Bringing the Theory and Measurement of Teaching into Alignment

Mark White^{1*}, Jennifer Luoto, Kirsti Klette, & Marte Blikstad-Balas

¹ Department of Teacher Education and School Research, University of Oslo

* Corresponding Author

Author Note

* Correspondence concerning this article should be addressed to Mark White,
Postboks 1099, Blindern, 0317 Oslo. Email: mark.white@ils.uio.no. Phone: +47 967 47 424.
OrcidID: <https://orcid.org/0000-0003-2394-3151>.

Abstract

There is a growing use of standardized observation systems to directly measure teaching quality in classrooms. These systems are based on rich, theoretical understandings of teaching quality that lead to carefully operationalized rubrics that decompose teaching quality into a number of distinct dimensions. In this paper, we argue that many of the ways observation systems are applied and analyzed do not fully conform to the theoretical understandings of teaching quality upon which the observation system is based. We further argue for the need to align theoretical understandings of teaching quality with analytical approaches to measure teaching quality. We discuss four 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 theoretical understandings of teaching quality. We end with a discussion of the importance of carefully aligning theoretical understandings with measurement approaches.

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 (Ball & Forzani, 2007). An important tool for measuring the quality of teaching is standardized classroom observation systems (henceforth, Observation Systems; Bell, Dobbelaer, Klette, & Visscher, 2019). Observation Systems take a community of practice's view of teaching quality and package that understanding within a measurement structure that provides observation scores that should align to that theoretical view. The foundation of an observation system is the rubric, which transforms the community of practice's view of teaching quality into a set of observable behaviors spread across a number of explicitly defined dimensions. The observation system 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 observation system. This allows observation systems to be broadly adopted and to codify enacted instruction into a set of scores that, in principle, represents the relevant community's understanding of teaching quality (Klette & Blikstad-Balas, 2018). This makes observation systems a useful tool across a range of purposes, including examining the relationship between teaching and learning (e.g., Bacher-Hicks, Chin, Kane, & Staiger, 2019; Blazar, Braslow, Charalambous, & Hill, 2017; Klette, Roe, & Blikstad-Balas, 2021; Lynch, Chin, & Blazar, 2017; OECD, 2020; Reddy et al., 2019); comparing enacted instruction across settings (e.g., Martinez, Taut, & Schaaf, 2016; Maulana et al., 2019, 2020; OECD, 2020; Praetorius, Rogh, Bell, & Klieme, 2019); and providing feedback to teachers (e.g., Cohen, Schuldt, Brown, & Grossman, 2016; Kimball, 2002; Kraft & Hill, 2020; Muijs et al., 2018; Steinberg & Sartain, 2015; van der Lans, van de Grift, & van Veen, 2018).

Observation systems exist at the boundary of theory, enacted practice, and

measurement, with the potential to contribute to each. This paper considers how well observation systems straddle these three realms, emphasizing theory and measurement. By theory here, we refer narrowly to the community of practice's principled understanding of teaching quality that serves as the basis for the creation of the observation rubric and the observation system, which may draw from a range of broader theoretical traditions. Enacted practice refers to the actual enactment of instruction. Measurement, in turn, refers to the procedure of applying the theory to enacted practice through using the observation system to generate scores that reflect the theoretical understanding of teachers' practice.

We start with two foundational premises: (1) the measurement of teaching with a given observation system should be driven by the community's theoretical view of teaching quality ~~and explicitly seek to test and/or refine that theoretical understanding of teaching quality (Shavelson & Towne, 2002)~~ and (2) every choice from data collection through data analysis makes assumptions about the nature of teaching quality (e.g., a correlation assumes linear relationships; adopting an etic coding system assumes an external perspective is meaningful; taking an average assumes the amount of something matters). From these two premises, we ask whether there is alignment between the community of practice's theoretical understanding of teaching quality, which undergirds the observation system, and the assumptions about the nature of teaching quality that are implied by choices in the measurement and analysis process. These must be in alignment if observation systems are to contribute to an improved understanding of teaching quality (Hoelscher, 2017; Kelcey, 2015; Shavelson & Towne, 2002; Stein et al., 2017).

Rather than passing judgement on specific observation systems, we identify a small set of assumptions that frequently characterize communities of practice's views of teaching quality and discuss the measurement implications of these assumptions. However, in order to concretize these arguments, we draw examples from two common observation systems: the

Classroom Assessment and Scoring System (CLASS; Pianta et al., 2008) and the Protocol for Language Arts Teaching Observation (PLATO; Grossman, Loeb, Cohen, & Wyckoff, 2013). We select these two systems as examples solely due to our greater familiarity with them and explicitly note here that any critique we make here could be made of many observation systems. Our intention is to focus on the relationship between theoretical assumptions and their implications rather than CLASS or PLATO.

Through this paper, we hope to encourage observation system developers to more clearly lay out the theoretical underpinnings of their community of practice's understanding of teaching quality and to provide clear guidance around data collection and analysis approaches that are consistent with that vision. This will allow the field to move beyond simple tests of basic empirical relationships (e.g., correlations) towards building and testing alternative theories that let us deepen our understanding of the complex construct of teaching quality (Stein et al., 2017). A key takeaway here is that we cannot straightforwardly borrow measurement approaches across fields (or even across observation systems) without very carefully considering the assumptions about teaching quality that are implicit in those measurement approaches. We believe that stronger, coherent connections between theoretical understandings of teaching quality and the measurement of teaching can lead to (a) stronger connections between observation scores and measures of student learning, (b) clearer guidance for teachers from observation scores, (c) joint refinement of both theories of instruction and methods for measuring instruction (Hoelscher, 2017; Kelcey, 2015; Stein et al., 2017).

Description of CLASS and PLATO

Before discussing the common assumptions about teaching quality, we introduce CLASS and PLATO in detail, focusing on their theoretical perspectives. We will later draw upon these theories to discuss their measurement implications. Note that some of the details

we describe (especially for PLATO) are inferred because we could not find them explicitly stated anywhere or different sources provided conflicting views.

CLASS starts from a developmental psychology perspective and posits a bio-ecological view of development that emphasizes the classroom as a key proximal setting for student learning and development (Hamre & Pianta, 2010). The goal of CLASS appears to be to measure the classroom as a setting that structures the classroom interactions supporting positive youth learning and development (Downer, Sabol, & Hamre, 2010; Hamre & Pianta, 2010; Hamre, Pianta, Mashburn, & Downer, 2007). This is most clearly stated by Pianta (2008), who discusses CLASS as a “theory of the classroom as a setting for youth development” (p. 24)¹.

CLASS posits three classroom-level domains that structure interactions: Emotional Support, Classroom Organization, and Instructional Support. Each domain is measured by a set of dimensions within a factor-analytic model (i.e. reflective model; Jarvis et al., 2003). For simplicity, we discuss only the Emotional Support domain of CLASS, which captures the emotional nature of interactions in the classroom and to what extent the teacher is attuned to students’ interests and needs.

PLATO was created in order to capture instructional practices that are supportive of learning complex skills in English language arts (ELA) at the middle school and high school level (Grossman et al., 2013; Grossman, Cohen, Ronfeldt, & Brown, 2014).

“The theory underlying PLATO articulates the importance of rigorous content and intellectually challenging tasks, the centrality of classroom discourse in developing sophisticated understanding of content and disciplinary skills, and

¹ Some discussions of the Teaching Through Interactions theory, which is the formalization of theory underlying CLASS, have placed more emphasis on the measurement of interactions (as opposed to the classroom context that structures those interactions; e.g., Hamre et al., 2013). For this paper, we assume that CLASS measures classroom contexts because the majority of the literature seems to imply this and all the empirical work on CLASS implicitly assumes this (discussed more below).

the critical role of teachers in providing instructional scaffolding for students to help them succeed.” (Grossman et al., 2014 p. 295)

Then, PLATO presents a less cohesive theoretical grounding on how teaching promotes student learning and development, but identifies an eclectic collection of important disciplinary and generic practices. PLATO focuses directly on instructional interactions (as opposed to examining classroom settings). PLATO includes four domains, described as “conceptual clusters” of variables—Instructional Scaffolding, Representation and Use of Content, Disciplinary Demand, and Behavior Management (Grossman et al., 2014). Each domain contains several dimensions (termed *elements* in PLATO), but the nature and importance of the domain structure is left mostly unspecified (Cohen et al., 2016).

