Teaching and Teacher Education 169 (2026) 105291

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate

Research paper

Three types of subject-specificity. Exploring differences in the measurement of teaching quality across Norwegian mathematics and language arts classrooms

Armin Jentsch [a,*], Mark C. White [a,b], Kirsti Klette [a,b]

[a] University of Cologne, Faculty of Human Sciences, Gronewaldstr. 2, 50931 Cologne, Germany
[b] University of Oslo, Department of Teacher Education and School Research, Postbox 1099 Blindern, 0317 Oslo, Norway

ARTICLE INFO

Keywords:
Teaching quality
Classroom observation
Equivalence
Generalizability theory
Mathematics
Language arts

ABSTRACT

Measuring teaching quality is important to better understand student learning in classrooms. To directly capture teaching quality constructs, subject-specific and subject-generic observation systems have been developed, with benefits and challenges on both sides. However, we show that subject-specificity may address different things and come with different implicit assumptions about what is subject-specific. As a common ground for empirical work, we draw on a framework of measurement equivalence and discuss three nested claims on subject-specificity. Then, we analyze videotaped lessons from 47 mathematics and 49 language arts classrooms which were scored with the Protocol for Language Arts Teaching Observation (PLATO). Results indicate that scores in Instructional Scaffolding show the largest differences between subjects. We argue that subject-specificity is a claim about a theoretical construct, which should be clearly addressed in future research.

1. Introduction

Measuring teaching quality directly through classroom observations is an important enterprise to explore student learning (Blömeke et al., 2022; Kyriakides et al., 2018), for school inspection purposes (Taut & Rakoczy, 2016), as well as for formative assessment in teacher education (Jenset & Brataas, 2024) and professional development (Magnusson et al., 2023). To do so, many observation systems have been developed in the past decades, some of which claim to be applicable *across* subjects, while others focus on teaching quality *within* subjects.

While teaching effectiveness research typically seeks to find patterns of learning that are valid across subjects (see Praetorius et al., 2025 for a summary), Grossman and Stodolsky (1995) remind us in their early work that teaching practices could vary systematically between subjects, as these come with their own culture, values, and norms. Accounting for such differences between subjects in measurement arguably allows for a more detailed view of what is happening in the classroom and could be particularly beneficial when related to student learning, evaluating teacher education or professional development. For this reason, scholars argue on a regular basis that *subject-specificity* should be considered more seriously when studying teaching quality (e.g., Cohen et al., 2018; Reusser & Pauli, 2021; Schlesinger & Jentsch, 2016).

However, a brief review of the literature shows that it is not at all self-explanatory what is being discussed under the concept of subject-specificity. Researchers use this term in many ways, including but not limited to subject-specific constructs of teaching quality (e.g., Charalambous & Praetorius, 2018), measures (e.g., Kunter & Ewald, 2016), rubrics (e.g., Schlesinger & Jentsch, 2016), raters, observations (Dreher & Leuders, 2021), and recently, empirical relations between scores and the constructs of interest (e.g., Wemmer-Rogh et al., 2024). We argue that the concept of subject-specificity can be meaningfully used in all these cases but may address different things and come with different implicit assumptions about what is subject-specific (and what is not).

Following this premise, the present study aims to contribute to the ongoing debate on subject-specificity in research on teaching quality in a twofold manner. First, we draw on a framework on measurement equivalence (Fischer et al., 2025) to organize different ways of thinking about subject-specificity and to help scholars be more explicit about the assumptions underlying their research. An important component of this contribution is to highlight that many aspects of subject-specificity are fundamentally non-empirical, theoretical claims. Second, to provide an example of how empirical aspects of subject-specificity could be investigated, we use the Protocol for Language Arts Observation (PLATO, Grossman et al., 2014) as well as Generalizability Theory (Cronbach

* Corresponding author.
E-mail addresses: ajentsc2@uni-koeln.de (A. Jentsch), m.c.white@ils.uio.no (M.C. White), kirsti.klette@ils.uio.no (K. Klette).

https://doi.org/10.1016/j.tate.2025.105291
Received 17 February 2025; Received in revised form 25 July 2025; Accepted 31 October 2025
Available online 4 November 2025

et al., 1972) and explore teaching quality scores across subjects in Norwegian mathematics and language arts classrooms.

1.1. Investigating teaching quality across subjects using classroom observation

While one could always compare observation scores across subjects, it is a far more difficult enterprise to compare teaching quality constructs *using scores*. This comparison of constructs is typically the intended goal. Drawing on a recent framework by Fischer and colleagues (2025), we argue that claims of subject-specificity can take various forms when studying teaching quality (Jentsch & Kirchhoff, 2024). Three nested levels of claims about the subject-specificity of constructs are presented in Fig. 1: (1) *Functional* subject-specificity, (2) *operational* subject-specificity, and (3) subject-specificity *in the measurement* of teaching quality. Claims of subject-specificity at one level precludes studying subject-specificity at lower levels. It is only when subject-specificity can be accounted for at each level that empirical comparisons of scores can be used to understand how constructs vary across subjects (Fischer et al., 2025).

(1) *Functional* subject-specificity suggests that constructs differ in how they are conceptualized within subjects or that constructions are only relevant for some subjects. For instance, Praetorius and colleagues (2015) refrained from assessing cognitive activation in German and English classrooms, because they argued that cognitive activation did not have the same conceptual understanding across subjects. This is not an empirical claim about the functioning of rubrics, but a conceptual claim about the very nature of the construct being measured. As such, empirical evidence cannot shed light on these claims because the claim explicitly precludes the capacity to measure the construct across subjects, implying the need for studies to be subject-specific. It is important to note that some sub-constructs of teaching quality may be conceptualized as subject-specific and others not. This is a vital point because while we often talk about the generic construct of teaching quality, it is often the sub-constructs that are deemed as subject-specific or not (Praetorius et al., 2015). Further, one group of researchers might make claims of subject-specificity for a construct while another group might reject that claim. Such disagreements are conceptual and reflect a lack of shared understanding of the nature of a construct (i.e., they are not empirically testable).

