals, content taught, assigned tasks etc.). This requires developing stronger knowledge and theories about how the instructional context moderates teaching quality, such that the right aspects of the context are measured. We know of little work that would provide a strong empirical basis for knowing which context features are most important. Second, lessons would be segmented based on the instructional context. Due to space, we cannot delve into how this would be done, but we do note here that this has been successfully done in the past (c.f., Andrews, 2007; Burns & Anderson, 1987; or Clarke et al., 2007). Third, analyses would explicitly use the instructional context as a moderator variable or separately conduct analyses across contexts. This approach allows for the statistical (or qualitative) exploration of how the instructional context moderates the impact of specific instructional practices, respecting the assumption that teaching is contingent upon the instructional context. This alternative approach, of course, comes with its own challenges. Namely, it may be very difficult to reliably define segments that align with instructional contexts (Luoto, Klette, & Blikstad-Balas, 2022).

3.1.1. Implications for understanding the relationship between teaching and learning

The contingency of teaching has important implications for linking teaching and learning. Namely, the two types of moderation mentioned above pose a threat to valid conclusions about the relationships between teaching and learning, as ignoring moderating effects leads to inaccurate estimates (Fairchild & MacKinnon, 2009). The first type of moderation involves the instructional context moderating the observed level of teaching quality, conflating instructional context and teaching quality and making it hard to differentiate the effect of a dimension of teaching quality as separate from the effect of the instructional context. For example, our experience is that some efforts to measure classroom discourse are highly conflated with activity format since almost all group or pair work involves student-led talk and very little whole class or individual work involves student-led talk. Then, any estimated effect of classroom discourse on an outcome could, in fact, be the result of the activity format (e.g., students learn better in small group instruction), unless the conflation of classroom discourse and activity format is addressed above supports such efforts to control for the instructional context.

The second type of moderation is when the instructional context moderates the relationship between teaching and learning. In this scenario, common practices of estimating average levels of teaching quality across a period of time (e.g., Kane et al., 2012, OECD, 2020) provide misleading results because they ignore the moderation (Fairchild & MacKinnon, 2009). The proposed alternative of aligning segments with the instructional context allows for moderation analyses. We believe that such explorations of the moderating effect of the instructional context has the potential to address concerns about the weak linkage between measured teaching quality using ObsSys and student learning (e.g., Kelly et al., 2020).

3.1.2. Implications for providing teacher feedback

When teaching is contingent, the instructional context must always be attended to. When segments are created based on the instructional context, the resulting scores can provide useful feedback on what instruction looks like across different contexts. For example, a teacher might receive separate scores characterizing how they introduced a topic and how they supported students in doing group work. This could allow, in principle, for deeper conversations and feedback to teachers that focus on how teachers can shape features of the instructional contexts (e.g., how long should the teacher spend introducing the topic), how that context in turn affects instructional interactions (e.g., how did working in groups help students learn), and how instructional interactions support student outcomes (e.g., what should group work look like). In essence, this sort of feedback emphasizes that teachers must both mold the instructional context to support learning and manage the interactions between themselves, students, and the content. The result could be more productive feedback discussions that help teachers see themselves as actively structuring lessons to meet specific goals (Cohen et al., 2016).

3.2. Assumption 2 Teaching takes place within broader institutional and cultural contexts

Assumption 2 is the assumption that instruction occurs within school systems and cultures, which provide a broader context that influences instruction (Heibert & Grouws, 2007). Note that the previous section focused on the instructional context, or the context around specific instructional interactions. Here, we focus on the lesson context, which is how the lesson is situated within the broader instructional program that exists within a school system and unfolds across the full school year. There is overlap between the first two assumptions. Namely, key points in both sections deal with potential moderators. However, only one of the two assumptions might fit one’s conceptual understanding of teaching, necessitating the discussion of each separately. The main difference between the assumptions is the grain size. Assumption 2 is about how individual lessons fit within the broader socio-cultural and institutional climates while Assumption 1 is a finer-grained assumption about the nature of instructional interactions.

