

Article

Observing Instructional Practice: Can We Consistently Measure Teaching Quality Constructs?

Mark White ^{1,*} , Armin Jentsch ² , Jennifer Luoto ³ and Kirsti Klette ¹

¹ Department of Teacher Education and School Research, University of Oslo, NO-0316 Oslo, Norway; kirsti.klette@ils.uio.no

² Department of Educational and Social Sciences, University of Cologne, 50923 Cologne, Germany; ajentsc2@uni-koeln.de

³ Department for Teacher Education and School Research, Yorkville University, Fredericton, NB E3B 0H2, Canada

* Correspondence: mark.white@ils.uio.no

Abstract

Classroom observation systems can be an important tool for understanding teaching quality. However, the wide range of possible lessons that can be observed raises concerns about whether fixed observation rubrics can measure the intended teaching quality constructs equally well in each lesson. This paper argues for exploring measurement invariance across lessons in observation systems, adopting an understanding of measurement invariance that emphasises the alignment between a theoretical construct and the measurement of that construct. We conceptualise teaching quality using the Protocol for Language Arts Teaching Observation (PLATO). PLATO's focus on individual items is misaligned to traditional measurement invariance approaches, so we emphasise the importance of detailed, qualitative considerations of how rubric operationalisations of a construct may or may not capture the intended teaching quality construct equally well across lessons with different characteristics. We discuss the affordances and limitations of this way of considering measurement invariance and argue for the importance of ensuring measurement invariance across lessons.

Keywords: measurement invariance; teaching quality; classroom observation; observation system

1. Introduction

Educational research must address the challenges of directly studying classroom instructional practices (Ball & Forzani, 2007), as classroom instructional practice is an important mediator of student outcomes and the core of how schools support student learning and development (Hamre et al., 2013). Developing an accurate, empirical understanding of teaching quality, or the components of classroom interactions that support student learning and development, requires its careful measurement, both within individual studies and across studies to facilitate the accumulation of knowledge about teaching quality (Klette, 2020).

There is a growing interest in formally conceptualising, defining, and operationalising specific teaching quality constructs to facilitate their systematic measurement and study (Klette, 2020). This interest has coalesced in observation systems (Bell et al., 2019), which combine manuals that operationalise teaching quality constructs and the broader systems



Academic Editor: Eila Jeronen

Received: 5 February 2026

Revised: 27 March 2026

Accepted: 31 March 2026

Published: 7 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

necessary to consistently measure those constructs. There is an enormous challenge in creating such observation rubrics, which can be understood by comparison with constructed response scoring rubrics. Constructed response rubrics are typically tailored to specific test questions, so that rubrics can be carefully designed to consider the full range of possible answers that might be scored (McClellan, 2010). However, observation rubrics are typically applied to lessons with a wide range of instructional features (e.g., content areas, activity formats, learning goals, student populations, and other factors that influence the nature and interpretation of instructional interactions). The diversity of typical instructional practice necessitates this breadth, since rubrics could not be developed for each individual lesson and observation systems often seek to characterise the range of instructional practice. The breadth of instructional practice that rubrics are applied to creates the real possibility that the scores may not always capture the intended teaching quality constructs equally well (i.e., measurement non-invariance; Engelhard & Wind, 2018). This would result in measurement quality being dependent on lessons in ways that could be quite difficult to measure and/or model, which could in turn lead study conclusions to be sample-dependent.

This paper argues that the difficulty in capturing the full breadth of instructional practice, which is inherent in building rubrics, generates the need to attend to measurement invariance (MI) issues at the level of lessons (i.e., the relationship between scores and the intended construct may vary across lessons; Fischer et al., 2025). MI issues are a common concern in cross-cultural research. For example, research has highlighted how some cultures engage in practices, such as choral responding, which might lead some rubric indicators (e.g., ‘the percentage of students talking’) to be poor measures of the construct of discourse (Xu & Clarke, 2018). We extend this concern to argue that the variation in instructional practice across lessons could lead to MI concerns at the between-lesson (and within-culture) level. For example, the construct of classroom discourse may manifest in different ways in or be differentially applicable to lessons that are organised as whole-class and small-group settings. Invariantly measuring the construct of classroom discourse requires observation manuals to address these differences in how the construct manifests and the construct’s applicability in different activity structures. Ensuring that observation manuals attend to such complexities is necessary to ensure that MI exists in scores, which is arguably a fundamental condition for measurement (Engelhard & Wind, 2018). To our knowledge, MI has never been formally explored at the lesson level. Due to this, the paper aims to make conceptual and methodological contributions. Further, we take a broad perspective on MI, viewing it as extending far beyond the psychometric analyses used to test for it, to include all aspects of ensuring that constructs are measured consistently across measurement occasions. This expanded view is consistent with more modern conceptions of MI and between-group bias in measurement (e.g., American Educational Research Association et al., 2014; Fischer et al., 2025).

This paper begins by introducing a conceptual framework that frames our thinking about MI, drawing on M. Kane’s (2013) validity theory. We turn then to a discussion of our conceptualisation of teaching quality, which is operationalised through the protocol for language arts teaching observation (PLATO; Grossman et al., 2014). Using this conceptualisation, we review standard approaches to capture MI, highlighting the fit (or lack thereof) of these approaches to our conceptualisation of teaching quality. This review leads us to explore alternative ways of understanding MI when trying to study theoretically and empirically meaningful dimensions of teaching quality. We end by discussing how this work informs other efforts to use observation systems.

2. Conceptual Framework

This paper is grounded in modern validity theory (M. Kane, 2013) and the *Standards for Educational Testing and Assessment* (American Educational Research Association et al., 2014). From this perspective, tests and other measures are first interpreted as representing an intended construct and so become useful for a given purpose through this interpretation (i.e., interpretation for use). Tests and measures are not directly validated, but interpretations of tests/measures are validated, along with the extent to which the interpretation supports a proposed use and the consequences of that use. Most uses of observation scores arguably require them to be interpreted as capturing the pre-defined construct of teaching quality (e.g., White et al., 2022). Therefore, the extent to which scores can be interpreted as representing the intended construct is highly relevant for most uses of observation scores.