Note on Terminology

There is little agreed upon terminology in the field (Praetorius & Charalambos, 2018; Schlesinger & Jentsch, 2016), so we explicitly define key terms here. A dimension of teaching quality is any facet of teaching quality that is directly scored by raters. A domain of teaching quality is a grouping of dimensions that share a common feature and are given a label. Dimensions can be broken down into observable behaviors that we call indicators, which serve to clarify the dimension meaning and guide rater scoring, but indicators are not scored. These definitions are not broad enough to serve all purposes and adequately characterize all observation systems, but they serve the purposes of this paper.

Uses of Observation Scores

In the discussion below, we try to emphasize the practical implications of our often technical points. To facilitate this, we focus on three uses of observation systems, examining the implications of each point on each use of observation systems: (1) empirically linking teaching and learning; (2) comparing teaching across settings; and (3) providing feedback to teachers. In discussing the linking of teaching and measures of student 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. For example, student learning gains are generally estimated across a full school year requiring one to estimate the average level of teaching quality across the same school year before one can link teaching quality and student learning. In discussing comparisons of teaching across settings (Use 2), we emphasize the need to carefully consider the contextual sensitivity of measures of teaching quality and to ensure that measured constructs are comparable across settings. In discussing providing feedback to teachers (Use 3), we emphasize that feedback should be as detailed, specific, and actionable as possible (Hill & Grossman, 2013; Wylie, 2020). Specifically, feedback focuses on individual lessons² and balancing three goals: (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 observation system as a conceptual tool that supports thinking about teaching rather than rules to follow (Cohen et al., 2016).

Assumptions about Teaching

In this section of the paper, we discuss four assumptions about the nature of teaching that commonly characterize views of teaching. As such, these assumptions are principled characterizations of what teaching is, according to some communities of practice (i.e. we make no claim that they are universally agreed upon). For each assumption, we highlight its measurement implications, pointing out where common practices implicitly violate the assumption and pointing towards alternative, aligned practices. The reader is left to decide if they agree with each assumption and provided with guidance to align the measurement of teaching to each assumption.

² Unlike in some past work that emphasized providing reliable feedback about teachers' average practice (e.g., Taut & Rakoczy, 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, 2021).

Assumption 1: Teaching is an intentional activity.

Teachers set and pursue specific instructional intentions when planning lessons and engage in activities planned to achieve those specific intentions (Danielson, 2007; OECD, 2020). Additionally, observation systems often take a theoretical perspective that privileges specific learning goals and de-privileges other learning goals (Bell et al., 2019; Grossman et al., 2014). Then, teachers will, at times, pursue intentions that are less valued by observation systems, which may lead them to productively engage in instructional actions that are not valued by observation systems (Luoto & Selling, 2021; White, 2021). For example, a teacher might implement sustained silent reading to help students build their reading fluency. Sustained silent reading typically has limited instruction, teacher support, and classroom discourse. Then, dimensions of teaching quality focused on direct instruction (e.g., PLATO's Representation of Content) or the cognitive demand of student activities (e.g., PLATO's Intellectual Challenge) will necessarily score very low, even though such instruction might successfully meet the intention set by the teacher and student needs. We discuss three challenges implied by teaching being an intentional activity: (1) the average score on a dimension may not be a meaningful metric; (2) the measurement of teaching might have to align with teachers' intentions so that these intentions can be accounted for; and (3) whether the same observation system can or should be used to score all types of instruction.

When teaching is viewed as an intentional activity such that different teacher actions promote different outcomes, each dimension captured by the observation system may not be relevant for each lesson (Cohen et al., 2016). In this case, the average level of a dimension may not be nearly as important as the alignment between the score and the learning goals and/or teacher intentions. Importantly, the average dimension score is, in principle, independent of the alignment between that score and instructional intentions so the average score is not even a useful proxy for alignment. Said using statistical language, when

instructional intentions strongly moderate the relationship between observation scores and student outcomes, the average score on a dimension can often be a poor measure of the teaching quality in a classroom.

Many existing observation systems divide lessons into smaller, equal-interval segments (e.g., 7.5, 15, or 20 minute segments) and independently score those segments. This approach stems from the desire to measure average teaching quality, the ease of the approach, and concerns about how cognitively demanding it would be to score lessons as a whole (Joe, McClellan, & Holtzman, 2014; OECD, 2020). This practice sets the base unit of analysis as an equal-interval segment, but instructional intentions are unlikely to be neatly divided into equal-interval segments. The result is the observation scores represent intervals of instruction that have little connection to instructional intentions, which obscures the connection between teacher intentions and the scores resulting from observation systems. The alternative would be to create segments based on teacher intentions such that the unit of analysis (i.e. the segment) aligns with teacher intentions. The result would be segments of different lengths, but where each segment represents the quality of instruction as the teacher pursued specific intentions. This would connect the observation system's unit of analysis and teacher intentions, allowing for more advanced analyses that can account for teacher intentions.³ This alternative approach, of course, comes with its own challenges. Namely, it may be very difficult to reliably define segments that align with teacher intentions (Luoto & Selling, 2021). Additionally, the construct of teachers' intentions needs far more specificity than we could provide here, as teachers' initial intentions both can shift throughout a lesson and will differ from their realized intentions.

Many observation systems are used to score whatever instruction happens to be

³ Due to space, we must remain agnostic as to exactly how one might use instruction intentions (or other observable or reportable features of instruction) to divide lessons into segments that provide a more meaningful unit of analysis, but we direct the interested readers to Andrews, 2007; Burns & Anderson, 1987; or Clarke et al., 2007.

occurring in the classroom on a given day. It is of interest to consider those that break from this trend. For example, the Instructional Quality Assessment (IQA; Matsumura et al., 2008) is only used to capture the quality of mathematical discussions around high-quality assignments (this is coupled with analyses of the assignment and student work not based on observations). This effectively limits the range of teacher intentions and learning goals that will be encountered when using IQA, which in principle allows the IQA to be more carefully tailored to those teacher intentions. If one takes a strong view about the importance of teacher intentions, measuring teaching quality accurately may require using specialized systems that narrowly capture only key aspects of teaching quality (Kelly et al., 2020).

Implications for linking teaching and learning

The intentionality of teaching implies that the alignment between instructional practices and teacher intentions could be a major determinant of teaching quality. Typical approaches for linking teaching and learning aggregate observation scores across a number of days through averaging to characterize the level of teaching quality across a period of time (e.g., Kane et al., 2012, OECD, 2020). However, as just discussed, when the effectiveness of instructional practices is moderated by teacher intentions and/or learning goals, the average score provides a potentially poor estimate of teaching quality, which necessitates alternative approaches for aggregating data across time.

One alternative would be to align the unit of analysis (i.e. the way segments are defined) to teacher intentions. This would allow for analytic models that explicitly account for the potential moderating effects of teacher intentions when aggregating observation scores. For example, one could separately estimate the average intellectual challenge of teachers' instruction when teacher intentions do and do not focus on promoting students' conceptual understanding and then model whether the relationship between intellectual challenge scores and student learning varies across teachers' intention.

Implications for comparing teaching across settings

If teaching is treated as an intentional activity, then the unit of analysis when making comparisons across settings needs to be carefully considered (Andrews, 2007; Clarke et al., 2007; Luoto & Selling, 2021; Luoto, Klette & Blikstad-Balas, in review). In comparative studies of instructional quality, instruction is often compared without considering teacher intentions (e.g., Klette, Blikstad-Balas, & Roe, 2017; Maulana et al., 2020; OECD, 2020). We can, in fact, specify three possible comparisons that arise when we assume teaching is an intentional activity: (1) comparing the frequency and type of teacher intentions across settings; (2) comparing teaching practices within each commonly found instructional intention (i.e. looking at within-intention teaching); or (3) comparing overall teaching practices (either with or without adjustments for intentions; White, Maher, & Rowan, 2021b). The first two of these options is likely to lead to more valid comparisons that respect the complexity of instruction under the assumption that teaching is an intentional activity (c.f. Clarke, 2013; White et al., 2021b).

Implications for providing teacher feedback

When teaching is viewed as an intentional activity, teachers are, at any given time, trying to achieve specific goals. Connecting feedback to those goals, then, is important to help teachers understand the extent to which they are providing (or could provide) the instruction needed to reach those goals. This would seem to include both identifying which dimensions of teaching quality are relevant for specific goals, which can help teachers to see observation systems as useful cognitive scaffolds rather than prescriptions of practice (Cohen & Grossman, 2016), and providing clear feedback on the quality of observed instruction on those dimensions to help teachers calibrate their self-perceptions of their own instruction (Kraft & Hill, 2020).