If teaching occurs within broader institutional and cultural contexts, individual lessons cannot necessarily be understood in isolation. This has been most clearly argued by Cohen and colleagues (2020). Cohen and colleagues (2020) examined changes in teaching practices resulting from the Common Core State Standards, concluding that the changes called for by the new standards could not be understood or studied using the standard sampling approach of observing single lessons, spaced across time. For example, the standards call for an equal focus on procedural learning, conceptual learning, and application. Determining if there is an equal focus on these three areas requires examining how individual concepts are developed across a series of lesson, as each individual lesson need not necessarily focus on all three aspects. The general point here is that understanding teaching quality within a
lesson, if Assumption 2 holds, requires understanding how an individual lesson fits within the broader instructional program as it unfolds within the nested institutional and cultural contexts in which the lesson is enacted (i.e., the lesson context).

There are two very concrete implications of the broader point just made. First, the lesson context can make some dimensions of teaching quality more or less relevant (i.e., moderation). For example, it has been argued that the first few days of the school year are especially important for establishing clear classroom routines and cultures that can then drive positive academic experiences across the rest of the school year (e.g., Bohn et al., 2004; Emmer et al., 1980). Similarly, differences in curricular foci or cultural norms can make specific dimensions of teaching quality more or less relevant in specific cultural contexts (Xu and Clarke, 2019) or at different points in the curricula. Second, the lesson context can make instruction more or less representative of teachers’ general practice at specific times of year. For example, a quasi-experiment in the USA has demonstrated that instruction just prior to standardized testing, when testing is highly salient, differs quite markedly from instruction after standardized testing, when testing is not salient (Plank & Condliffe, 2013). The representativeness of individual lessons is an important consideration when trying to generalize beyond the observed lessons to understand a teacher’s broader instructional practice.

3.2.1. Implications for understanding the relationship between teaching and learning

Estimating (average) teaching quality on the same time scale that student learning is measured is a necessary step for empirically comparing teaching and learning. The main implication of Assumption 2 is that this estimate can be less useful and more difficult than imagined. It can be less useful because it implies that instruction should be diverse and varied. This is the same argument for moderation as under Assumption 1, but focused on broader features of the lesson context. As in Assumption 1, the alignment between the lesson context and teaching quality may be more important than the average level of teaching quality. For example, if establishing a strong climate at the start of the year is vital (e.g., Bohn et al., 2004), then the observed quality in climate-related dimensions is far more important at the start of the year than later in the year (i.e., time of year moderates the relationship between climate-related teaching quality and student learning). We refer readers to the discussion under Assumption 1 for more details.

A second important implication is that estimating average teaching quality overall a period of time can be more difficult due to the exchangeability assumption in generalizability theory (Brennan, 2001). A number of studies have examined how many lessons are needed to generate reliable estimates of teaching quality across a period of time (e.g., Hill et al., 2012; Praetorius et al., 2014). These studies treat lessons (and segments, raters, etc.) as facets that drive variability in observed teaching quality. With few exceptions (e.g., White, 2017), generalizability theory studies have ignored the lesson context (e.g., time of year, content, day of the week, instructional goals), which, in effect, treats the lesson context as a “hidden facet”. The existence of hidden facets can lead to problematically high or low estimates of the reliability across lessons (Brennan, 2001). When the “hidden facets” of lesson context are not taken into consideration, the analyst assumes that all lessons are equal (on average) representations of teaching (i.e., lessons are exchangeable) in the time period one is generalizing across. This effectively assumes the systematic variation in teaching across lesson contexts is ignorable. The net effect is that estimates of average teaching quality may be far less reliable than standard generalizability theory studies suggest (White, 2017), which is problematic given the already low reliabilities of estimates of average teaching quality. The size of this potential bias should be examined empirically. For example, if a study collected observations across a large number of lessons, simulations using resampling could then simulate the actual reliability of common sampling plans based on estimates of their reliability.