To explore how well scores might represent the intended construct, this paper adopts a lens model framework (Brunswik, 1952; Engelhard & Wind, 2018) that is informed by McGrane and Maul's (2020) perspectives on measurement. Measurement is composed of three inter-linked and overlapping models: the substantive model, the data model, and the statistical model (McGrane & Maul, 2020). The substantive model provides a conceptualisation of the teaching quality construct, defining how it impacts (or manifests in) enacted instruction (see left side of rectangle in Figure 1). The data model describes how enacted instruction is transformed into a set of raw scores (see the top of the rectangle in Figure 1). Importantly, in observational tools, this occurs through indicators in the observation manual, which instruct raters on how observable features of enacted instruction correspond to different raw scores. The statistical model transforms raw scores into a summary score that is meant to represent the conceptualised construct (see right side of Figure 1). When measurement is successful, the summary score represents the conceptualised teaching quality construct (Markus & Borsboom, 2013; see dotted line at the bottom of Figure 1). Importantly, summary scores are interpreted as representing the pre-defined construct. Consistent with modern views, we define MI as the extent to which the summary scores represent the intended construct across measurement occasions equally well (here, across different lessons: Fischer et al., 2025; Engelhard & Wind, 2018). When MI exists, the summary scores represent the conceptualised construct equally well across lessons. When MI does not exist, the summary scores do not represent the conceptualised construct equally well across lessons, raising questions about whether the scores can be interpreted as representing the conceptualised construct and may lead results to be sample-dependent.

MI is a problem that exists at the intersection of the substantive model and the data model. The substantive model must correctly specify how constructs manifest in different types of lessons that may be encountered. The data model, as instantiated through rubric indicators that direct raters' attention, must correctly direct raters' attention to concrete, observable instructional interactions that reflect the levels of the target construct in enacted instruction. Ambiguity or a lack of alignment in these processes for some observed lessons will result in scores being driven by observed behaviours that do not appropriately reflect the target construct for some types of lessons, creating non-MI for those lesson types.

The difficulty of building manuals to apply to a wide range of instructional practices creates an enormous challenge in building rubric indicators. The complexity of teaching quality constructs combined with the practical demands that arise from building manuals that raters can reliably use under real-life data collection constraints creates further challenges. As a result, the operationalised manual may not always correctly map the enacted level of a construct to the raw scores. For example, the dimension of cognitive engagement may conceptually be focused on students' active cognitive processing (i.e., the brain state of a student), but manual indicators may focus on outward appearances of students' cognitive processing, due to practical measurement limitations (i.e., one cannot

measure brain states). This raises the important question of whether the manual indicators align with the conceptualised construct equally well across lessons. For example, manual indicators may lead scores to represent the cognitive engagement construct well when students are actively discussing content, but scores may not represent cognitive engagement as well when students passively listen to a lecture. That is, one can be highly certain that a student who is engaged in discussion is actively processing information (i.e., scores align well with the intended construct), but a student who is outwardly listening to a lecture may or may not be actively processing content (i.e., scores do not closely align with the intended construct).

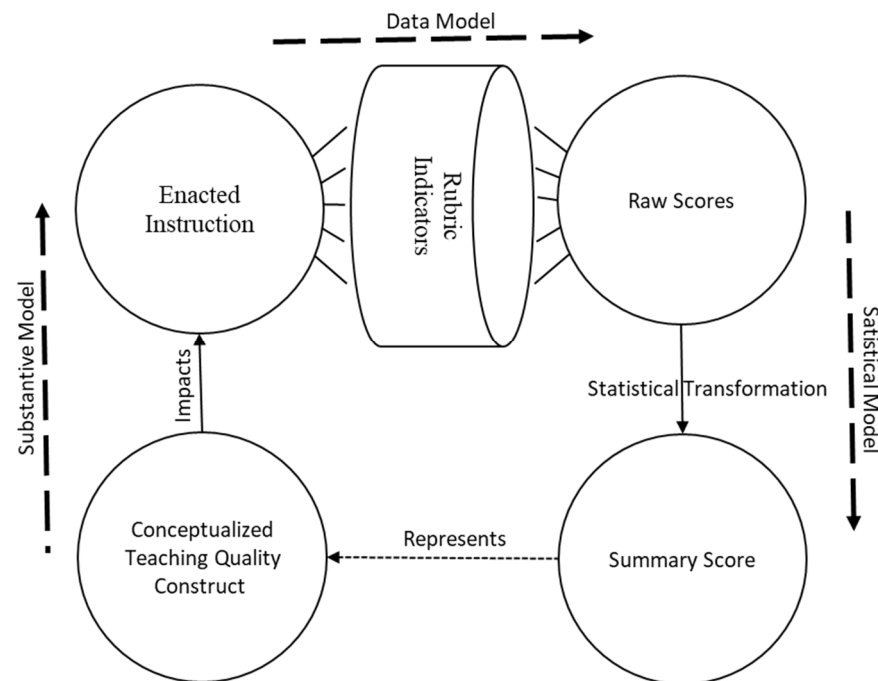


Figure 1. Three models capturing measurement.

This is a question of measurement invariance (MI; Engelhard & Wind, 2018). MI requires that the manual indicators (i.e., the source of scores) lead to scores that represent the intended construct (approximately) equivalently across measurements. This is a challenging requirement, especially when teaching quality constructs relate to individuals' internal states (such as for cognitive engagement), since manual indicators must be observable (i.e., reflect external manifestations of those internal states). Observable indicators may or may not invariantly represent the internal states across measurements.

Since MI involves the correspondence between scores and conceptualised constructs (Fischer et al., 2025), the way that teaching quality is conceptualised plays a vital role in exploring MI. Due to this, we turn now to describing our understanding of teaching quality, as operationalised through PLATO. This allows us to ground later discussions of MI in our specific understanding of teaching quality.

3. Conceptualising Teaching Quality Through PLATO

This paper arises from broader work that used the Protocol for Language Arts Teaching Observation (PLATO; Grossman et al., 2014) to conceptualise teaching quality (Klette et al., 2017, 2022). PLATO was chosen to be used for systematic coding due to both researchers' familiarity with PLATO and the belief that PLATO captured important and culturally appropriate aspects of instructional practice while supporting supplemental qualitative coding of instruction. Importantly, PLATO's focus on capturing specific discrete practices

related to students' engagement with content (Grossman et al., 2014) was a key factor in its adoption. Note that while we believe that our interpretation of PLATO is generally consistent with those of the developers, the presented conceptualisation of teaching quality should be understood as our own, not as that of the PLATO developers.