(2) *Operational* subject-specificity comes into play when different instruments, measures, tools, or rubrics are used to capture the same constructs of interest across subjects. For example, Cohen (2018) sought to measure direct instruction in both math and language arts, adapting the language-arts specific PLATO observation rubric to capture teaching quality in mathematics. In making this adaptation, the construct of Strategy Use, which provides a single score in language arts, was sub-divided to separately score procedural and conceptual strategy instruction. In doing so, Cohen and colleagues implicitly make the claim that Strategy Use has the same conceptual understanding in both subjects, but the way one would need to operationalize the construct varies across subjects.

Under this sort of conceptualization of subject-specificity, it becomes theoretically possible and meaningful to measure the same construct across subjects, but there exist many methodological difficulties that may make comparisons of scores invalid (e.g., ensuring that construct-underrepresentation and construct-irrelevant variation are comparable across the different operationalizations and that scores are measured on the same scale). For example, returning to Cohen (2018), questions arise as to whether the single score of strategy instruction in language arts should be compared with the procedural strategy instruction score in math, or the conceptual strategy score, or an aggregate of the two, or whether this depends on what types of strategies are taught in language arts (e.g., strategies that seem closer to conceptual strategies or ones that seem closer to procedural strategies). These are questions that empirical analyses cannot fully address because they intrinsically deal with the extent to which each operationalization captures the intended construct (i.e., levels of underrepresentation or construct-irrelevant variation in scores). That is, empirical analyses can inform researchers only to the extent to which scores can be compared, but not on ways in which constructs can be compared. Note that the same points would apply to two different observation manuals' operationalization of the same construct.

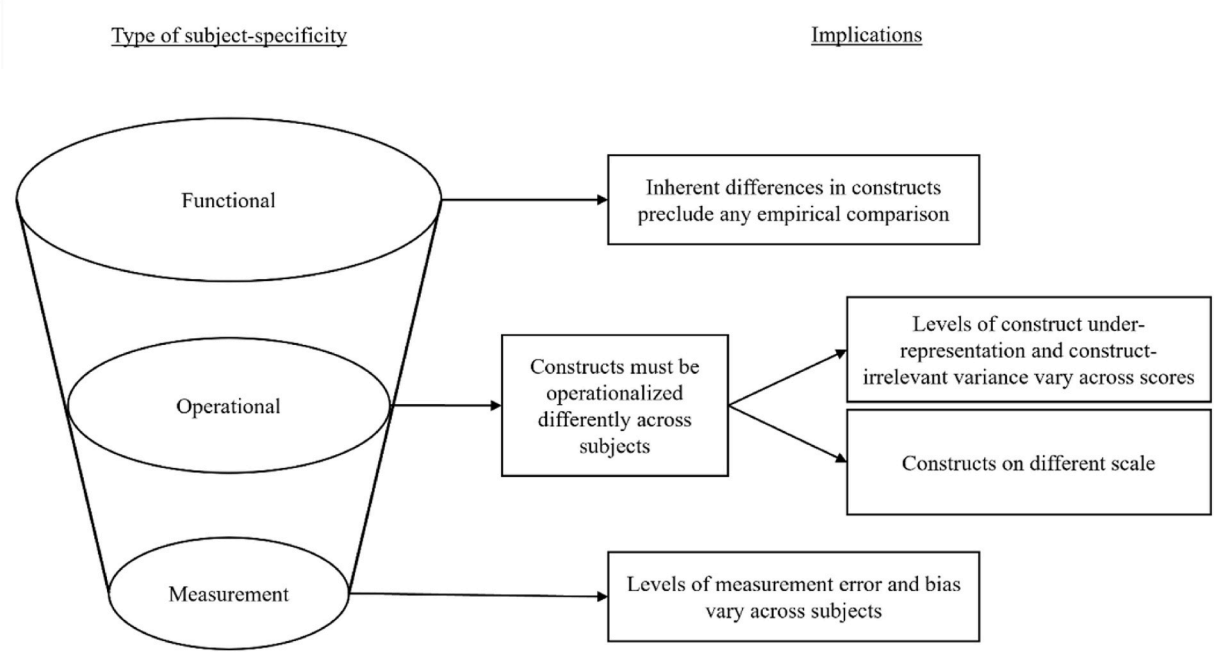


Fig. 1. Different types of subject-specificity and their implications for research practice.

(3) *Measurement* subject-specificity implies that levels of measurement error and variability in a construct vary across subjects. Here, there are technical challenges that must be addressed to compare levels of a construct across subjects, but these can be overcome through careful measurement practices. Claims of subject-specificity at this level involve adopting the same conceptual understanding of a construct, as well as using the same measurement tool, which is claimed to validly measure the intended construct in each subject. Even in this case, though, measurement artifacts that introduce subject-specific effects may remain. When measurement error is present, this error needs to be modelled and controlled for before comparing scores across subjects.

In the empirical part of this paper, we aim to explore variation in teaching quality constructs between Norwegian mathematics and language arts classrooms in light of the framework discussed here. To do so, we first consider functional and operational subject-specificity and then explore measurement subject-specificity in those elements that meet functional and operational equivalence.

1.2. The Protocol for Language Arts Teaching Observation (PLATO)

Teaching quality is assessed with the Protocol for Language Arts Teaching Observation (PLATO) observation system in this study. PLATO was first developed to explore effective teaching practices in language arts classrooms and their relationship to value added scores (Grossman et al., 2014). However, because of its broad focus on a variety of teaching practices, PLATO has shown to be applicable to other subjects (e.g., Cohen, 2018) and across various educational systems (i.e., countries, Klette et al., 2017). PLATO conceptualizes four instructional domains, which are *Instructional Scaffolding*, *Disciplinary Demand*, *Representation and Use of Content*, and *Classroom Environment*.

Instructional scaffolding addresses the extent to which teachers provide instructional support to students. It consists of the elements Modeling, Strategy Use, Feedback, and Accommodation for Language Learning. The domain refers to theories of explicit and guided instruction, where content is decomposed into smaller steps that are explicitly taught, as well as sustained by targeted feedback focusing on students' learning progress. The element of Accommodation for Language Learning was judged to be functionally subject-specific as it captures how teachers support language learning for those struggling with the language of instruction, particularly in first language instruction.