There are a number of ways to improve estimates of average teaching quality. The first is to reduce the lesson contexts across which one generalizes. This usually occurs by restricting the study to focus on learning specific content (e.g., Lipowski et al., 2009; OECD, 2020), limiting the lesson context across which one generalizes. Second, randomly sampling lessons will ensure the observed lessons (on average) are representative of the broader time period and that the hidden facets of the lesson context do not bias estimates of average teaching, but is difficult to accomplish. Due to the small number of days observed for each teacher, one should generally stratify on context features before random sampling to improve efficiency.

Of course, neither of these solutions get at the more fundamental problem that lessons, under Assumption 2, may each play a unique role in supporting achieving broader curricular and institutional goals. This would imply that the average level of teaching quality is not a useful characterization of teaching in a classroom. Novel analysis approaches that treat sequences of lessons as the unit of analysis, rather than individual lessons or segments within lessons, may be necessary to fully address the implications of Assumption 2 (see Cohen et al., 2020).

3.2.2. Implications for providing teacher feedback

The implication for teacher feedback is that feedback should consider the broader institutional and cultural practices. Two specific examples of this general principle pointed out above (i.e., considering the role of time of the year and institutional practices such as standardized testing). Feedback that is aware of how the observed lesson fits within the broader curricular and institutional context may require observers to observe multiple consecutive days or have deep knowledge of local contexts, which places restrictions on who provides feedback and how observations meant for formative assessment are organized. For example, external coaches without knowledge of the local context will likely be unable the broader context within which the lesson is situated, so may not be as useful sources of feedback under this assumption (though they may have other benefits).

3.3. Assumption 3 Teaching requires attending to multiple levels (lesson, classroom, etc)

Instructional effectiveness occurs at multiple levels (Creemers & Kyriakides, 2006). The assumption discussed here, then, is that teachers must simultaneously attend to more stable features of the classroom (e.g., routines; climate), features of lessons (e.g., overall lesson goals), and aspects of the specific instructional interactions (e.g., responding to questions). The implication of this assumption is that ObsSys can (or maybe should) measure teaching quality at different levels (i.e., measure dimensions that are conceptualized as functioning at different levels). A number of levels are possible: segment-level, lesson-level,

---

3 Stated more technically, the standard generalizability theory assumption is that instances of facets are exchangeable. Applied to sampling in ObsSys, this would imply that the sampling procedures ensure that any sample that can be achieved is equivalent (in expectation) to every other sample. When the lesson context impacts instruction, possible samples may be non-exchangeable in that some features of the lesson context are systematically under or over sampled. This is especially true when teachers choose lessons to be observed, a common feature of sampling plans.
classroom-level, or teacher-level (and possibly others). The level at which a dimension captures instruction has important implications for how it is analyzed.

The distinction being made here is equivalent to the more common distinction between traits and states in psychological measurement. Traits (e.g., personality) are hypothesized as stable features of a person. For example, an introverted person is always introverted, even if they act in more or less introverted ways across time. The very construct of introversion assumes stability (at least over modest time horizons). The assumption of stability in traits does not preclude measuring them repeatedly across time (e.g., through daily journal entries of a person’s social interactions) or at different grain sizes (e.g., measuring the amount of time spent along versus measuring reactions to specific scenarios). On the other hand, states (e.g., a happy mood) are hypothesized as variable constructs. Someone who is currently happy might not be happy in an hour. It is possible to aggregate a state variable across time (e.g., estimating one’s average level of happiness), but this requires conceptualizing the meaning of the aggregated construct. For example, “average happiness” is a fundamentally different construct than the state of happiness, which is defined by a specific brain state. Then, “average happiness” needs its own conceptualization as a construct.