The PLATO manual was originally developed to capture instructional practices that are important for learning in secondary language arts classrooms (Grossman et al., 2014), but it has since been used across subjects as a tool that captures important aspects of instruction (Cohen, 2018). PLATO contains four domains (Representation of Content, Disciplinary Demand, Management, and Instructional Scaffolding) and 12 elements. In contrast to other manuals that focus on capturing instructional quality broadly, PLATO arguably has a stronger focus on capturing specific, discrete practices that are understood as being highly important, rather than capturing the broader construct of teaching quality. That is, the focus of PLATO is on the 12 elements. The domain structure is less important and serves largely to group similar elements. This means that teaching quality constructs are defined at the item level.

This paper places a large emphasis on PLATO because research has highlighted the interdependence of construct conceptualisation and MI concerns (Fischer et al., 2025), making it necessary to consider how constructs are conceptualised when considering MI. However, PLATO serves simply as a case study to explore deeper issues of between-lesson MI. Similar points would apply to a wide range of observation systems. The fundamental challenges here are building a rubric that works well across the full range of observed instruction and operationalising complex teaching quality constructs, which are universal challenges across observation manuals. Further, we emphasise that, despite the challenges in establishing valid and reliable observation manuals, we believe that observation manuals can play an important role in improving our understanding of teaching quality. To achieve this goal, though, will require paying careful attention to the quality of measurement across lessons and instructional constructs, which is the challenge taken up by this paper.

4. Standard Approaches to Measurement Invariance (MI)

Measurement invariance, definitionally, is the property that the construct being measured remains the same across measurements (i.e., at every observation; Engelhard & Wind, 2018). MI is an important aspect of the fairness of tests in the AERA testing standards (American Educational Research Association et al., 2014). Applied here, MI requires that observation scores represent the pre-defined construct equally well across all observed lessons. That is, the indicators that manuals use to characterise specific levels of quality within each element correspond to the targeted level of quality on the intended construct in each observed instance of instruction.

Standard quantitative approaches to testing for MI depend on the use of a reflective measurement model (e.g., factor analysis, item response theory; Engelhard & Wind, 2018). Reflective measurement models work under the principle of repeated measurement, such that the sole source of shared variation across items (in PLATO, elements) is the intended construct (Markus & Borsboom, 2013; White et al., 2024; White, 2025). This allows the fitting of reflective measurement models that can examine whether the same construct is being measured across settings by examining patterns of covariation between items. These approaches are a problematic fit for our understanding of measuring instruction with PLATO.

Applied to PLATO, the standard MI logic would suggest that PLATO only captures the same construct across measurements when the covariance between PLATO elements (e.g., modelling and strategy use and instruction) is the same across measurements (or fits the same model-implied structure). However, as discussed above, our conceptual

understanding of PLATO emphasises that each element uniquely captures important features of teaching and teachers can choose amongst different approaches to achieve their goals. Then, teachers may choose to blend strategy instruction with modelling or use these instructional approaches separately, leading to different correlations between these elements across measurements without impacting our ability to understand PLATO scores as representing the teaching quality constructs. Rather, differences in element score correlations across lessons would merely provide an interesting characterisation of how teaching unfolds differently across lessons.

In fact, given our conceptual framework, we are focused on MI in individual element scores (i.e., for individual items, rather than groups of items combined into scales). This is because each of the PLATO elements was developed with the goal of capturing instructional interactions that are uniquely supportive of student learning (Grossman et al., 2013). The usefulness of each element is not understood as being dependent on its relationship to other elements but is dependent on the fact that the element itself captures specific instructional interactions that are thought to support student learning. Then, we are interested in the ways that enacted instruction (i.e., the object to be scored) is mapped to performance levels that represent the pre-defined construct (i.e., the second mapping of Fischer et al., 2025) or the data model in McGrane and Maul (2020).

It is interesting to point out here that typical uses of observation frameworks do not measure instruction in a single lesson, but rather aggregate across lessons to generate reliable scores to characterise a classroom (Gleason et al., 2017). The view of MI in this paper has important implications for such aggregation because aggregating may lead to little or no improvement in reliability if MI is not present. For example, take the previous example of cognitive engagement. We noted how cognitive engagement scores represent the construct in discussions but may not represent the construct in lectures, where students' state of engagement is not visible to observers, equally well (i.e., levels of error may be very high in lectures). Assume that this is true and the scores from lectures have very high levels of error. In such a situation, observing a single lesson with discussions may give a more reliable measure of cognitive engagement than observing one lesson with discussions and one or more lectures, as the lectures may do little more than contribute measurement error.

5. Exploring Measurement Invariance in Observational Measures of Instruction

This review of MI approaches leads us to conduct more qualitative explorations of how well PLATO scores represent the intended constructs across lessons. Due to space constraints, we focus on examples from two PLATO elements that we believe highlight important challenges and considerations when examining how well one might support the inference that PLATO scores represent the intended constructs. In each section, we highlight the teaching quality constructs that the targeted element intends to measure, the PLATO manual indicators (i.e., surface-level features), and evaluate the degree to which scores represent the intended construct invariantly across different potential types of lessons.

We know of no other papers that discuss issues of MI across lessons, nor suggestions that a lack of MI across lessons would be an important challenge for observational measures of teaching. Due to this, we did not originally plan the study to systematically explore between-lesson MI in PLATO. Rather, questions about MI across lessons arose within broader discussions between the authors regarding MI across subjects. We therefore do not have systematic information about MI across lessons but rather provide exemplars that are meant to raise awareness of this potential threat and show us how to consider the threat. This makes the paper more of a conceptual and methodological contribution, rather than an empirical contribution. The exemplars were discussed and verified by all authors, ensuring

that the points raised relate to our understanding of PLATO and not errors in individual raters' scoring. Examples are drawn from our experiences in coding classroom instruction, but do not necessarily represent any specific event that occurred in a specific lesson. Rather, they represent categories of instructional events that stimulated discussions in our author group (e.g., if one-on-one discussions were counted as discourse). Through presenting these examples, we hope to demonstrate the importance of exploring MI at the lesson level, since MI must exist to draw strong conclusions about theoretical teaching quality constructs, according to standard measurement theory ([American Educational Research Association et al., 2014](#); [Fischer et al., 2025](#); [Praetorius et al., 2019](#)).