A number of studies have sought to use Item Response Theory (IRT) models when examining data from observation systems (e.g., Jones & Bergin, 2019; van de Grift, Houtveen, van den Hurk, & Terpstra, 2019; van der Lans et al., 2018). IRT models place

teacher “ability estimates” and dimension “difficulty estimates” on the same scale. This has led authors to interpret some dimensions as more “difficult” than others and, importantly for teacher feedback, to identify teachers’ “zone of proximal development”-- i.e. those dimensions which a teacher must focus on mastering next (van de Grift et al., 2019). The assumption that teaching is an intentional activity undermines this approach and these claims. In traditional uses of IRT, all respondents (i.e. teachers) would be trying to score as high as possible on all items (i.e., dimensions). However, if teaching is intentional, then, in any given lesson, teachers may focus on doing well in only a subset of dimensions, conflating teachers’ ability to demonstrate a skill and teacher intentions. Then, items may be “difficult” because they are hard to achieve *or* because they are an infrequent focus of instruction. This introduces major conceptual challenges to identifying teachers’ “zones of proximal development” or interpreting specific dimensions as more or less difficult than others.

Assumption 2: Teaching Quality is a latent construct that cannot be directly observed

All observation systems break down teaching into a set of analytically separable dimensions and most observation systems organize dimensions into domains (Downer et al., 2010; Grossman et al., 2014). Establishing a domain structure is important for at least two reasons. First, there is broad consensus that a small number of domains support student learning and development (e.g. Klette, 2021; Praetorius & Charalombous, 2018). Each domain, though, is too broad to be measured directly by a single dimension, which naturally leads to a theoretical structure where dimensions are nested within domains. Second, the domain structure provides an analytical scaffold that limits challenges around multicollinearity and multiple comparisons (e.g., inflated false-positive rates; Simmons, Nelson, & Simonsohn, 2011), which occur in studies when each dimension is treated as independently meaningful (e.g., Cohen & Brown, 2016; Grossman et al., 2013; Klette et al., 2021).

There are two basic analytical models for a domain-dimension structure: reflective

models or formative models (Jarvis et al., 2003). Importantly, these models are theoretical descriptions of the construct being measured, which then lend themselves towards analytic techniques. The distinction between the two is best made with analogy. The reflective model can be understood with reference to Plato's cave. In Plato's cave, the unseen real object (i.e. the domain) casts shadows (i.e. the dimensions) against a wall. One assumption here is that the dimension scores (i.e. the shadows) are caused by the domain construct (i.e. the object casting the shadows) and dimension specific factors (e.g., the light source illuminating the object) that are considered error. The key construct in reflective models is the unobservable domain (i.e. the object casting the shadows) and inferences are made about the domain using the observed shadows (i.e. the dimensions). The complexity of trying to make inferences about the unobservable domain based on dimension scores naturally leads to concerns around factor structures and measurement invariance in the reflective model, as inferences about the domain assume conformity to such models. For example, CLASS, which posits a reflective model, has had numerous papers published examining its factor structure (e.g., Hamre et al., 2013, 2014; Hafen et al., 2014) and measurement invariance (e.g., Downer et al., 2012; Hu, Fan, Gu, & Yang, 2016).

The formative model can be understood with analogy to a model car. The model car (i.e. the domain) is created by combining a variety of parts, such as the wheels, engine, etc. (i.e. the dimensions). Under this model, the domain is created by the combination of the dimensions, but it takes on meaning beyond the sum of its parts, just as a model car is more than just a collection of pieces. Like in the reflective model, the domain is the focal construct. Both dimensions and domains are assumed to be directly observable in the formative model, whereas the reflective model assumes the domain is inherently unobservable. As a result, issues of factor structure, measurement invariance, Cohen's alpha, and other standard measurement approaches are often irrelevant in formative models (Jarvis et al., 2003) since

there is no need to rely on those approaches to support inferences about an unobservable construct. Empirically, we would use principal components analysis or simple averages to estimate domain scores under formative models, but more complicated structural equation models are also possible (Diamantopoulos & Winklhofer, 2001).

The distinction between reflective and formative models has important implications for observation systems. Consider CLASS's Emotional Support domain. Figure 1 shows both a reflective and formative representation of this domain, using conventions from structural equation modeling (Kline, 2011). The arrows imply that the object pointed from is causally prior to the object pointed to. The models differ due to the arrows' direction and the location of the measurement error.

Figure 1: CLASS's Emotional Support Domain.

To highlight the conceptual difference between these models, it helps to imagine one is trying to support teachers in improving the Emotional Support of their classroom. Under the formative model, one would directly intervene to help teachers improve, for example, Teacher Sensitivity. Higher Teacher Sensitivity would, by definition, imply higher levels of Emotional Support since Emotional Support is composed of Teacher Sensitivity in part.

Under the reflective model, one could not directly intervene to increase Teacher Sensitivity because Teacher Sensitivity is entirely determined by Emotional Support and an idiosyncratic Teacher Sensitivity error term. Rather, one could either intervene on the error term, which would increase Teacher Sensitivity, but have no impact on Emotional Support, or one could intervene directly on the construct of Emotional Support, which would tend to increase Teacher Sensitivity (and the other dimensions). Then, the reflective model assumes that any intervention meant to impact student learning must directly act at the domain level, rather than the dimension level. We believe this fact is under-recognized as the MyTeaching Partner intervention that is built around CLASS (e.g., Allen et al., 2011), appears to focus on the CLASS dimensions as the target of intervention (e.g., Gregory et al., 2014), which violates the proposed reflective model. While this distinction may feel like an overly technical point, the direct implication is that setting the target of intervention at the dimension level and assuming increases in dimension scores will lead to increases in domain scores implies a formative model. Further, factor structures and measurement invariance are less meaningful for formative models so switching to a formative model would fundamentally undermine the conceptual basis for the large body of work that examines the factor structure of CLASS observation scores (e.g. Allen et al., 2013; Blazar et al., 2017; Buyse et al., 2008; Downer et al., 2012; Hafen et al., 2014; Hamre et al., 2007, 2013, 2014; Jensen et al., 2020; Nguyen et al., 2015; Pakarinen et al., 2010; Virtanen et al., 2018).

There are also benefits of the reflective model over the formative model. The reflective model is more flexible in some sense. Under the reflective model, additional dimensions can be added or removed without changing the interpretation of the domain score. This might happen in order to increase the reliability with which a domain is measured or remove poorly performing items. Importantly, though, care must be taken when adding or removing domains so that the full range of the domain construct is still being measured.

Under the formative model, any change to the dimensions being measured will fundamentally shift the interpretation of the domain score. Further, the formative model assumes that all relevant features of the domain construct are being measured and those features not measured (i.e. the error term) are uncorrelated to student learning. The reflective model, on the other hand, can accurately estimate domain scores without measuring every possible piece of the domain construct.

Implications for linking teaching and learning

Both the reflective and the formative models explicitly assume that Teacher Sensitivity is related to student learning *only* through the Emotional Support domain. Connecting this to CLASS theory, the assumption here is that student learning is impacted by the degree to which the classroom is a safe, responsive, and engaging environment, as captured by the Emotional Support domain. The dimensions are not directly related to student outcomes. Then, analyses that look at the relationship between individual dimension scores and student learning (e.g., Cohen & Brown, 2016; Henry & Guthrie, 2015) are violating both the formative and reflective models. There is nothing in principle wrong with proposing and testing an alternative theory that has dimensions directly linked to student outcomes, but researchers should explicitly acknowledge when deviating from the theory set forth by developers of observation systems.