If teaching requires attending to multiple levels, then ObsSys should also measure “traits” and “states” of instructional quality. Here, traits are stable classroom-level features of instruction, such as the classroom climate and the presence of routines. Take classroom climate as an example. The climate is always present and always influences classroom behavior. If two consecutive lessons are measured as having different levels of a climate, this does not mean that climate has changed across lessons, as the very construct of climate assumes stability that precludes rapid change (Marsh et al., 2012). Rather, differences in measured climate across consecutive days are, by definition, measurement error, possibly due to the climate manifesting in observable behaviors differently across days. State measures, on the other hand, can vary greatly in their presence within- or across-lessons. They are thus best considered as segment-level constructs and often capture specific instructional interactions, such as scaffolding techniques, discussion, and ways of questioning. For state measures, differences in scores across consecutive segments or lessons could be true changes in the construct and/or measurement error (i.e., a teacher could provide scaffolding in one segment, but not the next).

Importantly, the analyses appropriate for classroom-level (i.e., traits) and segment-level (i.e., states) dimensions differ, as do interpretations of analyses. Since classroom-level dimensions are assumed to be stable features of classrooms, high levels of within-classroom variation are potentially problematic. For example, the emotional support domain of the classroom assessment scoring system (CLASS) captures whether the classroom provides a safe and predictable environment that supports student engagement and exploration (Hamre et al., 2013). If the level of emotional support experienced by the students fluctuated on a daily basis, the theory of action for how the emotional support domain supports students would be undermined. Similarly, a single day of high emotional support would not lead to the purported benefits of emotional support since a single day does not lead to a safe and predictable environment. Rather, high levels must be sustained over time before students will feel safe and secure and so benefit from emotional support. Said differently, emotional support is a feature of classrooms, not a feature of individual interactions, though individual interactions can contain indicators that provide evidence for the classroom-level construct of emotional support. Then, statistical models that examine emotional support should focus on estimating a single classroom-level score for emotional support while examining whether the emotional support received by students remains relatively constant. Other classroom-level constructs are the same, the focus should be on estimating a classroom level score, treating within-classroom variation as error. There are no strong discussions of these points with classroom observations, but Marsh and colleagues (2012) discuss these issues for survey measures of classroom climate that aggregate across students (rather than segments).

The opposite is generally true for segment-level scores. Variability in segment-level constructs across days and segments could be positive (see discussions of the moderating effect of the instructional context under Assumption 1), at least if this variability is systematically related to the instructional or lesson contexts. For example, it would generally be a sign of high-quality instruction if the level of scaffolding varied across segments, remaining high when the difficulty of content is high and low when the difficulty of content is low. Note the connection to Assumption 1 here, segment-level dimensions (i.e., state dimensions) are those dimensions where alignment with the instructional context matters (like the scaffolding example just raised).

We encourage researchers to engage with the potentially productive idea of studying segment-level constructs at the segment level, rather than rushing to estimate teacher or classroom level scores. As discussed under Assumption 1, if a segment-level dimension is contingent on the instructional context, then alignment of the dimension to the context could be more important to teaching quality than the average level, such as in the scaffolding example just discussed. When there is the understandable desire to aggregate segment-level constructs to a higher level, analysts should carefully conceptualize the nature and usefulness of this aggregation. For example, it is not immediately clear why it matters whether a classroom has high average levels of scaffolding. After all, high average levels of scaffolding does not imply that scaffolding exists when students need it to support their learning nor does it suggest that scaffolding has been faded to support students’ independent mastery of complex content. Through carefully measuring the instructional context, one could generate a classroom level aggregate of the level of scaffolding that existed when the content to be learned is such that scaffolding would be useful, a much more useful metric of classroom quality than the overall average level of scaffolding.

3.3.1. Implications for understanding the relationship between teaching and learning

Typical studies that seek to understand the relationship of teaching and learning through linking measured teaching quality to measured student learning are easier using classroom-level dimensions, as such dimensions are more stable across time and are conceptualized at the same level of analyses as the student learning outcome. Similarly, these dimensions (especially management and climate) are often viewed as enabling factors that must be in place to facilitate the instructional interactions captured by segment-level dimensions (Cohen, Schultd, Brown, & Grossman, 2016; Luoto, Klette, & Blikstad-Balas, 2022). In combination, these factors would tend to lead these classroom-level dimensions to have higher correlations with student learning (e.g., Gil et al., 2016). These higher correlations are somewhat misleading, as they do not account for the complex ways that domains affect each other (e.g., moderation) or measurement error. Only by positing a full theory (i.e., including moderated and mediated effects) and testing that theory through targeted models can one get a true sense of which dimensions are most important for student learning and development.