5.1. Classroom Discourse

This section discusses possible MI challenges in the PLATO element classroom discourse. This example arises from internal discussions in our project team about whether PLATO measures uptake equally well in both one-on-one discussions and whole-class instructional formats. Importantly, we determined at the end that there were not MI concerns here, given the alignment of conceptualisation and measurement, but we believe that the example is still useful, as it shows the sorts of careful thought processes around MI issues that we are promoting in this paper.

5.1.1. Conceptualised Teaching Quality Element

This element is based on research on productive discourse and assesses the opportunities that students have to engage in content-related conversations, as well as the nature of the conversations: whether perfunctory and minimal or rich and elaborated ([Grossman et al., 2013](#)). Classroom discourse is broken down into two sub-elements, uptake of student responses (uptake) and opportunities for student talk. We focus here on uptake, as it provides a clearer example of potential MI challenges. Importantly, PLATO explicitly defines uptake as something that occurs during discourse and defines discourse as occurring only when two or more students are involved ([Grossman et al., 2014](#)).

The uptake component is grounded in research on classroom discussions that can promote learning (e.g., [Nystrand, 1996](#)). In other words, it is hypothesised that teachers can support student learning through discursive moves that clarify student thinking and connect student ideas to each other, to the content at hand, and to academic terminology ([Alexander, 2008](#)). Research in classroom discourse has identified teacher moves that are especially beneficial to engage students in discourse, such as revoicing, taking up student ideas, asking for clarification, and justification ([Nystrand, 1996](#); [Cazden, 2001](#); [Wells, 1999](#)). The focus on such uptake moves reflects PLATO's emphasis on explicit instruction, as such uptake moves make student ideas accessible to other students, allowing students to learn from the thinking and reasoning of others (e.g., [Anthony & Walshaw, 2009](#); [Mercer & Sams, 2006](#)). The uptake component is meant to capture the extent and frequency with which these moves occur.

5.1.2. Manual Indicators

The PLATO manual directs raters to focus on any examples of uptake (e.g., revoicing, elaboration, asking for clarification or justification) that occur within the context of discourse (i.e., talk between two or more students [and possibly the teacher]), regardless of who engaged in the uptake. Uptake scores are based on the quality and consistency of these observed uptake practices with more consistent, high-quality uptake, indicating higher scores.

5.1.3. The Degree to Which Scores Represent the Conceptualised Element

PLATO is quite clear and consistent in how they define discourse and how they operationalise uptake, scoring only uptake moves that occur within the context of discourse. Then, the intended construct of uptake seems to be correctly and consistently measured in PLATO (i.e., measurement invariance exists), both in group discussions (where uptake can occur) and in one-on-one conversations (where uptake cannot occur, according to PLATO).

It is important to note here, though, that uptake, like most constructs in education, is contested (Blikstad-Balas, 2014). Other conceptualisations of uptake or observation manuals may argue for a different understanding of uptake that includes uptake moves in one-on-one conversations with students, since uptake moves could help individual students to refine and clarify their own ideas by linking those ideas to academic content and terminology and/or challenging students to elaborate and expand on their thinking (e.g., Alexander, 2008). For example, such one-on-one conversations are common in math instruction, so may be important to capture (Luoto et al., 2022). This fact does not affect PLATO's measurement of its own conceptualisation of uptake, but it does raise an important point to consider for comparisons of uptake scores across studies and manuals.

5.2. Representation of Content

This section discusses possible MI challenges in the PLATO element of representation of content (ROC). This example arose out of discussions for how to score cases in which students provide clear representations of content at the behest of teachers.

5.2.1. Conceptualised Teaching Quality Element

Representation of content (ROC) focuses on the (1) quality and (2) conceptual richness of the explanations the teacher provides to students (Grossman et al., 2014). This involves the accuracy and clarity with which the teacher presents content (e.g., concepts, definitions, strategies, examples). The literature suggests that the quality of explanations is a relevant aspect of teaching quality, as it is closely connected to students' conceptual understanding (Learning Mathematics for Teaching Project, 2011; Lipowsky et al., 2009; Schlesinger & Jentsch, 2016). PLATO focuses specifically on the 'teacher's ability and accuracy in representing ELA content' (Grossman, 2015).

5.2.2. Manual Indicators

The PLATO manual is divided into two sub-elements. The Quality of Instructional Explanations sub-element focuses on the degree to which teachers provide examples, analogies, or explanations that are accurate and clear while addressing student misunderstandings and the nuances of the content. Higher scores are for more elaborated and nuanced explanations, and lower scores are for superficial or incorrect explanations. The conceptual richness of instructional explanations sub-element focuses on the extent to which explanations are focused on building students' conceptual understanding of ELA content. Higher scores focus more on building the conceptual understanding of content (e.g., connecting different concepts and ideas, focusing on interpretations), while lower scores are for instruction focused on rules, procedures, and labels. As one might notice from the names of the two sub-elements, the PLATO manual focuses on verbal 'explanations' given by the teacher, and ways that students represent content are explicitly excluded from counting towards PLATO scores.

5.2.3. The Degree to Which Scores Represent the Conceptualised Element

In applying PLATO to a range of lessons, we noted several occasions where PLATO scores seemed to poorly capture how teachers represented content. For example, imagine

a lesson where students work together on problems in small groups while the teacher circulates and provides guidance and feedback. The teacher might then pick several groups to present their work to the class, ensuring that multiple approaches to solving the problem are presented, and highlighting specific aspects of solutions through questioning. Throughout this process, the teacher provides few explicit verbal explanations of content, so PLATO ROC scores would be low. However, we would argue that teachers can represent content a great deal in such lessons (through their choice of problems, choice of who will present their work, and questioning).

These more indirect ways of representing content seem to fit PLATO's definition of ROC, which emphasises the teacher's ability to represent content effectively and clearly, but PLATO scores do not capture this indirect representation of content. Then, PLATO scores may fail to represent the intended construct in cases where the teacher is representing content indirectly (e.g., through structuring opportunities for students to engage with content). This is, we believe, a straightforward case of measurement non-invariance (i.e., PLATO scores do not represent the intended construct in lessons where teachers directly and verbally represent content and lessons where teachers indirectly represent content equally well).