The distinction between reflective and formative models also has important implications for which analyses are theoretically coherent. For example, it is common to estimate a correlation between a dimension and student learning (e.g., Grossman et al., 2014; Klette et al., 2021). This correlation assumes a direct connection between the dimension and student learning. Under a formative model, this direct connection can be interpreted as the overall effect (both direct and indirect [through the domain]) of the dimension on student learning (Kline, 2011)—though note that the standard formative model assumes the direct effect is zero. Under the reflective model, however, this direct connection is conceptually

ambiguous. Consider Figure 1 and the result that would occur if we added a connection from positive climate to student learning. Then, both the Emotional Support domain and the Positive Climate error term would then be viewed as acting through Positive Climate to affect student learning. The direct connection (e.g., the correlation of Positive Climate and student learning) captures a mixture of these two effects, making it difficult to explain what the correlation captures⁴. Then, under the reflective model, correlations between dimensions and student learning are ambiguous to interpret and highly discouraged (though not uncommon, e.g., Cohen & Brown, 2016; Henry & Guthrie, 2015)

The PLATO instrument is useful to consider at this point. While PLATO has a domain structure, that structure, unlike CLASS, is not well specified. As an example, we take the instructional scaffolding domain, which PLATO decomposes into four dimensions: Modeling, Strategy Use and Instruction, Feedback, and Accommodations for Language learning. As far as we can tell, PLATO does not clearly conceptualize how these dimensions should relate to each other or the domain as a whole. At the most basic level, three possibilities exist: (1) a formative model; (2) a reflective model; or (3) each dimension independently impacts student learning. Empirical research with PLATO has implicitly made each assumption: each dimension independently influences learning (e.g., Cohen & Brown, 2016; Cohen & Grossman, 2016; Grossman et al., 2013; Klette et al., 2021), the domains related to dimensions in a reflective model (e.g., Cor, 2011; Grossman et al., 2014), and the domains relate to dimensions in a formative model (White, Maher, & Rowan, 2021a). Importantly, the use of these different models across studies hinders the accumulation of evidence by making it harder to compare results across studies (or to know if the choice

⁴ Because of this ambiguity, reflective models would typically connect the error term on Positive Climate to student learning, assuming that Positive Climate is composed of Emotional Support and a second factor, which impacts student learning independently of Emotional Support. Importantly, though, the correlation between Positive Climate and student learning does not capture this relationship between a second factor and student learning-- though bifactor type models could.

between models was driven by p-hacking; Gehlbach & Robinson, 2017). If PLATO developers provided clearer guidance for the appropriate theoretical models, it would be easier to accumulate knowledge of PLATO scores across studies, which could support the development of stronger theories around PLATO.

Implications for comparing teaching across settings

The nature of the relationship between domains and dimensions is also important for making comparisons across teachers/classrooms. Since the domain score in the reflective model is inferred through examination of the dimension scores, assuming a reflective model means one must establish measurement invariance before making comparisons of domains across contexts (i.e. one must examine the shadows and conclude that the object casting the shadows is the same across contexts; e.g., Downer et al., 2012; Hu et al., 2016; Kelcey, 2014; Kelcey et al., 2014; Maulana et al., 2019, 2020; van de Grift, 2014). Under the formative model, there is no need for measurement invariance to be established before comparing teaching quality at the domain level across settings, because the domain scores are directly observable (and not inferred). There will still likely be the interest in examining the relationship between dimensions within each setting, which, in this case, simply provides information about how teaching quality manifests across settings.

To this point, we have ignored the indicator level. However, just like the relationship between dimensions and domains, the relationship between indicators and dimensions can be specified as a reflective or formative model. In much the same way, specifying a reflective model implies the need for measurement invariance across indicators when comparing dimension scores across settings. This is, of course, much more problematic since the indicators are not scored, but are meant to support rater scoring. In order to validate observation systems to make comparisons across settings, this issue of dimensions' measurement invariance needs careful attention (White, 2021a).

Implications for providing teacher feedback

The main implication for teacher feedback here relates to helping teachers see observation systems as conceptual tools (Cohen et al., 2016). Any model organizing dimensions into domains helps to clarify for teachers the relationship between the many dimensions and can help narrow teachers' focus to a small number of domains that are directly related to student outcomes. Additionally, as just discussed, the reflective model posits the more abstract domain as teachers' improvement goal, identifying the dimensions as useful ways of tracking domain-level changes, while the formative model posits the more concrete dimensions as teachers' improvement goal. Then, depending on whether domains are assumed to be unobservable variables (i.e. reflective model) or observable variables (i.e. formative model), the feedback teachers receive for how they would go about changing their practice should change.

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

Educational 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 specific questions). The implication of this assumption is that observation systems can (or maybe should) measure teaching quality at different levels (i.e. measure dimensions and domains that are conceptualized as functioning at different levels). A number of levels are possible: segment-level, lesson-level, classroom-level, or teacher-level (and possibly others). The level at which a construct is defined is an important component of its theorization. For example, a (trait) construct defined at the classroom level is conceptualized as a stable feature possessed by the classroom (and so present in all lessons).

Consider CLASS and PLATO. As discussed above, CLASS measures teaching quality at the classroom-level (i.e. the ability of the classroom environment to serve as a

setting that promotes the high-level interactions that facilitate student learning and development). PLATO measures teaching quality at the segment level, which is the lowest level at which scores exist. This theoretical distinction can be seen in empirical work as analyses of CLASS scores typically use classroom-level aggregates (e.g., Hamre et al., 2007) and analyses of PLATO scores often examine segment level distributions (e.g., Cohen & Grossman, 2011; Klette et al., 2021).

The level at which a dimension/domain is conceptualized has both theoretical and measurement implications. Classroom-level dimensions are generally conceptualized as stable over time and so they are represented by a single score for the classroom that typifies the level of the construct in the classroom (e.g., averaging the score assigned across all scored segments). As such, classroom-level constructs are often those that enable other high-quality practices (e.g., Charalambous & Praetorius, 2020; Cohen et al., 2016), such as classroom climate (Marsh et al., 2012) or management (when management is focused on routines; e.g., Lester, Allanson, & Notar, 2017). Unless within-classroom variability is separately theorized, the daily fluctuations in classroom-level scores are best understood as measurement error. Segment level constructs, on the other hand, generally capture those instructional practices that are responsive to student actions or that vary based on content or classroom activities so these constructs are assumed to vary across time within classrooms and lessons.

Consider CLASS's Emotional Support and PLATO's Instructional Scaffolding domains as examples. The theoretical bases of Emotional Support are attachment theory and self-determination theory⁵, both of which emphasize the importance of stable, predictable environments that lead students to feel a "sense of security that allows for exploration of novel, stretching experiences and a feeling of connectedness to teachers and schools" (Downer et al., 2010, p. 704). That is, the theory itself emphasizes the importance of stable,

⁵ While self-determination theory is not necessarily a setting-level construct, the references cited when discussing self-determination theory focus on stable setting features.

supportive contexts that would support students' feelings of security and safety.

Theoretically, there is little apparent basis from which to interpret daily fluctuations. In fact, daily fluctuations, especially if those fluctuations were not highly predictable for students, are potentially problematic theoretically for Emotional Support since they reflect instability that might damage students' sense of security. Empirically, then, we would estimate a single Emotional Support score to represent a classroom over a period of time, which would capture the stable ways that the classroom supports students developing a sense of security.

Contrast this with Instructional Scaffolding, which captures the ways that teachers support students in mastering specific content, emphasizing the varied types of support provided by teachers (Grossman et al., 2014). Daily fluctuations in Instructional Scaffolding scores, then, are theoretically expected and justified as teachers providing different levels and types of support across different content and activities. This theory then leads to the empirical question of whether teachers are targeting scaffolding to those situations where it is needed. For example, a study could compare segments (within-teacher) with high Instructional Scaffolding to those with low Instructional Scaffolding to understand where, when, and why teachers engage in scaffolding—analyses that would be difficult to justify when measuring stable classroom-level constructs like Emotional Support. If teachers are reasonably targeting their scaffolding practices, then high levels of variation across days would be a positive finding. In fact, theoretically, we might expect low levels of variation in Instructional Scaffolding scores to be a negative finding as it implies teachers are targeting their scaffolding efforts to those situations it is most needed. Unlike with classroom-level constructs, segment-level constructs, then, would be much more likely to vary systematically across teacher intentions (See Assumption 1).

Classroom-level and segment-level dimensions must be combined into domains using different approaches. Segment-level domains are defined at the segment-level, so dimensions

should be aggregated into domains at that level (see Assumption 2). Afterwards, the domain scores can be aggregated across segments if required. A different approach is needed for classroom-level domains. The standard approach in the vast majority of studies with the CLASS system is to first aggregate data to the classroom level and run a factor analysis of classroom-level average domain scores (e.g., Allen et al., 2013; Blazar et al., 2017; Downer et al., 2012; Hafen et al., 2014; Hamre et al., 2014; Jensen et al., 2020). McCaffrey, Yuan, Savitsky, Lockwood, & Edelen (2014) show the error of this approach. Rather, multi-level factor analysis is necessary (Muthén, 1994). Multi-level factor models simultaneously aggregate scores across segments and dimensions to examine domain structures, assuming a reflective model. Note that we know of no similar approaches for formative models.