Disciplinary demand is concerned with the extent to which students are engaged in cognitively challenging activities and talk about disciplinary content (Grossman et al., 2014). This domain includes the elements Intellectual Challenge, Classroom Discourse, and Text-Based Instruction. Intellectual Challenge distinguishes between tasks that require rote and recall, and higher-level thinking. Classroom discourse highlights instruction that gives students an opportunity to engage in discussions. Text-based instruction captures students' opportunities to work with authentic texts. We judge this element to be functionally subject-specific since it captures specific aspects of learning to work with disciplinary texts (e.g., poems versus novels).

Representation and Use of Content targets how teachers present content to students. This domain includes the elements Representation of Content, Connections to Prior Knowledge, and Purpose. Representation of Content focuses on the quality and conceptual richness and depth of teachers' explanations of content. The element of Connections to Prior Knowledge addresses connections between new and previously learnt knowledge. Finally, Purpose captures how clearly teachers present and frame learning objectives.

Classroom environment captures how efficiently a teacher manages the behavior of students and time on task in the classroom. This domain involves the elements Behavior Management and Time Management, where poor management means less time on task or a disorderly classroom, and good management is an orderly classroom such that instructional time is maximized.

1.3. The present study

Measuring teaching quality is important to better understand student learning in classrooms within and across subjects. To directly capture teaching quality constructs, subject-specific and generic observation systems have been developed, with benefits and challenges on both sides. As noted above, while other studies have explored subject-specificity, they are often unclear about what claims are being made (e.g., functional, operational, or measurement), limiting their potential to contribute to our understanding of teaching quality. Importantly, subject-specific observation systems preclude the capacity to measure teaching quality constructs *across* subjects from the start, because of implicit assumptions on functional or operational differences between subjects (Fig. 1). Adopting the same conceptualization and operationalization (i.e., thereby avoiding functional and operational subject-specificity) allows us to explore subject-specificity empirically, that is, in the measurement of teaching quality.

In social science research, such questions are typically addressed with reflective measurement models (e.g., factor analysis) and invariance testing when latent constructs are investigated (see Fischer et al., 2025 for an overview). Invariance testing with factor models involves exploring whether the empirical relations between indicators of the construct of interest are similar across settings. However, new research questions the usefulness of such testing for observation systems, as items in observation systems are not exchangeable (e.g., White, 2025; White et al., 2024). PLATO emphasizes element (or item-level) scores that are directly assigned by observers, making the typical reflective modelling approach inapplicable. Due to this, we conduct a Generalizability Study (G Study; Brennan, 2001). G Studies decompose variation in item-level scores to sources that contain construct-relevant variation (i.e., signal or true variation in teaching quality) and construct-irrelevant variation (i.e., measurement error). They also allow for overall estimates of mean scores that incorporate all sources of variation for the respective units of measurement (say, a single lesson or a classroom, Casabianca et al., 2013).

To test for measurement invariance, we draw on ideas of configural, metric, and scalar invariance from reflective modeling and apply them to Generalizability Theory. Configural invariance is the extent to which the same sub-aspects of constructs are captured across subjects. This is implied by the claim that there is no functional or operational subject-specificity and using the same rubric operationalization across subjects. Metric invariance is the extent to which a fixed change in observed scores corresponds to a fixed change in the latent construct. This occurs if the percentage of variance attributable to true score variation (i.e., construct-relevant variation) is equivalent across subjects. In other words, and in line with Fig. 1, we call the absence of metric invariance across subjects *measurement subject-specificity*. Finally, scalar invariance is the extent to which mean differences in observed scores correspond to mean differences in the latent construct. This is definitionally true if sources of error are random, as opposed to systematic (e.g., if raters are equally lenient or harsh across subjects). We address the following research questions:

(1) How do sources of variability compare between subjects?
(2) What are the implications regarding measurement error and reliability for lesson or classroom scores, respectively?

In this study, we use the PLATO observation system to investigate measurement subject-specificity in Norwegian mathematics and language arts classrooms. PLATO involves similar scoring procedures to many other observation systems (Casabianca et al., 2013). That is, raters are confronted with (videotaped) evidence from classrooms and employ a rubric with a set of observable indicators to provide an integrated

judgement of teaching behaviors and/or teacher-student interactions based on rubric descriptors.

2. Methods

2.1. Videotaped lessons, raters, and procedures

Data were collected in years 2014–2015 within the Linking Instruction and Student Achievement (LISA) study (Klette et al., 2017). Our sample consists of videotaped lessons from Norwegian classrooms in grade 8, each one taught by a unique teacher (47 in mathematics and 49 in language arts). From each subject and classroom, an average of three largely consecutive lessons have been sampled and videotaped. To videotape lessons, two cameras and two microphones were used. A camera was set on the class using a wide angle, and one camera focused on the front of the class where the teacher typically stands. Similarly, one microphone was used to capture ambient audio of the whole classroom, and the teacher wore another microphone.

Lessons have been scored in segments of 15 min as recommended by the PLATO developers, which results in a full data set of 599 scored lesson segments in mathematics, and 630 segments in language arts. An average of three segments per lesson was scored, because lessons tended to vary in length. Lesson scoring was then conducted by trained raters who were also certified to use PLATO. All raters had expertise in the subject being scored. Raters were researchers, PhD students, or student teachers who had obtained at least a bachelor's degree. The training involved certification with master raters, with whom raters need to reach 80 % of agreement.

Raters assign their scores on a four-point scale for each of the PLATO elements regarding specific behavioral indicators (from 1: almost no evidence, through 4: consistent strong evidence). However, we refrained from using the elements Text-based Instruction and Accommodations for Language Learning, as these are considered functionally subject-specific (i.e., they are only relevant in language arts).