Importantly, PLATO has a strong focus on explicit instruction (e.g., [Cohen, 2018](#); [Klette et al., 2022](#)), so the operationalisation of ROC fits with the broader manual. Further, it is likely to be quite challenging to observationally capture the ways that teachers have planned to indirectly represent content, which is to say that this may be an explicit choice by PLATO to minimise the scoring errors and ambiguities that would arise from trying to capture indirect teacher representations. Regardless of the origin of this non-invariance, PLATO scores may not invariantly represent the intended construct across lessons. Two solutions come to mind here. The first would be to narrow the intended construct from 'teacher's ability to represent content' to 'teacher's direct, verbal representations of content'. While this seems like a minor shift, it can have important implications for how PLATO is used, how scores are interpreted, and how PLATO constructs relate to other constructs and student learning. The second solution would be to expand the PLATO manual to capture teachers' indirect representations, although finding ways to capture all of these indirect representations may be extremely difficult.

5.3. The Problem of Double-Barrelled Items

In applying PLATO to our context and exploring the extent to which the PLATO scores seemed to invariantly represent the intended constructs, a class of challenges arose. Namely, the PLATO manual often captures a range of related ideas within the same element. This has been noted as a problem more broadly with observation manuals, as many manuals capture a blending of frequency and quality ([Praetorius & Charalambous, 2018](#)). The challenge here can be understood using the example of double-barrelled survey items, a more familiar problem ([Gehlbach, 2015](#)). An example of a double-barrelled survey item is 'My teacher is nice and helpful'. Here, respondents provide an answer that blends information about both how nice their teacher is and how helpful they are.

Double-barrelled items are typically understood from the perspective of multidimensionality ([Gehlbach, 2015](#)). For example, the previously discussed item would measure both the construct of niceness and helpfulness. Here, we explicitly connect multidimensionality to MI. In a classroom where a teacher is super nice, but only modestly helpful, the survey item may mostly reflect the teacher's niceness, but the same item may mostly reflect the teacher's helpfulness in classrooms where the teacher is super helpful, but modestly nice. That is, when items are double-barrelled and so multidimensional, the relationship between scores on the item and the construct being measured can be non-constant, since scores could

be driven by multiple different constructs. In multi-item scales, this limitation could potentially be overcome by modelling item scores as being multi-dimensional (though in practice, this is rarely done for observation systems), but when single items are considered, there is no way to remove the confounding of the two constructs. For some interpretations/uses of answers, all dimensions composing an item may have the same relationship with criterion variables, and so the double-barrelled nature of items may have limited practical impact, though it still contributes to conceptual ambiguity. However, for interpretations/uses of scores where the different dimensions being measured may not be functionally equivalent, the lack of MI that can come from different lessons reflecting the two measured dimensions to different extents can create challenges.

Several PLATO elements contain this sort of double-barrelled scoring. For example, the modelling element includes both modelling (i.e., demonstrations of how to do the task at hand) and use of models (i.e., teachers presenting completed examples of the task at hand). Scores can be based on either modelling or the use of models. There is, then, an implicit assumption that the process of modelling and the use of models are functionally equivalent (i.e., that the process of modelling and the use of models are effectively the same construct). To the extent that these are not the same construct, then the item is confounding multiple constructs in the same score without any ability to remove the confound. In this case, and since teachers typically either model or use models, but not both, scores in different lessons will alternatively reflect only either modelling or using models (i.e., the relationship between scores and constructs is non-constant), creating a lack of MI.

The same challenge occurs with strategy use instruction (SUI), which includes the prompting of strategies and the explicit teaching of strategies (though this problem is isolated to the score category of two in SUI). In each of these elements, two distinct concepts are blended to provide a single score to characterise the element. We might also note here that this problem is even more prevalent in other manuals, such as the Classroom Assessment Scoring System (CLASS; Hamre et al., 2013), which includes several discrete indicators that capture different aspects of the item construct within each item.

There is a real question of when and where this ‘double-barrelled’-ness of elements would have practical implications. This likely depends on the intended interpretation/use of scores. Recently, PLATO has mostly been used to descriptively characterise the nature of teaching quality and to select lessons for deeper qualitative inquiry (e.g., Christensen & Mathé, 2023; Nissen, 2023; Sigurjónsson et al., 2022). For this use, the double-barrelled-ness of elements adds to the coarseness of measurement but does not appear to impact the validity of conclusions. However, if one were focused on empirically linking PLATO scores to measures of student learning, this double-barrelled-ness could create problems, especially if the two ‘barrels’ had different impacts on student learning (e.g., if the impact of ‘modelling’ and ‘using models’ on student learning differed). Namely, the double-barrelled-ness would create unmodelled variation in the relationship between PLATO scores and student learning measures that reduces the replicability of results and creates conceptual confusion about which teaching practices support student learning. Then, the double-barrelled-ness of some elements could be understood as a challenge related to the non-invariance in the PLATO manual. The solution to this challenge would be to collect scores for each “barrel”, allowing for empirical testing of the impacts of double-barrelled items.

6. Discussion

This paper sought to raise and discuss fundamental questions about MI in observation manuals. Observation manuals painstakingly define specific constructs related to teaching quality and measure those constructs, with uses of observation manuals typically requiring

interpreting scores as representing those constructs (White et al., 2022). This makes the interpretation of observation scores as capturing the intended construct an important aspect of most uses of observation manuals, so they are relevant to consider (American Educational Research Association et al., 2014; M. Kane, 2013). Since manuals are used across a wide range of lessons and teaching quality constructs are often complex, there is a risk that the manual indicators developed to operationalise the teaching quality constructs may not reflect the intended construct equally well across lessons (i.e., there is a need to explore the second mapping from Fischer et al., 2025). This is, in fact, the likely cause of the non-invariance issues discussed in this article. Namely, the complexity of instruction leads manual indicators to be built with certain assumptions about the type of instruction that might be observed (e.g., that teachers will directly represent content). When these assumptions are violated, manual indicators fail to capture the intended construct. There is a need to systematically study the measurement (non-)invariance of observation manuals across lessons with different characteristics (e.g., subjects, activity formats).

This is not to deny the usefulness or importance of observation manuals. The explicit operationalisation of teaching quality constructs in manuals helps us to concretise what can be abstract constructs and supports the systematic accumulation of evidence on the nature and impact of instructional practice. Such attempts at systematic measurement are necessary to advance our understanding of instruction. However, the full benefits of observation manuals will never be realised without frank discussions about the difficulties of measuring instructional constructs reliably and consistently across lessons, which this paper is meant to inspire. Measurement non-invariance represents a threat to our understanding of instructional constructs, necessitating direct explorations of the extent to which our operationalisations allow for the invariant measurement of the intended construct.