Implications for linking teaching and learning

When examining the link between teaching and learning, researchers will need to carefully consider the relationship between segment-level and classroom-level domains. For example, Charalambous & Praetorius's (2020) MAIN-TEACH model, while not fully flushed out, seems to describe three intermediate domains (i.e. classroom-level domains) that set the stage for instructional interactions (i.e. segment-level domains), enabling those interactions. Measurement-wise, the implication seems to be that the classroom-level, intermediate domains moderate the relationship between the segment-level domains and student outcomes. Others have similarly argued that classroom-level management domains must meet a specific threshold before dimension-level instructional domains are to impact learning (Cohen et al., 2016; Klette et al., 2021).

Researchers should note that since classroom-level domains are likely more stable over time (i.e. more reliably estimated) and often conceptualized as enabling factors (i.e. as bolstering or moderating the impact of other domains on student learning), we should expect these domains to have higher correlations with student learning (e.g., Gill, Shoji, Coen, & Place, 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 effects) and testing that theory through targeted models can we better understand how teaching relates to learning.

As discussed briefly above, segment-level constructs make possible more complex research designs than we have observed before because day-to-day fluctuations should be associated with day-to-day differences in student learning. Then, studies that measure student learning at frequent intervals and explore how day-to-day fluctuations of teaching quality impact day-to-day, within-student variation in learning should be feasible. For classroom-level constructs, such studies would violate the assumed nature of the construct.

Implications for comparing teaching across settings

The most direct implication for comparing teaching across settings is that segment-level dimensions/domains should be compared at the segment level and classroom-level dimensions/domains at the classroom level. This likely makes classroom-level dimensions/domains easier to use for comparisons, as the unit of analysis is more clearly defined. When comparing segment-level dimensions/domains, researchers will need to carefully consider contextual factors (e.g., sampling decisions regarding instructional intentions and activities, see Assumption 1 and 4) in order to make clear and equivalent comparisons and interpretations.

Implications for providing teacher feedback

The level of the construct has implications for providing teachers feedback. Segment level constructs can be directly observed and scored from a single lesson, such that a skilled observer can provide teachers feedback on the level of segment level constructs within a lesson after observing a single lesson. Then, feedback on segment level constructs is quite direct, as the observation score captures the level of the construct in the observed lesson.

Classroom-level constructs, though, exist at the classroom level so a single lesson can at most be a window into the level of that construct. For example, an observer can watch a

lesson and score Emotional Climate within a given lesson, but this score is at best a guess for the extent to which the classroom provides a “sense of security that allows for exploration of novel, stretching experiences” (i.e., for the construct of Emotional Climate; Downer et al., 2010, p. 704). That is, a single lesson does not provide a broad enough view of the classroom to support strong conclusions about the level of a classroom construct. Then, feedback on classroom-level constructs based on a single observed lesson cannot reveal to teachers the precise level of a given construct, since a single lesson provides imprecise estimates of classroom-level constructs (e.g., Kane et al., 2012). Rather, teachers will need to be supported in understanding that observers are making inferences about the level of the classroom-level construct based on what they observe. Then, feedback on classroom-level constructs should probably deemphasize specific scores and/or aggregate scores across many lessons when emphasizing the feedback value of scores.

Assumption 4: Teaching varies Systematically based on Contextual Features of a Lesson

Lessons occur within school systems and cultures, focus on specific content, and are provided to specific groups of students. Each of these factors, and others, contributes to a broader lesson context that can influence instruction (Hiebert & Grouws, 2007). Note that the lesson context influences teachers’ choices and intentions so the points raised here overlap somewhat with those under Assumption 1. The emphasis here is on how the lesson context invariably impacts teaching while Assumption 1’s emphasis was on teachers’ intentions. The main difference, then, is that here we focus on systematic effects across teachers while Assumption 1 focused on teachers’ individual choices and intentions.

The lesson context can impact the measurement of teaching quality in at least three ways. First, it can make some dimensions of teaching quality more or less relevant based on contextual features of lessons. For example, it has been argued that the first few days of instruction 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, Roehrig, & Pressley, 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 (Luoto et al., under review; Xu & Clarke, 2019). 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 US 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). Then, instructional practice during the testing season might not be reflective of teachers' broader practices.

Third, the effects of the lesson context on teaching can make it quite difficult to generalize estimates of teaching quality beyond the observed lessons, especially for individual teachers. Most research focusing on the generalizability of observation scores has focused solely on the question of how many lessons one must observe to estimate the average teaching quality in a classroom (e.g., Brophy, 2006; Kane et al., 2012; Praetorius et al., 2014). This research has assumed, usually implicitly, that all days of instruction are exchangeable. That is, that every lesson is just like any other and each provides (on average) the same view into the teaching quality in the classroom. But, if one assumes the lesson context impacts the quality of teaching observed in classrooms, then lessons are not exchangeable because each lesson provides a different view into the classroom's instructional quality.

Consider, for example, the Linking Student Instruction to Achievement (LISA; Klette et al., 2018) project. The LISA project filmed each teacher on four consecutive days. Then, if one believes that the content of instruction or time of year (both features of lesson context) impact teaching, one could not generalize to estimate the average teaching quality for any individual teacher since the study only observes teachers teaching a specific topic at a specific

point in the school year. Similarly, estimates of the within-teacher variability in teaching quality from the LISA study will likely under-estimate the within-teacher variability in teaching quality that exists since teachers are observed on consecutive days, at least if one believes teaching varies systematically by time of year or content. On the other hand, since the LISA study collected data across the school year, if we assume that the specific 4-lesson block that a given teacher was observed is unrelated to how that specific teacher teaches on those days (i.e. no bias in the sampling of days of instruction across teachers), then the LISA study can generalize to the average teaching quality of teachers in the sample, even when we assume lesson contexts impact teaching. A full treatment of the sampling of lessons is beyond the scope of this paper (see White, 2021 for a broader, yet still incomplete treatment). The main takeaway of this section is that if one assumes that aspects of the lesson context impact teaching, when measuring teaching quality, one will need to carefully consider how to sample days of teaching and how the lesson context might impact the conclusions one hopes to draw.

Implications for linking teaching and learning

Two implications for linking teaching and learning are clear from the assumption that lesson context matters. First, linking teaching and learning requires estimating a statistic that characterizes the quality of teaching from the pre-test time point to the post-test time point (i.e., so teaching and learning are measured across the same time period), which generally requires generalizing from observed lessons to instruction across a broader period of time. The lack of exchangeability of days of instruction makes this generalization process difficult, and researchers will need to take great care in sampling lessons so that the observed lessons are fully representative of teachers' broader practice within the time period. Randomly sampling lessons will achieve this, but is difficult to accomplish. This process is made easier when restricting the measure of learning to specific content, thereby narrowing the lesson contexts across which one must generalize (Lipowsky et al., 2009; OECD, 2020).

Second, if features of the lesson context make some dimensions of teaching quality

more or less relevant in specific lessons, the model one uses to connect teaching and learning will need to account for those features. It is reasonable to think that at least some features of the lesson context will moderate the relationship between observed teaching quality and student learning. Such moderation should be explicitly modeled and tested so that we can deepen our understanding of teaching quality.

Implications for comparing teaching across settings

If teaching is impacted by the lesson context, this has large implications for comparing teaching across settings, as lesson contexts are likely to vary systematically across settings. One could, in principle, still estimate average levels of teaching in each setting to compare average teaching quality across settings, though note the above discussions of the difficulties the non-exchangeability of lessons creates for this. Rather, a better approach in this case might be to carefully consider and control for features of the lesson context in sampling so that comparisons of classrooms across contexts are made for lessons that share as many features of the lesson context as feasible (c.f., White, 2021). This approach is somewhat common. For example, the Pythagoras study focused on students' first introduction to the Pythagorean Theorem across countries, controlling for many features of the lesson contexts (Klieme & Reusser, 2009).