2.2. Generalizability and decision studies

To answer our research questions, we conduct separate G Studies for both subjects and PLATO element or domain, respectively. Since segments and lessons are nested within classrooms and raters are partially cross-classified (i.e., multiple raters scored each segment and raters score multiple but not all segments), our G Study models divide the variation in scores across classrooms (c), lessons within classrooms (l:c), segments within lessons within classrooms (s:l:c), raters (r), and the interactions between raters and lessons ([l:c]r). We refrain from adding a rater-by-classroom interaction to reduce model complexity, and because raters were not assigned to score all lessons from a classroom. Further, these variance components are not exhaustive but reflect those typical of such studies and represent what is considered important in measurement decisions.

Variance components are obtained by estimating mixed models using the lme4 package (Bates et al., 2024) in R statistical software. The obtained variance components, then, give rise to a Decision Study (D Study) which estimates measurement subject-specificity in scores at different levels of aggregation (Shavelson & Webb, 1991). A summary of the estimated effects and their interpretation is displayed in Table 1.

We consider classroom, lesson, and segment effects to capture true variation in the measured teaching quality construct. Classroom effects reflect variation that is due to differences between classrooms. The lesson effect captures variation around classroom averages that is due to deviations of lesson scores from the classroom average, which we assume captures teachers focusing lessons on different aspects of the broader teaching quality construct. Similarly, segment effects capture variation within lessons, and provide information on how patterns of teaching quality vary within lessons (Casabianca et al., 2013). Variance components related to raters include main effects (i.e., some raters scoring systematically higher or lower than others, also termed rater leniency or harshness) and rater-by-lesson interactions (i.e., raters assigning different scores to the same lesson). Finally, our G Study involves a residual variance component, which reflects raters assigning different scores to the same individual segment (White & Ronfeldt, 2024). Rater and residual components, therefore, contribute to measurement error.

Table 1
Estimated variance components with examples of measurement subject-specificity.

Variance component	Notation	Description and example of measurement subject-specificity
Classroom, c	σ^2_c	Variation in average scores across classrooms. For example, Representation of Content could reflect larger shares of classroom variance in math, because errors by the teacher are captured.
Lesson, l:c	$\sigma^2_{l:c}$	Variation in average scores across lessons within classrooms. For example, Classroom Discourse could vary more across language arts lessons because topics change more frequently than in math.
Segment, s:l:c	$\sigma^2_{s:l:c}$	Variation in scores across segments within lessons. For example, behavior issues could arise throughout math lessons, but only in the beginning of lessons in language arts.
Rater, r	σ^2_r	Variation in average scores assigned by different raters, called also rater leniency or harshness. For example, differences in the way that teachers provide feedback across subjects could result in raters not properly noting the feedback provided in language arts, leading to lower scores in language arts than in math.
Rater-by-lesson, (l:c)r	$\sigma^2_{(l:c)r}$	Variation in average scores assigned by different raters to the same lesson. For example, differences in how lessons are structured may lead to more rater disagreement in lesson scores for math lessons than for language arts lessons.
Residual, (s:l:c) r, res	σ^2_{res}	Variation resulting from raters assigning different scores to individual segments. For example, raters may disagree on Classroom Discourse scores for individual segments more in math because segments are more likely to have scores near thresholds of adjacent score categories.

To formally test for measurement subject-specificity, we pooled the data from mathematics and language arts classrooms and fit another set of linear mixed models. The first set of these models allowed variance components to vary across subjects and the second set of models fixed variance components to be equivalent across models. The same variance components were used (see Table 1). Likelihood ratio tests were performed to test whether variance components were equally distributed across subjects (i.e., by comparing the restricted model against a model with subject-specific variance components, for a similar approach see Casabianca et al., 2013). Because likelihood ratio tests are generally sensitive, and we wish to control accumulated type-1 error on the domain level, we consider statistical significance at $p < .010$.

To further evaluate measurement subject-specificity in PLATO scores, we conduct a D Study that examines the percentage of true-score variation (i.e., reliability) of lesson-level and classroom-level average scores because these are typically the units of measurement in educational research or practice (e.g., Casabianca et al., 2013). Note that calculations are based on three lessons and two raters per classroom as well as three segments and one rater within a lesson.

3. Results

In Table 2, we present means and standard deviations for the PLATO element scores. We show a raw difference to highlight differences in actual practice. Since there are typically five lessons in a week, scoring one point higher once a week corresponds to an average difference of 0.2

Table 2
Descriptives by subject for all PLATO elements.

	Mathematics		Language Arts		Raw mean difference
	M	*SD*	*M*	*SD*	
Elements					
Purpose	2.14	0.41	2.18	0.43	−0.04
Intellectual Challenge	2.49	0.50	2.14	0.59	0.35
Classroom Discourse	2.03	0.52	1.98	0.64	0.05
Representation of Content	2.48	0.66	2.17	0.57	0.31
Connections to Prior Knowledge	1.74	0.56	1.65	0.44	0.09
Modelling	2.16	0.74	1.33	0.40	0.83
Strategy Use	1.90	0.79	1.41	0.43	0.49
Feedback	2.07	0.45	2.03	0.48	0.04
Behavior Management	3.76	0.35	3.73	0.42	0.03
Time Management	3.49	0.59	3.33	0.66	0.16

(i.e., 1/5). A raw difference of 0.05, therefore, corresponds to a single lesson scoring one point higher about once a month (i.e., once every four weeks) while a raw difference of 0.3 corresponds to one and a half lessons scoring one point higher each week. Except for Purpose, scores tend to be higher in mathematics than in language arts (see Table 3).

The scores shown in Table 2 are raw differences in the assigned scores across subjects for the elements that we believe do not show functional and operational subject-specificity. The natural question is to what extent any of the observed differences reflect true differences in the theoretical constructs that we intended to measure. Making such comparisons requires scalar invariance, the most difficult form of invariance to demonstrate (van de Vijver et al., 2019). In our case, this difficulty arises because we can only examine rater error by looking at rater disagreement (as is typical with observational data). If math or language arts raters were, on average, more or less lenient than each other, this would be invisible in our data (White & Ronfeldt, 2024). Then, we would be unable to establish scalar invariance, and comparisons of mean levels of the intended construct cannot confidently be made. This limitation of our analysis is present in most observational studies since studies are typically limited to examining rater disagreement rather than rater error. This is simply hidden in analyses that do not explicitly model rater error (e.g., Wemmer-Rogh et al., 2024). Having the same raters score all videos could help to address this problem but may not solve it since within-rater differences in subject-matter expertise could lead to different average levels of rater error within-raters across-subjects.