We used the PLATO manual as an example to discuss this challenge because of the authors' familiarity with this manual. In our discussion, like other recent studies, we rejected traditional approaches to MI (e.g., Welzel et al., 2022; Welzel & Inglehart, 2016; White, 2025). In our case, this was because of the lack of fit of reflective models to our conceptualisation of teaching quality and the lack of existing methods for considering MI at the item (here, element) level (c.f., White et al., 2024).

This led us to focus on more conceptual and qualitative explorations that identified several examples of apparent measurement non-invariance across lessons to which PLATO was applied. In the uptake example, we concluded that PLATO would likely measure the intended construct invariantly in one-on-one discussions and larger discussions, based on its own definition of uptake. In the ROC example, we concluded that PLATO would likely not invariantly measure ROC across lessons where teachers represented content directly versus those where content was represented indirectly. We also noted several examples where the PLATO manual contains double-barrelled elements, the implications of which will likely vary across interpretations/uses of scores. These results raise important questions about where PLATO can validly be used. The similarity of PLATO to other observation systems suggests that these concerns are likely to impact other observation systems (e.g., Bell et al., 2019).

An important question to consider here is how much the identified non-invariance matters, and for what purposes. Standard MI approaches have established ways of examining the size of the non-invariance, which has important implications for its practical impact (e.g., Meade, 2010). There are no established methods for the sort of argumentation modelled here (see Section 6.4 on next steps for establishing such methods).

In addition to determining the size of MI violations, the impact of the measurement non-invariance discussed here will likely depend strongly on the research question. Different interpretations of scores will be more or less impacted by MI issues and interpretations

of scores will depend heavily on the use to which scores are put ([American Educational Research Association et al., 2014](#); [M. Kane, 2013](#)). For example, problems with the ROC element would be especially problematic when comparing a curriculum based on traditional lectures with a curriculum based on problem-based learning, since the second curriculum is probably more likely to have teachers indirectly represent content. The element may be less problematic, though, when comparing two lecture-based curricula. Thus, the impact of identified non-MI will likely depend on the distribution of the lesson features and how this distribution interacts with the research questions.

6.1. Generalisability of Findings to Other Manuals

Since PLATO contains many of the same features as other observation manuals, these results point towards the potential for there to be a lack of MI in other observational approaches to measure teaching quality. There is no reason to think that the problems discussed here are unique to PLATO. For example, it has been pointed out that many observation manuals conflate the measurement of quality and frequency, making manual items double-barrelled ([Praetorius & Charalambous, 2018](#)). The sorts of analyses discussed in this paper, then, would be broadly applicable to understand MI in other observation systems, as well as other measures of teaching quality ([White et al., 2024](#)).

PLATO is widely used in both the US and several Nordic countries ([T. J. Kane et al., 2012](#); [Klette et al., 2017](#); [Tengberg et al., 2022](#)). Additionally, generalisability studies ([Cor, 2011](#); [Jentsch & Klette, 2024](#)) and measurement invariance studies examining PLATO across subjects ([Jentsch et al., 2026](#)) support that PLATO can validly measure teaching quality in a Norwegian and Nordic context (see also [Blikstad-Balas et al., 2022](#); [Klette et al., 2026](#)), for which there are many qualitative analyses that confirm PLATO scores and demonstrate positive uses of PLATO. Therefore, the discussion of PLATO in and of itself is beneficial, above and beyond the main focus of using PLATO to discuss wider issues in the field.

6.2. Building a Complete Measurement of Instruction

There is a broader point following on from this work. As discussed above, the measurement process consists of at least three models: the substantive model that details how the hypothesised teaching quality construct manifests in instruction; the data model that details how the manifestation of the construct maps to specific raw data/scores, and the statistical model that maps the raw data back to levels of the theoretical construct (see Figure 1; [McGrane & Maul, 2020](#)). Most current efforts to observationally measure teaching quality do not provide clear and distinct substantive and data models. Rather, they conflate these models, which conflates theoretical conceptualisations and the measurement of teaching quality constructs. This arguably is one factor contributing to the difficulty in determining how much similar constructs in different observation manuals overlap ([Charalambous & Praetorius, 2020](#)). For example, the overlap between PLATO's uptake construct and how uptake is conceptualised in other frameworks can be unclear.

Much of the work underlying this paper has stemmed from our efforts to distinguish the substantive model that defines how the conceptualised construct of teaching quality (i.e., the intended construct) manifests across different lessons and the data model that defines how one can develop manual indicators to measure that construct. For example, our discussion of the element of uptake really revolved around how uptake was substantively conceptualised. Namely, is uptake about specific teacher moves that support students in refining their ideas? Or is uptake a socio-cultural phenomenon related to direct instruction that occurs when teachers enact specific moves within groups of students? Or both? This distinction not only has important implications for our understanding of uptake as a

construct that supports student learning, but it has important implications for the scoring rules used by the PLATO manual and the quality of PLATO scores.

There is, of course, no one right answer to how one should conceptualise uptake. When the goal of a project is to understand teacher moves, uptake might be defined as capturing verbal teacher statements, while, when the goal is to connect measures of teaching and learning, uptake might rather be conceptualised by students' response to instances of uptake (i.e., the success of the teacher move; e.g., do students adopt academic language after a teacher links their language to more academic language; [Hiebert & Stigler, 2023](#)). Explicitly distinguishing between such subtle differences in the way that a construct is conceptualised is important both for developing a clearer conceptualisation of teaching quality and understanding whether we are measuring the constructs we intend to measure.

6.3. Implications for the Field and Future Research

We are advocating for future research to take seriously the idea that observation systems must adequately measure the intended construct in all settings and guard against the threat of lack of MI across lessons, a point that fits squarely within the fairness concerns of the AERA testing standards ([American Educational Research Association et al., 2014](#)), which state that the same construct should be measured across all occasions if interpretations of scores rely on measuring a given construct. This requires work at the intersection of conceptualising constructs and measurement. We make the following recommendations to improve measurement practice with observation manuals.