Implications for providing teacher feedback

Similar to our implications for teacher intentions (see Assumption 1), the implication for teacher feedback here is that not all dimensions of teaching quality are equally relevant for teachers to receive feedback on at all times. For example, classroom-level variables (e.g., routines and climate) may be especially important at the start of the year (Bohn et al., 2004). Feedback to teachers, then, must consider the lesson context and reflect a contextualized understanding of what happened during the lesson.

Discussion

In this paper, we have argued for the importance of aligning the theoretical

understandings of teaching quality that undergird observation systems with the analytic and measurement choices made when measuring teaching quality. There needs to be alignment between theory and measurement in order to both improve our measurements and to create evidence that can test and validate theories. We see the alignment of theory and measurement as being important in three uses of observation systems. First, the alignment of theory and measurement can hopefully help improve the empirical relationship between teaching and learning, which has been disappointingly low and inconsistent (e.g., Kane et al., 2012; Klette et al., 2021; OECD, 2020). At the very least, the areas of problematic alignment that we identified here are potential explanations of this weak relationship. Second, the alignment of theory and measurement can help improve the precision and clarity of comparisons across contexts by ensuring that those comparisons account for potentially important contextual factors (e.g., lesson context and teacher intentions). Such alignment will hopefully protect against researchers and/or policymakers drawing inappropriate conclusions from comparisons. Third, the alignment of theory and measurement can improve the feedback provided to teachers by ensuring that teacher feedback aligns with theoretical 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 theory and measurement requires observation systems to more fully and completely develop the theoretical views that serve the basis for the dimensions of teaching quality that they capture. Such deeper specification of theory needs to be done with an eye towards measurement implications so that researchers using observation systems have guidance in terms of how to collect and analyze observational data. We have highlighted four assumptions that we believe are common in more theoretical conceptualizations of teaching and shown how these assumptions are implicitly violated in many common data collection and analysis approaches. 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.

Towards Programmatic Approaches of Studying Teaching Quality

Despite our critiques of the way observation systems 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. Standardized observation systems allow systematic inquiries into teaching quality that support the accumulation of research knowledge across time that is necessary in order to test and validate theories of teaching quality (Klette, 2020). A key takeaway from this paper is that this accumulation of empirical evidence to test and validate theories does not happen automatically with the use of observation systems. Rather, researchers must collect and analyze data from observation systems in ways that are aligned with the theoretical understanding of teaching quality contained in the observation system. This requires observation system developers to provide more clear guidance around appropriate approaches to the collection and analysis of observation scores and to align that guidance with the theoretical understandings underlying their observation system. It also requires researchers using observation systems 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 theories of teaching quality captured by an observation system.

Take PLATO as an example. Research using PLATO has made a wide range of implicit assumptions about the nature of teaching quality: i.e., that dimensions directly impact student learning (e.g., Cohen & Brown, 2016; Cohen & Grossman, 2016; Klette et al., 2021), that a set of domains under a reflective model impact student learning (e.g., Cor, 2011; Grossman et al., 2014), and that a set of domains under a formative model impact student

learning (White et al., under review). In taking these different analytic approaches, we have not seen authors highlight the assumptions they are making about the nature of teaching quality nor have we seen authors compare their given approach with alternative models to compare which fits the data better. Given the different assumptions about the nature of teaching quality and the different analytic models, it becomes difficult to accumulate research knowledge across studies. Research using CLASS has been stronger on this front with many studies using consistent analytic approaches to explore CLASS data across studies (e.g., Hamre et al., 2007). Of course, care must be taken to ensure that the tested model matches the underlying theory (see Assumption 2). Moving towards programmatic research on teaching quality requires not just using PLATO (or other such systems) across studies, but adopting common theoretical understandings of teaching quality and analytic models across studies.

Our push for programmatic research does not mean that all research should be restricted to the standard, common models that we argue observation system 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 theoretical, empirical, or practical benefits of the proposed alternative conceptualization. CLASS research that compares multiple theoretical conceptualizations again provides a good guide here (e.g., Hafen et al., 2014), though we would encourage research that moves beyond simply examining model fit statistics and moves towards exploring the practical implications of different models. For CLASS, this 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 observation systems in order to support researchers in going back to re-examine older data using newly proposed models.

Consequences for the Measurement of Teaching Quality

An important implication of the points made in this paper is that different observation systems need to collect data in different ways and be analyzed with different approaches depending on the theoretical underpinnings of the observation system. In fact, many observation systems could propose different theoretical models for different dimensions within the same observation system, especially if constructs are defined at different levels. For example, the MAIN-TEACH meta-model proposes segment-level and classroom-level constructs (Charalambous & Praetorius, 2020), which will generally need different analytic models. This will increase the complexity of analyzing data from observation systems, but will be necessary to align the analyses with underlying theories so that those theories can be tested. Again, observation system developers should provide clear guidance to help clarify underlying theories and guide others in appropriately analyzing data from their systems.

Bibliography

- Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement. *Science*, 333(6045), 1034–1037. DOI: 10.1126/science.1207998
- Allen, J. P., Gregory, A., Mikami, A., Lun, J., Hamre, B. K., & Pianta, R. C. (2013). Observations of Effective Teacher-Student Interactions in Secondary School Classrooms: Predicting Student Achievement With the Classroom Assessment Scoring System--Secondary. *School Psychology Review*, 42(1), 76–97.
- Andrews, P. (2007). Mathematics teacher typologies or nationally located patterns of behaviour? *International Journal of Educational Research*, 46(5), 306–318.
<https://doi.org/10.1016/j.ijer.2007.10.001>
- Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, 73, 101919. DOI: [j.econedurev.2019.101919](https://doi.org/10.1016/j.econedurev.2019.101919)
- Ball, D. L., & Forzani, F. M. (2007). 2007 Wallace Foundation Distinguished Lecture—What Makes Education “Research Educational”? *Educational Researcher*, 36, 529–540.
<https://doi.org/10.3102/0013189X07312896>
- Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An Argument Approach to Observation Protocol Validity. *Educational Assessment*, 17(2–3), 62–87. <https://doi.org/10.1080/10627197.2012.715014>
- 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.
<https://doi.org/10.1080/09243453.2018.1539014>
- Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2017). Attending to General and

- Mathematics-Specific Dimensions of Teaching: Exploring Factors Across Two Observation Instruments. *Educational Assessment*, 22(2), 71–94. DOI: [10.1080/10627197.2017.1309274](https://doi.org/10.1080/10627197.2017.1309274)
- 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.
- Brophy, J. (2006). Observational Research on Generic Aspects of Classroom Teaching. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 755–780). Lawrence Erlbaum Associates Publishers.
- Burns, R. B., & Anderson, L. W. (1987). The Activity Structure of Lesson Segments. *Curriculum Inquiry*, 17(1), 31–53. <https://doi.org/10.1080/03626784.1987.11075276>
- Buyse, E., Verschueren, K., Doumen, S., Van Damme, J., & Maes, F. (2008). Classroom problem behavior and teacher-child relationships in kindergarten: The moderating role of classroom climate. *Journal of School Psychology*, 46(4), 367–391. [10.1016/j.jsp.2007.06.009](https://doi.org/10.1016/j.jsp.2007.06.009).
- 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.2020.100894>
- Clarke, D., Mesiti, C., O’Keefe, C., Xu, L. H., Jablonka, E., Mok, I. A. C., & Shimizu, Y. (2007). Addressing the challenge of legitimate international comparisons of classroom practice. *International Journal of Educational Research*, 46(5), 280–293.
- Clarke, D. (2013). International Comparative Research into Educational Interaction: Constructing and Concealing Difference in: Interaction in Educational Domains. In *Interaction in Educational Domains* (pp. 5–22). Sense Publishers.
- Cohen, J. J., & Brown, M. (2016). Teaching Quality Across School Settings. *The New*