Using segment-level dimensions in efforts to link teaching and learning leads to both more complications and more opportunities. The complications, as discussed above, stem from the higher within-classroom variability, which leads to lower levels of reliability when aggregating to the classroom level and the need to separately conceptualize segment-level constructs at both the segment level and the classroom level. The opportunities come from the ability to look at the alignment of segment-level dimensions with features of the instructional context (e.g., alignment of scaffolding with student struggle) and to generate novel approaches to examining the impact of these dimensions on learning. This requires novel research designs that are more aligned with conceptual understandings of segment-level scores. For example, a study could compare daily fluctuations in learning to daily fluctuations in the level of scaffolding (or other segment-level dimensions), generating more targeted evidence about how scaffolding supports student learning.
learning. Such analyses would violate the assume nature of classroom-level dimensions.

3.3.2. Implications for providing teacher feedback

Classroom-level and segment-level dimensions have different implications for teacher feedback. Recall here that we focus on providing feedback to teachers immediately after a lesson based on that single lesson (as opposed to estimating average teaching quality). Providing feedback on segment-level dimensions is simpler, as the entire set of instructional interactions that determine the level of teaching quality are observed. For example, the rater directly observes the level of, type of, and timing of instructional scaffolding provided by the teacher. They can then provide feedback directly on each of these aspects. Since the goal is providing feedback on the level of scaffolding within the given lesson, high within-classroom variability is not problematic here.

At the same time, providing feedback on classroom-level dimensions is more challenging. A single lesson provides at most a single snapshot of a classroom-level dimension. For example, observers do not directly observe the classroom climate, but must make inferences about the climate based on observed interactions. Those observed interactions might be more or less representative of typical interactions and the classroom climate. Providing teachers feedback, then, on the classroom climate requires making an inference about the climate from the brief snapshot. For this reason, it might be more reasonable to not provide feedback to teachers on classroom-level dimensions based on a single lesson, but to provide feedback only after aggregating scores across time. At the very least, this limitation of feedback on classroom-level dimensions should be made clear to teachers.

4. Discussion

We have argued for the importance of aligning the conceptual understandings of teaching quality that undergird ObsSys with measurement choices made when measuring teaching quality. This alignment will help improve measures of teaching quality and support the testing and validation of conceptualizations of teaching quality. Further, two specific uses of ObsSys can be improved through aligning conceptualizations and measurement. First, the alignment of conceptualizations and measurement can hopefully lead to stronger empirical relationships between teaching and learning, as these relationships have been disappointingly low and inconsistent in past work (e.g., Luoto, Klette, & Blikstad-Balas, 2022; Kane et al., 2012; OECD, 2020). At the very least, not considering the instructional and lesson contexts, as well as the confusion between classroom-level and segment-level dimensions, represent potential causes of this weak relationship. Second, the alignment of conceptualizations and measurement can improve the feedback provided to teachers by ensuring that teacher feedback aligns with conceptual understandings of teaching quality, which should provide an additional layer of assurance that the feedback is valid, especially as our understanding of teaching quality deepens over time.

Achieving the alignment of conceptual understandings and measurement requires ObsSys developers to more fully and completely develop the conceptual understandings of teaching that serve as the basis for ObsSys. This needs to be done with an eye towards measurement implications so that researchers using ObsSys have guidance in terms of how to collect, code, and analyze observational data. We have highlighted three common assumptions that characterize understandings of teaching and showed how some common measurement approaches implicitly violate these assumptions. In doing so, we hope to make researchers more aware of the implicit assumptions that standard approaches to collecting and analyzing data make, especially where these implicit assumptions might violate researchers’ conceptualizations of teaching quality.