Observation manuals should first ensure that all constructs are clearly and explicitly defined (see [White et al., 2025](#)). Second, both developers and users of such manuals should consider how constructs will manifest across the full breadth of lessons (e.g., across individual work, small group, and whole class instruction; across introductions of new content, reviews of previously learned content, and practicing; and across lessons focused on teaching basic skills and those focused on abstract or conceptual knowledge; e.g., [Gleason et al., 2017](#)). This is effectively considering how the substantive model applies to specific types of lessons. Users of observation systems should specifically consider the breadth of lessons that might be encountered in their local context. Third, for each type of lesson that could be expected to be encountered, careful consideration should be given to whether the operationalisation of the construct in the observation rubric will adequately capture the intended construct. This effectively explores how the data model applies to specific types of lessons. For manual developers, this should be done explicitly as part of the creation of the rubric to guide future users in understanding where and when manuals can be applied validly, but this should also be done by manual users to ensure that manuals apply to the local context and for their specific research question. The results of these analyses should be compared using the sort of reasoning applied in the illustrative examples in this paper. If there is reason to think that the rubric operationalisation will not lead scores to represent the intended construct in some types of lessons, the approaches discussed next can be applied.

There are several approaches that might address identified issues, including adjusting construct definitions (see recommendation for ROC above), adjusting manuals/rubrics to improve measurement across lesson types, developing separate rubrics for different types of lessons (see, e.g., TRU framework; [Schoenfeld, 2018](#)), and/or narrowing the scope of manuals (see, e.g., [Matsumura et al., 2008](#)). Alternatively, where these solutions are not feasible, users of manuals could identify characteristics of lessons that are potentially problematic and, in addition to scoring manuals, code lessons as to whether they fall into categories that are problematic. This would allow for statistical analyses that explore the prevalence of potentially problematic lessons and the extent to which construct scores

from problematic lessons have the same correlation with criterion scores (i.e., test for moderation effects/predictive bias; American Educational Research Association et al., 2014). By having evidence regarding the frequency of lessons that potentially lack MI and regarding predictive bias, authors can begin to make the case that, even if a lack of MI exists, it does not impact the study's conclusions. Further, this would allow researchers studying this challenge of MI across lessons to look across studies to understand where MI concerns might be strongest and develop stronger guidelines for addressing MI risks across lessons.

6.4. Thinking Differently About Measurement Invariance

Measurement invariance is generally considered with psychometric models at the level of combining items into scales. This is possibly because this is the level at which clear and specific procedures to test for and address MI exist (Markus & Borsboom, 2013). However, as we discussed throughout this paper, this is not the only level at which MI can be a problem (c.f., Fischer et al., 2025). Measurement invariance requires a stable correspondence between the cues that determine the score assigned and the underlying construct (Engelhard & Wind, 2018). Non-invariance, then, is an important measurement consideration at the level of individual items, as well as at the level of scales. This presents a challenge, because there are no standard procedures for assessing measurement non-invariance at the item level. The field needs to create clear guidelines and procedures for establishing the claim of MI at the item level to ensure that item scores correspond in clear and stable ways to the constructs that those scores are meant to represent.

Additionally, cross-cultural research has shown that MI is not just a methodological issue, but also a substantive issue that can promote theoretical development (Fischer et al., 2025). Through studies like this one that examine how teaching quality manifests in similar and different ways across lessons, important insights can be gained that explain challenges in the field. For example, if one measures only some of the ways that teachers represent content, this would, in principle, reduce the potential relationship between observation scores and measures of student learning, making the construct of how teachers represent content appear to be less important than it is. This could explain the weak and inconsistent relationships between observation scores and student learning (Klette et al., 2022).

Author Contributions: Conceptualization, M.W. and A.J.; Methodology, M.W. and A.J.; Formal analysis, M.W., A.J. and J.L.; Writing—original draft, M.W., A.J., J.L. and K.K.; Writing—review & editing, M.W., A.J., J.L. and K.K.; Funding acquisition, K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Research Council of Norway grant number 300791.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: There is no data associated with this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Alexander, R. (2008). *Towards dialogic teaching* (4th Revised ed.). Dialogos.
- American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. Available online: <http://www.apa.org/science/programs/testing/standards.aspx> (accessed on 30 March 2026).
- Anthony, G., & Walshaw, M. (2009). *Effective pedagogy in mathematics*. International Academy of Education.

- Ball, D. L., & Forzani, F. M. (2007). 2007 wallace foundation distinguished lecture—What makes education ‘research educational’? *Educational Researcher*, 36, 529–540. [CrossRef]
- 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. [CrossRef]
- Blikstad-Balas, M. (2014). Vague concepts in the educational sciences: Implications for researchers. *Scandinavian Journal of Educational Research*, 58(5), 528–539. [CrossRef]
- Blikstad-Balas, M., Klette, K., & Tengberg, M. (Eds.). (2022). *Ways of analyzing teaching quality: Potentials and pitfalls*. Scandinavian University Press.
- Brunswik, E. (1952). *The conceptual framework of psychology*. University of Chicago Press.
- Cazden, C. (2001). *Classroom discourse: The language of teaching and learning* (2nd ed.). Heinemann.
- Charalambous, C. Y., & Praetorius, A.-K. (2020). Creating a forum for researching teaching and its quality more synergistically. *Studies in Educational Evaluation*, 67, 8. [CrossRef]
- Christensen, A. S., & Mathé, N. E. H. (2023). Higher order thinking in social science education—An empirical study with classroom observations from Denmark and Norway. *Sammenlignende Fagdidaktik*, 7, 11–36. [CrossRef]
- Cohen, J. (2018). Practices that cross disciplines? Revisiting explicit instruction in elementary mathematics and English language arts. *Teaching and Teacher Education*, 69, 324–335. [CrossRef]
- Cor, K. (2011, April 1–5). *The measurement properties of the PLATO rubric*. Annual Meeting of the American Educational Research Association, New Orleans, LA, USA.
- Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments* (1st ed.). Routledge.
- Fischer, R., Karl, J. A., Luczak-Roesch, M., & Hartle, L. (2025). Why we need to rethink measurement invariance: The role of measurement invariance for cross-cultural research. *Cross-Cultural Research*, 59(2), 147–179. [CrossRef]
- Gehlbach, H. (2015). Seven survey sins. *The Journal of Early Adolescence*, 35(5–6), 883–897. [CrossRef]
- Gleason, J., Livers, S., & Zekowski, J. (2017). Mathematics classroom observation protocol for practices (MCOP2): A validation study. *Investigations in Mathematics Learning*, 9(3), 111–129. [CrossRef]
- Grossman, P. (2015). *Protocol for Language Arts Teaching Observation (PLATO) manual* (version 5.0). Graduate School of Education at the University of Pennsylvania. Available online: <https://www.gse.upenn.edu/content/protocol-language-arts-teaching-observation-plato> (accessed on 30 March 2026).
- 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. [CrossRef]
- 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. [CrossRef]
- 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 4000 classrooms. *The Elementary School Journal*, 113(4), 461–487. [CrossRef]
- Hiebert, J., & Stigler, J. W. (2023). Creating practical theories of teaching. In A.-K. Praetorius, & C. Y. Charalambous (Eds.), *Theorizing teaching: Current status and open issues* (pp. 23–56). Springer.
- Jentsch, A., & Klette, K. (2024, April 11–14). *Using PLATO to assess teaching quality in Scandinavian mathematics and language arts classrooms*. American Educational Research Association (AERA) Annual Meeting 2024, Philadelphia, PA, USA.
- Jentsch, A., White, M. C., & Klette, K. (2026). Three types of subject-specificity. Exploring differences in the measurement of teaching quality across Norwegian mathematics and language arts classrooms. *Teaching and Teacher Education*, 169, 105291. [CrossRef]
- Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, 42(4), 448–457. [CrossRef]
- 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. Available online: <https://files.eric.ed.gov/fulltext/ED540960.pdf> (accessed on 30 March 2026).
- Klette, K. (2020). Towards programmatic research when studying classroom teaching and learning. In L. Ligozat, A. Rakhkockhine, & J. Almqvist (Eds.), *Thinking through didactics in a changing world. European perspectives on learning, teaching and the curriculum* (pp. 137–158). Routledge Education. [CrossRef]
- 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. [CrossRef]
- Klette, K., Magnusson, C. G., & Sigurjónsson, J. Ö. (Eds.). (2026). *Investigating teaching quality through video capture: Equity and quality in Nordic classrooms*. Springer Nature.