- Educator*, 12(2), 191–218. DOI:10.1080/1547688X.2016.1156459
- Cohen, J. J., & Grossman, P. (2011, April). Of Cabbages and Kings: Classroom Observations & Value-Added Measures. Annual Meeting of AERA.
- Cohen, J. J., & Grossman, P. (2016). Respecting complexity in measures of teaching: Keeping students and schools in focus. *Teaching and Teacher Education*, 55, 308–317. <https://doi.org/10.1016/j.tate.2016.01.017>
- 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*, 35.
- Cor, M. K. (2011). Investigating the Reliability of Classroom Observation Protocols: The Case of PLATO. Stanford University School of Education.
- Creemers, B. P. M., & Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement*, 17(3), 347–366. DOI:10.1080/09243450600697242
- Danielson, C. (2007). *Enhancing Professional Practice: A Framework for Teaching*, 2nd Edition (2nd edition). Association for Supervision & Curriculum Development.
- Diamantopoulos, A., & Winklhofer, H. M. (2001). Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research*, 38(2), 269–277. DOI:10.1509/jmkr.38.2.269.18845
- Downer, J. T., López, M. L., Grimm, K.J., Hamagami, A., Pianta, R. C., & Howes, C. (2012). Observations of teacher-child interactions in classrooms serving Latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly*, 27(1), 21–32. [10.1016/j.ecresq.2011.07.005](https://doi.org/10.1016/j.ecresq.2011.07.005)
- Downer, J., Sabol, T. J., & Hamre, B. K. (2010). Teacher–Child Interactions in the

- Classroom: Toward a Theory of Within- and Cross-Domain Links to Children's Developmental Outcomes. *Early Education and Development*, 21(5), 699–723.
<https://doi.org/10.1080/10409289.2010.497453>
- Emmer, E. T., Evertson, C. M., & Anderson, L. M. (1980). Effective Classroom Management at the Beginning of the School Year. *The Elementary School Journal*, 80(5), 219–231.
<https://doi.org/10.1086/461192>
- Gehlbach, H., & Robinson, C. D. (2017). Mitigating Illusory Results through Preregistration in Education. *Journal of Research on Educational Effectiveness*, 11(2), 1–20.
DOI:10.1080/19345747.2017.1387950
- Gill, B., Shoji, M., Coen, T., & Place, K. (2016). The content, predictive power, and potential bias in five widely used teacher observation instruments (Indexes; Offices REL 2017-191). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. <http://ies.ed.gov/ncee/edlabs>
- Gregory, A., Allen, J. P., Mikami, A. Y., Hafen, C. A., & Pianta, R. C. (2014). Effects of a Professional Development Program on Behavioral Engagement of Students in Middle and High School. *Psychology in the Schools*, 51(2), 143–163. 10.1002/pits.21741
- Grossman, P., Loeb, S., Cohen, J. J., & Wyckoff, J. (2013). Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores. *American Journal of Education*, 119(3), 445–470. <https://doi.org/10.1086/669901>
- Grossman, P., Cohen, J. J., Ronfeldt, M., & Brown, L. (2014). The Test Matters The Relationship Between Classroom Observation Scores and Teacher Value Added on Multiple Types of Assessment. *Educational Researcher*, 43(6), 293–303. DOI: 10.3102/0013189X14544542

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2014).

Teaching Through Interactions in Secondary School Classrooms Revisiting the Factor Structure and Practical Application of the Classroom Assessment Scoring System—Secondary. *The Journal of Early Adolescence*, 0272431614537117.

Hamre, B. K., Hatfield, B., Pianta, R. C., & Jamil, F. (2014). Evidence for General and Domain-Specific Elements of Teacher–Child Interactions: Associations With Preschool Children’s Development. *Child Development*, 85(3), 1257–1274. [10.1111/cdev.12184](https://doi.org/10.1111/cdev.12184)

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M., & Hamagami, A. (2013). Teaching through Interactions: Testing a Developmental Framework of Teacher Effectiveness in over 4,000 Classrooms. *The Elementary School Journal*, 113(4), 461–487. <https://doi.org/10.1086/669616>

Hamre, B. K., Pianta, R. C., Mashburn, A., & Downer, J. T. (2007). *Building a Science of Classrooms: Application of the CLASS Framework in over 4,000 U.S. Early Childhood and Elementary Classrooms*. 1–35.

Hamre, B. K., & Pianta, R. C. (2010). Classroom Environments and Developmental Processes. In *Handbook of Research on Schools, Schooling and Human Development* (pp. 25–41). Routledge.

Henry, G. T., & Guthrie, J. E. (2016). Using multiple measures for developmental teacher evaluation. In *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 143–155). Teachers College Press.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students’ learning. *Second Handbook of Research on Mathematics Teaching and Learning*, 1, 371–404.

Hill, H. C., & Grossman, P. (2013). Learning from Teacher Observations: Challenges and

- Opportunities Posed by New Teacher Evaluation Systems. *Harvard Educational Review*, 83(2), 371-384,401.
- Hoelscher, M. (2017). "It's the Method, Stupid." Interrelations between Methodological and Theoretical Advances: The Example of Comparing Higher Education Systems Internationally. *Research in Comparative and International Education*, 12(3), 347–359. <https://doi.org/10.1177/1745499917724287>
- Hu, B. Y., Fan, X., Gu, C., & Yang, N. (2016). Applicability of the Classroom Assessment Scoring System in Chinese Preschools Based on Psychometric Evidence. *Early Education and Development*, 27(5), 714–734. [10.1080/10409289.2016.1113069](https://doi.org/10.1080/10409289.2016.1113069)
- Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research. *Journal of Consumer Research*, 30(2), 199–218. DOI:10.1086/376806
- Jensen, B., Martínez, M. G. P., Medina, A. M. G., Martínez, J. F., Cox, C. B., & Larsen, R. (2020). An ecological analysis of the Classroom Assessment Scoring System in K-1 Mexican classrooms. *Early Years*, 0(0), 1–20. DOI: [10.1080/09575146.2020.1749035](https://doi.org/10.1080/09575146.2020.1749035)
- Joe, J. N., McClellan, C., & Holtzman, S. L. (2014). Scoring Design Decisions: Reliability and the Length and Focus of Classroom Observations. In *Designing teacher evaluation systems: New guidance from the measures of effecting project* (pp. 415–443). Jossey-Bass.
- Jones, E., & Bergin, C. (2019). Evaluating Teacher Effectiveness Using Classroom Observations: A Rasch Analysis of the Rater Effects of Principals. *Educational Assessment*, 24(2), 91–118. <https://doi.org/10.1080/10627197.2018.1564272>
- 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. (2014). *Noninvariant Measurement in Rater-Mediated Assessments of Teaching Quality*. <http://eric.ed.gov/?id=ED562804>
- Kelcey, B. (2015). Measuring Teachers' Instruction with Multilevel Item Response Theory. *International Journal of Research in Educational Methodology*, 7(1), 1037–1047. <https://doi.org/10.24297/ijrem.v7i1.3857>
- Kelcey, B., McGinn, D., & Hill, H. (2014). Approximate measurement invariance in cross-classified rater-mediated assessments. *Frontiers in Psychology*, 5. DOI:10.3389/fpsyg.2014.01469
- Kelly, S., Bringe, R., Aucejo, E., & Fruehwirth, J. C. (2020). Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives*, 28, 62. 10.14507/epaa.28.5012
- Kimball, S. M. (2002). Analysis of Feedback, Enabling Conditions and Fairness Perceptions of Teachers in Three School Districts with New Standards-Based Evaluation Systems. *Journal of Personnel Evaluation in Education*, 16(4), 241–268. DOI: 10.1023/A:1021787806189
- Klette, K. (2020). Towards programmatic research when studying classroom teaching and learning. In L. Ligozat, A. Rakhkochkine, & J Almqvist (Eds.), *Thinking through Didactics in a Changing World. European Perspectives on Learning, Teaching and the Curriculum*. Routledge Education.
- Klette, K., & Blikstad-Balas, M. (2018). Observation manuals as lenses to classroom teaching: Pitfalls and possibilities. *European Educational Research Journal*, 17(1), 129–146. <https://doi.org/10.1177/1474904117703228>
- Klette, K., Blikstad-Balas, M., & Roe, A. (2017). Linking Instruction and Student Achievement. A research design for a new generation of classroom studies. *Acta*

Didactica Norge, 11(3), 19. <https://doi.org/10.5617/adno.4729>

Klette, K., Roe, A., & Blikstad-Balas, M. (2021). Do all student groups benefit from what is regarded as high-quality teaching? *Analysing Teaching Quality: Perspectives, Principles and Pitfalls*. Scandinavian University Press.

Klieme, E., & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann Publishing Co.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (3rd ed). The Guilford Press.