3.1. Generalizability studies

G Studies can help to shed light on issues of measurement subject-specificity (i.e., whether a fixed change in the observed score corresponds to a fixed change in the construct). Measurement subject-specificity would be indicated when the percentage of variance in a score that is due to true variation in the construct differs across subjects (i.e., variance components that capture construct-relevant variation, which are classroom, lesson, and segment effects in our case). However, measurement subject-specificity also depends on the level to which scores are aggregated, which shall be discussed in the next section.

Likelihood ratio tests show statistically significant differences in the distribution of variance components for Intellectual Challenge (χ^2 = 22.60, p = .003), Representation of Content (χ^2 = 22.49, p = .004), Modeling (χ^2 = 95.36, p < .001), Strategy Use (χ^2 = 63.73, p < .001), and Feedback (χ^2 = 23.69, p = .001), with the most pronounced differences occurring for Modeling and Strategy Use. In particular, we see that lesson-level variation is close to zero for Modelling and Strategy Use in language arts classrooms, which indicates a larger amount of true score variation for math classrooms. This means that measurement subject-specificity exists for all captured elements of Instructional Scaffolding, suggesting that such teaching practices follow a different pattern in math and language arts classrooms.

The G Studies can also help elucidate the nature of the construct and

Table 3
Variance decomposition of scores by subject (percentages of total variability).

	Elements									
	PUR	IC	CD	ROC	CPK	MOD	SUI	FEED	BM	TM
Mathematics										
Classroom, c	0.01	0.05	0.06	0.10	0.03	0.08	0.04	0.02	0.01	0.06
	(4.2)	(9.4)	(9.3)	(14.1)	(5.2)	(7.9)	(4.3)	(6.6)	(2.6)	(7.8)
Lesson, l	0.07	0.04	0.00	0.14	0.00	0.20	0.08	0.00	0.03	0.04
	(24.6)	(7.0)	(0.0)	(20.0)	(0.0)	(20.3)	(8.4)	(0.0)	(14.1)	(5.1)
Segment, s	0.08	0.06	0.20	0.18	0.14	0.23	0.07	0.00	0.08	0.29
	(25.3)	(11.7)	(33.4)	(25.0)	(20.2)	(23.7)	(7.5)	(0.0)	(32.2)	(40.0)
Rater, r	0.01	0.05	0.01	0.06	0.10	0.06	0.29	0.04	0.03	0.11
	(3.4)	(10.4)	(1.7)	(8.3)	(14.9)	(6.6)	(29.2)	(9.4)	(13.5)	(15.4)
(l:c)r	0.02	0.06	0.11	0.00	0.07	0.04	0.10	0.08	0.02	0.00
	(6.0)	(11.7)	(18.2)	(0.0)	(10.0)	(4.5)	(10.2)	(22.6)	(6.4)	(0.0)
(s:l:c)r, res	0.11	0.26	0.23	0.24	0.33	0.36	0.41	0.23	0.08	0.23
	(36.6)	(49.8)	(37.4)	(32.6)	(49.6)	(37.0)	(40.6)	(61.4)	(31.2)	(31.7)
Language Arts										
Classroom, c	0.02	0.07	0.04	0.08	0.02	0.03	0.03	0.04	0.05	0.07
	(5.8)	(10.9)	(5.7)	(11.8)	(3.6)	(7.9)	(6.5)	(9.7)	(15.8)	(7.9)
Lesson, l	0.02	0.03	0.13	0.04	0.00	0.00	0.00	0.08	0.05	0.08
	(5.7)	(5.1)	(17.1)	(6.1)	(0.0)	(0.0)	(0.0)	(18.5)	(14.4)	(9.2)
Segment, s	0.09	0.20	0.12	0.27	0.20	0.10	0.06	0.11	0.10	0.47
	(25.2)	(29.9)	(16.1)	(41.4)	(35.7)	(24.3)	(12.7)	(26.5)	(31.9)	(51.0)
Rater, r	0.03	0.12	0.09	0.06	0.04	0.00	0.03	0.00	0.01	0.04
	(8.4)	(18.0)	(12.7)	(9.2)	(7.0)	(1.1)	(7.2)	(0.0)	(4.7)	(4.7)
(l:c)r	0.06	0.02	0.04	0.05	0.04	0.03	0.06	0.02	0.03	0.03
	(16.7)	(3.2)	(5.2)	(8.4)	(6.5)	(7.9)	(12.6)	(4.6)	(10.8)	(3.7)
(s:l:c)r, res	0.13	0.21	0.32	0.15	0.27	0.24	0.27	0.17	0.07	0.21
	(38.2)	(32.9)	(43.2)	(23.1)	(47.3)	(58.7)	(61.0)	(40.7)	(22.4)	(23.5)
LR test										
χ^2	16.06	22.60	11.25	22.49	7.23	95.36	63.73	23.69	5.38	18.29
p	0.042	0.003	0.188	0.004	0.406	<0.001	<0.001	0.001	0.496	0.019

Note. PUR = Purpose, IC = Intellectual Challenge, CD = Classroom Discourse, ROC = Representation of Content, CPK = Connections to Prior Knowledge, MOD = Modelling, SUI = Strategy Use, FEED = Feedback, BM = Behavior Management, TM = Time Management.

its measurement across subjects. Fig. 2 shows the results of the G Studies graphically, separating what contributes to true score variance from what contributes to error variance. The error variances tell us about the challenges of measuring the construct in each subject. Here, we are most interested in the rater and rater-by-lesson effects, as these capture more systematic errors. The rater effect captures within-subject differences in rater leniency/harshness. If raters systematically show different levels of leniency/harshness within subjects, it is more likely that there are average differences in rater leniency/harshness across subjects, which would imply a lack of scalar invariance and call into question comparisons of mean scores. It might also lead to challenges in interpreting within-subject mean differences in scores if raters are not randomly assigned to score. For example, the high levels of the rater effect for Strategy Use in mathematics would raise questions about whether different raters are using the same scale when assigning scores.