- Klette, K., Roe, A., & Blikstad-Balas, M. (2022). 6. Observational Scores as Predictors for Student Achievement Gains. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Ways of analyzing teaching quality: Potentials and pitfalls* (pp. 174–203). Scandinavian University Press.
- Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14(1), 25–47. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- Luoto, J., Klette, K., & Blikstad-Balas, M. (2022). Possible biases in observation systems when applied across contexts: Conceptualizing, operationalizing, and sequencing instructional quality. *Educational Assessment, Evaluation and Accountability*, 35(1), 105–128. [\[CrossRef\]](#)
- Markus, K. A., & Borsboom, D. (2013). *Frontiers in test validity theory: Measurement, causation, and meaning* (1st ed.). Routledge.
- 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. [\[CrossRef\]](#)
- McClellan, C. A. (2010). *Constructed-response scoring—Doing it right*. ETS.
- McGrane, J. A., & Maul, A. (2020). The human sciences, models and metrological mythology. *Measurement*, 152, 107346. [\[CrossRef\]](#)
- Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, 95(4), 728–743. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mercer, N., & Sams, C. (2006). Teaching children how to use language to solve maths problems. *Language and Education*, 20(6), 507–528. [\[CrossRef\]](#)
- Nissen, A. (2023). Cognitive activation as an aspect of literature instruction. *L1-Educational Studies in Language and Literature*, 23, 1–20. [\[CrossRef\]](#)
- Nystrand, M. (1996). *Opening dialogue: Understanding the dynamics of language and learning in the English classroom* (Language and literacy series). Teachers College Press.
- Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM*, 50(3), 535–553. [\[CrossRef\]](#)
- Praetorius, A.-K., Rogh, W., Bell, C., & Klieme, E. (2019). Methodological Challenges in Conducting International Research on Teaching Quality Using Standardized Observations. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative studies in education* (pp. 269–288). SAGE Publications Ltd.
- Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM*, 48(1), 29–40. [\[CrossRef\]](#)
- Schoenfeld, A. H. (2018). Video analyses for research and professional development: The teaching for robust understanding (TRU) framework. *ZDM*, 50(3), 491–506. [\[CrossRef\]](#)
- Sigurjónsson, J. Ö., Sigurðardóttir, A. K., Gísladóttir, B., & van Bommel, J. (2022). Connecting student perceptions and classroom observations as measures of cognitive activation. *Nordic Studies in Education*, 42(4), 328–346. [\[CrossRef\]](#)
- Tengberg, M., van Bommel, J., Nilsberth, M., Walkert, M., & Nissen, A. (2022). The quality of instruction in Swedish lower secondary language arts and mathematics. *Scandinavian Journal of Educational Research*, 66(5), 760–777. [\[CrossRef\]](#)
- Wells, G. (1999). *Dialogic inquiry: Towards a sociocultural practice and theory of education*. Cambridge University Press.
- Welzel, C., & Inglehart, R. F. (2016). Misconceptions of measurement equivalence: Time for a paradigm shift. *Comparative Political Studies*, 49(8), 1068–1094. [\[CrossRef\]](#)
- Welzel, C., Kruse, S., & Brunkert, L. (2022). Against the mainstream: On the limitations of non-invariance diagnostics: Response to Fischer et al. and Meuleman et al. *Sociological Methods & Research*, 52(3), 1438–1455. [\[CrossRef\]](#)
- White, M. (2025). A peculiarity in psychological measurement practices. *Psychological Methods*. Advance online publication. [\[CrossRef\]](#)
- White, M., Edelsbrunner, P. A., & Thurn, C. M. (2024). The conceptualisation implies the statistical model: Implications for measuring domains of teaching quality. *Assessment in Education: Principles, Policy & Practice*, 31(3–4), 1–25. [\[CrossRef\]](#)
- White, M., Göllner, R., & Kleickmann, T. (2025). Reconsidering the measurement of teaching quality using a lens model: Measurement dilemmas and trade-offs. *School Effectiveness and School Improvement*, 36(2), 192–210. [\[CrossRef\]](#)
- White, M., Luoto, J., Klette, K., & Blikstad-Balas, M. (2022). Bringing the conceptualization and measurement of teaching into alignment. *Studies in Educational Evaluation*, 75, 101204. [\[CrossRef\]](#)
- Xu, L., & Clarke, D. (2018). Validity and comparability in cross-cultural video studies of classrooms. In L. Xu, G. Aranda, W. Widjaja, & D. Clarke (Eds.), *Video-based research in education* (pp. 19–33). Routledge.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.