More fine-tuned ways of using observation systems, which we suggest here, could also increase the relevance of observation systems for understanding equity issues related to teaching quality. We know that observation ratings vary across teacher and student characteristics (Campbell & Ronfeldt, 2018), and more fine-grained and context-close understanding of quality teaching may illuminate reasons for differences in relationships between teaching and learning for different teacher/student groups.

4.1. Towards programmatic approaches of studying teaching quality

Despite our critiques of the way ObsSys are used in the literature, we wish to emphasize here that we see an important role for such systems in the study of teaching and learning. ObsSys allow systematic inquiries into teaching quality that support the accumulation of research knowledge across time that is necessary to test and validate conceptualizations of teaching quality (Klette, 2020). A key takeaway from this paper is that this accumulation of empirical evidence to test and validate conceptualizations of teaching quality does not happen automatically with the use of ObsSys. Rather, researchers must collect and analyze data from ObsSys in ways that are aligned with the conceptual understanding of teaching quality contained in the ObsSys. This requires ObsSys developers to provide more clear guidance around appropriate approaches to the collection and analysis of observation scores and to align that guidance with the conceptual understandings underlying their ObsSys. It also requires researchers using ObsSys to adopt these recommended approaches, which will lead to more standardization in the analysis of observation scores. Such standardization is needed to accumulate empirical evidence on the conceptualizations of teaching quality captured by an ObsSys.

Our push for programmatic research (Klette, 2020; Grossman & McDonald, 2008) does not mean that all research should be restricted to the standard, common models that we argue ObsSys developers should promote for use with their protocols. However, deviations from standard models should explicitly be labeled as such and should emphasize comparing the novel model to the standard approach and highlighting the conceptual, empirical, or practical benefits of the proposed alternative conceptualization. A good guide for this suggestion is research on the Classroom Assessment and Scoring System (CLASS; Piianta et al., 2008). CLASS clearly proposes a factor analytic model that reflects their conceptual understanding of teaching quality, but much research has explicitly compared this standard model to alternatives (e.g., Hafen et al., 2014). Though such comparisons would be stronger if they routinely emphasized the practical implications of different models, rather than focusing on model fit statistics. For CLASS, demonstrating the practical implications of models might take the form of demonstrating (and replicating evidence for) divergent validity across CLASS domain scores in order to show the benefit of measuring three domains instead of a single responsive teaching domain—though note our assumption here that there will be domain-specific effects may violate CLASS theory (Downer et al., 2010). Note that this type of systematic comparison of proposed new models to standard models would be highly supported by the archiving of data from ObsSys in order to support researchers in going back to re-examine older data using newly proposed models.

4.2. Consequences for the measurement of teaching quality

An important implication of the points made in this paper is that different ObsSys need to collect data in different ways and be analyzed with different approaches depending on the conceptual views of teaching quality underpinning the ObsSys. In fact, many ObsSys could propose different models for different dimensions within the same ObsSys, especially if constructs are defined at different levels. For example, an ObsSys might posit the importance of classroom-level constructs like the classroom climate and routines, as well as segment-level constructs like instructional scaffolding and student talk (see e.g., the Protocol for Language Arts Teaching Observation; Grossman et al., 2013). This will increase the complexity of analyzing data from
ObSsys, but is necessary to align the analyses with underlying conceptions of teaching quality. Again, ObSsys developers should provide clear guidance to help clarify underlying conceptualizations and guide others in appropriately analyzing data from their systems.

Funding

Project was funded by Nordforsk, grant number: 87663 and Research Council of Norway, grant number: 300791.

CRediT authorship contribution statement

All authors contributed to the Conceptualization and writing of this work.

Declarations of interest

None.

References


Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. Learning and Instruction, 46, 45–60. https://doi.org/10.1016/j.learninstruc.2016.08.003