The rater-by-lesson effect captures systematic differences in how raters score individual lessons. Such differences would imply that scores are not being invariantly assigned to individual lessons, calling comparisons of mean scores across lessons into question. For example, the rater-by-lesson effect for Classroom Discourse is much larger in mathematics than the corresponding lesson and classroom effects. This would imply that more variation in scores is associated with the choice of which rater is assigned to score a lesson than with the lesson itself, which raises serious questions about the ability to use observation scores to compare levels of the latent constructs.

On the other hand, the sources of true score variance provide us with model-based estimates of the extent to which the latent construct varies across segments, lessons, and classrooms. For example, if we look at Intellectual Challenge, the variation of the construct across individual segments (within lesson and classrooms) in language arts is much greater than the overall variation of the construct in math classrooms (including all true score sources of variation). This might say something about how language arts classrooms are organized, with alternating periods of higher and lower levels of cognitive demand, versus math classrooms where construct levels are more constant within the studied setting. At the same time, for Intellectual Challenge, there is approximately the same amount of variation between classrooms across subjects, suggesting the same levels of between-classroom variation in the latent construct across settings. By segregating sources of variation that reflect the intended construct from variation due to sources of error, the G Study results can provide interesting information about the latent constructs even if measurement subject-specificity prevents direct comparisons of raw observation scores. It is important to note, though, that, as with any statistical model, these are model-based estimates that are only as good as the models are accurate.

3.2. Decision studies

With respect to our framework (see Fig. 1), large differences in the precision of scores indicate that the amount of error and signal varies across subjects. This means that it is likely for the respective constructs to exhibit measurement subject-specificity. To summarize the results from the G Studies, Table 4 presents standard errors and reliabilities (D Studies) for the PLATO elements aggregated to either the lesson-level or the classroom-level, which involves generalizing across lessons. Note that these are *absolute* measures which take rater bias into account (i.e., some raters consistently assigning higher scores than others).

In line with the concept of approximate measurement invariance (e.g., Senden et al., 2023), we apply a stricter (0.10) and a more lenient (0.20) cut-off value to discuss differences in reliabilities across subjects for the PLATO elements. If the difference is larger than the cut-off value, we conclude measurement subject-specificity for the respective element. We see that the estimated reliabilities for *lesson-level* scores are very similar across subjects for Intellectual Challenge, Classroom Discourse, and Strategy Use. That is, the difference in reliabilities across subjects is less than 0.10 for these elements. Using the same cut-off, all other PLATO elements indicate measurement subject-specificity with respect to lesson-level scores. Given a more lenient value of approximate measurement invariance, the number of PLATO elements indicating subject-specificity drops to four (i.e., Purpose, Modeling, Feedback, and Behavior Management).

Interestingly and for the most part, (non-)invariant elements vary with respect to the unit of measurement. For classroom-level scores, we see that half of the PLATO elements could be regarded as invariant with a strict cut-off value of 0.10. Considering 0.20 instead, four elements indicate measurement subject-specificity (i.e., Classroom Discourse, Modeling, Strategy Use, and Behavior Management), two of which are the same for lesson-level scores. For an example, looking at the element of Classroom Discourse, there is a large amount of lesson-level variation in scores in language arts and more segment-level variation in scores in mathematics. If we were to aggregate scores to the classroom-level (assuming three lessons per classroom, three segments per lesson, and two raters), the relative percentage of true score variation would be 51 % in math and 24 % in language arts (Table 4), which would imply that after aggregating to the lesson level, a one-point change in observed scores would not correspond to the same change in true score variation, which suggests measurement subject-specificity. Taken together, this suggests that measurement subject-specificity does not only depend on the broader teaching quality construct, but also the level to which scores are aggregated.

4. Discussion

One of our aims in this paper is to help clarify the concept of subject-specificity that is frequently used in research on teaching quality, but in the literature is applied to constructs, observation systems, teaching practices, and scores stemming from classroom observation. From this we conclude that the meaning of this concept is not entirely clear (Jentsch & Kirchhoff, 2024). To move forward, and to promote exploring subject-specificity with empirical methods (Wemmer-Rogh et al., 2024), we have presented a framework from differential psychology that unpacks measurement equivalence in detail (Fischer et al., 2025) and used this to discuss three different types of subject-specificity (i.e., functional, operational, measurement). These three types are organized hierarchically, that is, they make more assumptions on "genericness" of teaching quality across subjects from top to bottom. For instance, when empirically exploring subject-specificity in measurement one needs to make the claim that teaching quality constructs and measures are equivalent across subjects. We believe that this framework could help researchers be more explicit about how they think about subject-specificity and subject-genericness in their respective studies.

One important takeaway from this paper is the need to specify precisely what is meant by subject-specificity. Developers and users of observation systems should specify whether the constructs being measured are functionally or operationally subject-specific from a theoretical perspective. Being clear in the claims made will help facilitate conversations across researchers working in different subjects. It will also help clarify efforts to build generic, synthetic frameworks. For example, the MAIN-TEACH model is a recent effort to synthesize an overarching framework that covers all (or at least the majority) of constructs measured by observation systems (see Charalambous & Praetorius, 2020; Litke et al., 2021). Efforts like this would seem to implicitly assume no functional subject-specificity since they place constructs from subject-specific observation systems into generic categories, though the potential for operational subject-specificity appears to be acknowledged by the framework (Litke et al., 2021). Functional subject-specificity would undermine the very nature of MAIN-TEACH by precluding the possibility of grouping together constructs measured in different subjects. After explicitly making and supporting theoretical claims about functional and operational subject-specificity, developers and users of observation systems should explicitly empirically test for measurement subject-specificity for those constructs where there is