Kraft, M. A., & Hill, H. C. (2020). Developing Ambitious Mathematics Instruction Through Web-Based Coaching: A Randomized Field Trial. *American Educational Research Journal*. DOI:10.3102/0002831220916840

Lester, R. R., Allanson, P. B., & Notar, C. E. (2017). Routines Are the Foundation of Classroom Management. *Education*, 137(4), 398–412.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, 19(6), 527–537. <https://doi.org/10.1016/j.learninstruc.2008.11.001>

Luoto, J. & Selling, A. (2021). Exploring the potential in using teachers' intended lesson goals as a context-sensitive lens to understanding observational scores of instructional quality. *Analysing Teaching Quality: Perspectives, Principles and Pitfalls*. Scandinavian University Press.

Luoto, J., Klette, K., & Blikstad-Balas, M., (under review). Possible biases in observation systems when applied across contexts: Conceptualizing, operationalizing, and sequencing

instructional quality.

- Lynch, K., Chin, M., & Blazar, D. (2017). Relationships between Observations of Elementary Mathematics Instruction and Student Achievement: Exploring Variability across Districts. *American Journal of Education*, 123(4), 615–646. DOI:10.1086/692662
- Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation*, 49, 15–29. DOI: [10.1016/j.stueduc.2016.03.002](https://doi.org/10.1016/j.stueduc.2016.03.002)
- Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom Climate and Contextual Effects: Conceptual and Methodological Issues in the Evaluation of Group-Level Effects. *Educational Psychologist*, 47(2), 106–124. <https://doi.org/10.1080/00461520.2012.670488>
- Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward Measuring Instructional Interactions “At-Scale.” *Educational Assessment*, 13(4), 267–300. DOI: 10.1080/10627190802602541
- Maulana, R., Smale-Jacobse, A., Helms-Lorenz, M., Chun, S., & Lee, O. (2019). Measuring differentiated instruction in The Netherlands and South Korea: Factor structure equivalence, correlates, and complexity level. *European Journal of Psychology of Education*. DOI: [10.1007/s10212-019-00446-4](https://doi.org/10.1007/s10212-019-00446-4)
- Maulana, R., André, S., Helms-Lorenz, M., Ko, J., Chun, S., Shahzad, A., Irnidayanti, Y., Lee, O., de Jager, T., Coetzee, T., & Fadhillah, N. (2020). Observed teaching behaviour in secondary education across six countries: Measurement invariance and indication of cross-national variations. *School Effectiveness and School Improvement*, 1–32. DOI: [10.1080/09243453.2020.1777170](https://doi.org/10.1080/09243453.2020.1777170)
- McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2014). Uncovering Multivariate Structure in Classroom Observations in the Presence of Rater

- Errors. *Educational Measurement: Issues and Practice*, 34(2), 34–46.
- Muijs, D., Reynolds, D., Sammons, P., Kyriakides, L., Creemers, B. P. M., & Teddlie, C. (2018). Assessing individual lessons using a generic teacher observation instrument: How useful is the International System for Teacher Observation and Feedback (ISTOF)? *ZDM*, 50(3), 395–406. DOI: 10.1007/s11858-018-0921-9
- Muthén, B. O. (1994). Multilevel Covariance Structure Analysis. *Sociological Methods & Research*, 22(3), 376–398. <https://doi.org/10.1177/0049124194022003006>
- Nguyen, C., Adams, R. J., Tayler, C., Cloney, D., & Cleveland, G. (2015). *The Psychometric Properties of the Classroom Assessment Scoring System (CLASS)*. AERA, Chicago, IL.
- OECD. (2020). *Global Teaching InSights: A Video Study of Teaching*. OECD Publishing. <https://doi.org/10.1787/20d6f36b-en>
- Pakarinen, E., Lerkkanen, M.-K., Poikkeus, A.-M., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., & Nurmi, J.-E. (2010). A Validation of the Classroom Assessment Scoring System in Finnish Kindergartens. *Early Education and Development*, 21(1), 95–124. DOI: [10.1080/10409280902858764](https://doi.org/10.1080/10409280902858764)
- Pianta, R. C. (2008). Building capacity for positive youth development in secondary school classrooms: Changing teachers' interactions with students. In *Toward positive youth development: Transforming schools and community programs*. Oxford University Press.
- Pianta, R. C., Hamre, B. K., Hayes, N., Mintz, S., & LaParo, K. M. (2008). *Classroom Assessment Scoring System–Secondary (CLASS-S)*. University of Virginia. Charlottesville, VA.
- Plank, S. B., & Condliffe, B. (2013). Pressures of the Season: An Examination of Classroom Quality and High-Stakes Accountability. *American Educational Research Journal*, 50(5), 1152–1182. <https://doi.org/10.3102/0002831213500691>
- Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for

studying instructional quality: Looking back and looking forward. *ZDM*, 50(3), 535–553.

DOI:10.1007/s11858-018-0946-0

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

Praetorius, A.-K., Rogh, W., Bell, C., & Klieme, E. (2019). Methodological Challenges in Conducting International Research on Teaching Quality Using Standardized Observations. In *The SAGE Handbook of Comparative Studies in Education*. SAGE Publications Ltd. DOI:10.4135/9781526470379

Reddy, L. A., Hua, A., Dudek, C. M., Kettler, R. J., Lekwa, A., Arnold-Berkovits, I., & Crouse, K. (2019). Use of observational measures to predict student achievement. *Studies in Educational Evaluation*, 62, 197–208. DOI:10.1016/j.stueduc.2019.05.001

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM*, 48(1), 29–40. DOI:10.1007/s11858-016-0765-0

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific Research in Education*. National Academy Press. <https://doi.org/10.17226/10236>

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science*, 22(11), 1359–1366. DOI: 10.1177/0956797611417632

Stein, M. K., Correnti, R., Moore, D., Russell, J. L., & Kelly, K. (2017). Using Theory and Measurement to Sharpen Conceptualizations of Mathematics Teaching in the Common Core Era. *AERA Open*, 3(1). DOI:10.1177/2332858416680566

Steinberg, M. P., & Sartain, L. (2015). Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*, 10(4), 535–572. DOI: 10.1162/EDFP_a_001738.

- Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction*, 46, 45–60. DOI: [10.1016/j.learninstruc.2016.08.003](https://doi.org/10.1016/j.learninstruc.2016.08.003)
- van de Grift, W. J. C. M., Houtveen, T. A. M., Hurk, H. T. G. van den, & Terpstra, O. (2019). Measuring teaching skills in elementary education using the Rasch model. *School Effectiveness and School Improvement*, 30(4), 455–486. 10.1080/09243453.2019.1577743
- van de Grift, W. J. C. M. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement*, 25(3), 295–311. DOI:[10.1080/09243453.2013.794845](https://doi.org/10.1080/09243453.2013.794845)
- van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2018). Developing an Instrument for Teacher Feedback: Using the Rasch Model to Explore Teachers' Development of Effective Teaching Strategies and Behaviors. *The Journal of Experimental Education*, 86(2), 247–264. DOI: 10.1080/00220973.2016.1268086
- Virtanen, T. E., Pakarinen, E., Lerkkanen, M.-K., Poikkeus, A.-M., Siekkinen, M., & Nurmi, J.-E. (2018). A Validation Study of Classroom Assessment Scoring System–Secondary in the Finnish School Context. *The Journal of Early Adolescence*, 38(6), 849–880. DOI:[10.1177/0272431617699944](https://doi.org/10.1177/0272431617699944)
- White, M. (2021a). A Validity Framework for the Design and Analysis of Studies using Standardized Observation Systems. *Analysing Teaching Quality: Perspectives, Principles and Pitfalls*. Scandinavian University Press.
- White, M. (2021b). Considerations on Schools' Use of Observation Scores from Teacher Evaluation Systems. University of Oslo.
- White, M., Maher, B., & Rowan, B. (2021a). Teaching Practices Before and After the Common Core: 2010-2018 Changes in the Reading and Writing Instruction of Experienced Teachers in Six Urban School Districts. University of Michigan.
- White, M., Maher, B., & Rowan, B. (2021b). A Framework for addressing Instrumentation

Biases when using Observation Systems as Outcome Measures in Instructional Interventions. University of Michigan.

- Wylie, E. C. (2020). Observing Formative Assessment Practice: Learning Lessons Through Validation. *Educational Assessment*, 25(4), 251–258. DOI:10.1080/10627197.2020.1766955
- Xu, L., & Clarke, D. (2019). Speaking or not speaking as a cultural practice: analysis of mathematics classroom discourse in Shanghai, Seoul, and Melbourne. *Educational studies in mathematics*, 1-20. doi:10.1007/s10649-019-09901-x