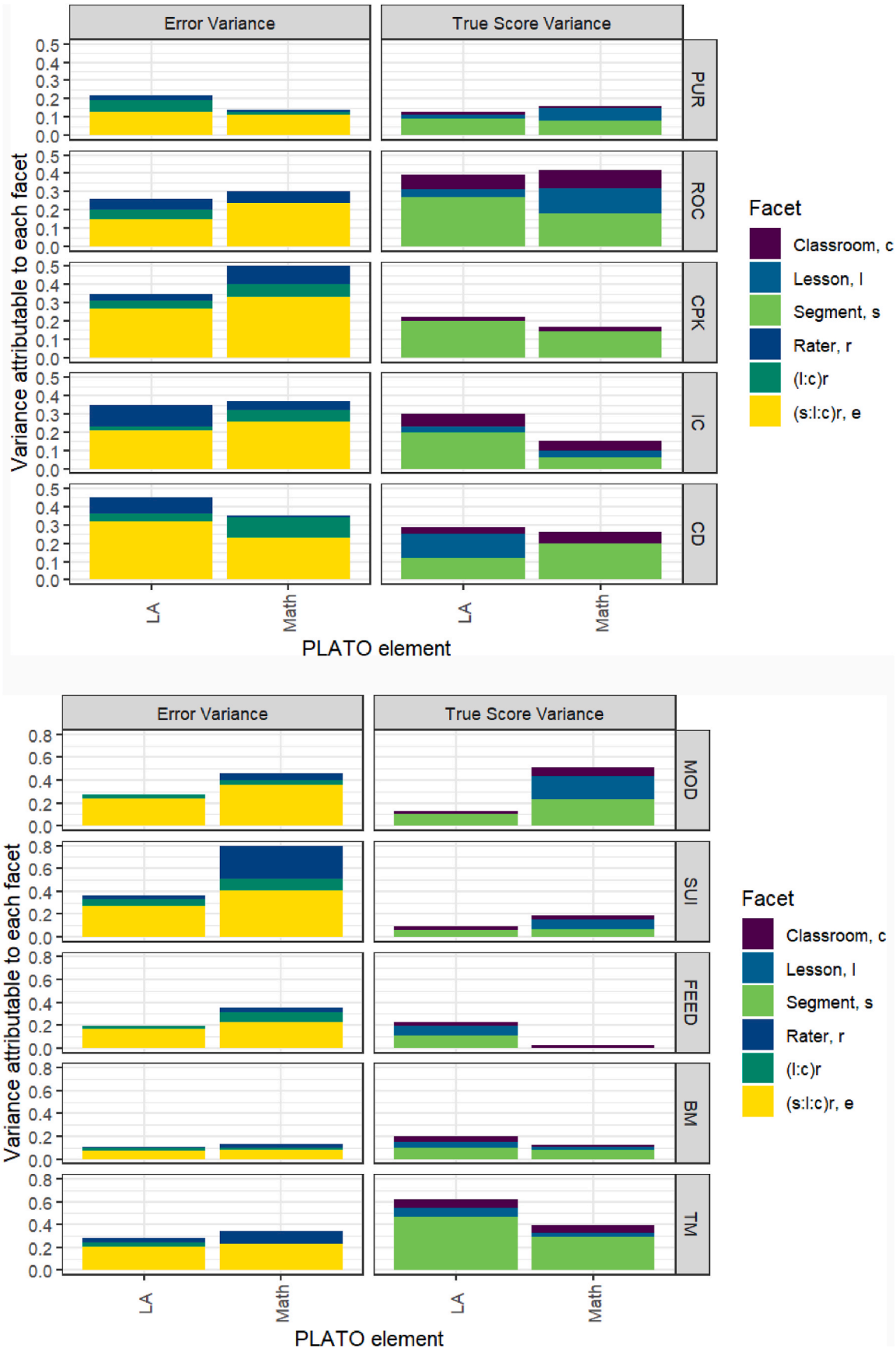


Fig. 2. Construct relevant and irrelevant variance across subjects and PLATO elements
PUR = Purpose, IC = Intellectual Challenge, ROC = Representation of Content, CPK = Connections to Prior Knowledge, MOD = Modelling, SUI = Strategy Use, FEED = Feedback, CD = Classroom Discourse, BM = Behavior Management, TM = Time Management.

Table 4
Lesson-level and classroom-level Decision Studies by subject for all PLATO elements involving three segments and one rater per lesson as well as three lessons and two raters per classroom.

	Elements									
	PUR	IC	CD	ROC	CPK	MOD	SUI	FEED	BM	TM
Lesson-level										
Mathematics										
Reliability	0.62	0.36	0.39	0.68	0.21	0.62	0.21	0.09	0.47	0.51
Language Arts										
Reliability	0.34	0.44	0.47	0.57	0.34	0.37	0.22	0.67	0.68	0.69
Difference across subjects	0.28	−0.08	−0.08	0.11	−0.13	0.25	−0.01	−0.58	−0.21	−0.18
Classroom-level										
Mathematics										
Reliability	0.18	0.42	0.51	0.48	0.24	0.35	0.15	0.30	0.19	0.35
Language Arts										
Reliability	0.29	0.39	0.24	0.47	0.24	0.50	0.39	0.44	0.55	0.38
Difference across subjects	−0.11	0.03	0.27	0.01	0.00	−0.15	−0.24	−0.14	−0.36	−0.03

Note. PUR = Purpose, IC = Intellectual Challenge, ROC = Representation of Content, CPK = Connections to Prior Knowledge, MOD = Modelling, SUI = Strategy Use, FEED = Feedback, CD = Classroom Discourse, BM = Behavior Management, TM = Time Management.

neither functional nor operational subject-specificity.

The second aim was to explore measurement subject-specificity in scores stemming from an established observation system that has been used to explore differences in teaching quality across subjects (Cohen et al., 2018). Since PLATO emphasizes the use of element-level (i.e., item-level) scores, typical reflective modelling approaches to invariance testing were not possible in this case. This led us to turn to a G Study, which isolates and estimates variation in scores that should reflect true variation in the latent construct (i.e., true score variation) as well as variation that reflects sources of error. This approach takes for granted that elements are operationalized correctly while looking at rater error as a source for subject-specificity. This provides a very different insight into subject-specificity in contrast to past studies that adopted reflective modelling or have ignored rater error and instead tried to understand subject-specificity based on intercorrelations of different item scores (e.g., Wemmer-Rogh et al., 2024). New research has called into question whether such approaches are appropriate for observation scores (e.g., White, 2025; White et al., 2024). Based on the framework presented here, assumptions on functional and operational subject-specificity involve theoretical considerations on the composition of latent constructs rather than empirical questions, but we were able to test for measurement subject-specificity for PLATO elements that are not functionally or operationally subject-specific using G Theory.

We found small to modest differences in mean scores across subjects, but it is unclear to what extent these differences reflect true differences in the latent construct as we lack the ability to examine if raters are, on average, more or less lenient across subjects (i.e., we cannot establish scalar invariance). This limitation is generally true for all observational studies, as most research focuses on rater disagreement rather than rater accuracy and average levels of rater harshness/leniency are invisible in such study designs (White & Ronfeldt, 2024). That said, the results of the G Studies show relatively large variation for many elements due to the rater effect (i.e., leniency/harshness for individual raters), which would indicate the potential for average differences in rater leniency/harshness across subjects. The elements of Purpose and Behavior Management had relatively little variation due to the rater effect, which would make mean comparisons more trustworthy.

Beyond these mean differences, we highlight that measurement subject-specificity is dependent on the level to which scores are aggregated. For individual elements, this is true when scores have the same reliability (or percentage of true score variation). Reliability, in turn, depends on levels of aggregation. This is important because scores could potentially be informative across subjects to provide feedback to teachers on a single lesson but exhibit measurement subject-specificity when aggregated to the classroom level, limiting their usefulness across subjects for long-term decision-making (and vice versa).

Regardless of whether measurement subject-specificity occurs, G Studies can provide highly meaningful data that directly impacts the usefulness of scores. For example, as noted above, the high levels of the rater effect in some elements raises important questions about the extent to which mean scores can be compared when those scores are assigned by different raters. Since the rater effect is estimated based on the average difference of scores between raters, this relies on the random assignment of raters to properly estimate rater effects. Non-randomly assigned raters would pose serious questions for the interpretation of this effect by conflating true values of the construct with rater harshness/leniency. The rater-by-lesson effect also poses important challenges for using observation scores as it implies that raters are not using the same scale when assigning scores to different lessons (Engelhard & Wind, 2018). This would lead to difficulties in comparing lesson or classroom scores across raters.

Since G Studies separate out sources of true score variation from error, even in the case of measurement subject-specificity, we can use those true scores to understand variation in the latent constructs. Here, we see an overall large amount of variation in the constructs across segments within-lessons, suggesting that most constructs are not equally present throughout entire lesson periods, but present for certain aspects of lessons. For example, the element Connections to Prior Knowledge varies mostly within lessons, as this construct captures specific instances that teachers link new content to previously learned content and such connections tend to be done during the introduction of new material rather than consistently across lessons. As another example, we see much higher variation of Intellectual Challenge in language arts compared to math, suggesting that language arts lessons have more within-lesson variability in the cognitive demand of tasks that students engage in while the between-lesson variability in Intellectual Challenge is approximately the same across classrooms, suggesting that different classrooms have approximately the same variation in average cognitive demand of tasks across subjects. This information provides useful understandings of the constructs being measured by PLATO.

4.1. Benefits and challenges in using PLATO across subjects

It is a benefit of PLATO that it has demonstrated applicability to a variety of purposes and contexts so far. For instance, there is evidence that the observation system can be used in the Norwegian context without much adaption (Klette et al., 2017), and it seems to be at least moderately transferable between subjects (Cohen et al., 2018). However, more research is needed to explore the extent to which PLATO must be adapted if it should be applied to additional subjects in the

future or, in contrast, focus more closely on its original purpose (i.e., student learning in language arts). Bearing in mind that the observation system involves elements exhibiting functional subject-specificity, another way to move forward could be to explore whether PLATO could help shed light on the extent to which teaching patterns vary across topics and/or lessons *within* language arts.

A critical aspect of any observation system is that it relies on qualified observers with some knowledge of the subjects and, equally important, with experience in the respective educational context. This means that observers are required to do some hours of training until they can use the rubrics correctly and efficiently. Thus, applying the instrument in practice requires many resources which might not always be available. At the same time, more studies are needed on ways of effectively training raters, as well as recommendations for calibration (e.g., to what extent master raters should be involved, how often raters should be re-trained, which cut-off values for rater agreement are reasonable).

Another important aspect of PLATO involves the overall conceptualization of teaching quality dimensions. While an obvious point is that the observation system to this date does not contain elements specifically designed for mathematics but for language arts, a critical discussion of how PLATO is conceptually intertwined with the context it was first developed for is beyond this paper.

4.2. Limitations

It is important to highlight some limitations of this study to help make sense of our findings. First, we remind readers that within each subject, classrooms were confounded with teachers. That is, every classroom was taught by a unique teacher (see Cohen et al., 2018, for a different study design, in which teachers taught both subjects). Although we consider teaching quality to be co-constructed by teachers and students rather than being a characteristic of teachers, we are not able to distinguish between teacher and classroom effects in this study.

Then, we note that a large amount of variance remained unexplained for most, if not all PLATO elements, indicating that raters disagreed on scores for specific segments after accounting for other types of error (White & Ronfeldt, 2024). It would be useful to further explore this issue (e.g., have raters score the same segments across spaced intervals). As a result of this and other rater errors, the estimated reliabilities were only moderate, which implies that scores should not be used for high-stake decision making. This would also entail that using classroom-level scores might suffer from low precision that could potentially affect correlations between scores of teaching quality and student outcomes.

Future studies might want to consider further aspects of teaching and learning that could have an impact on scores. For instance, any didactical decision a teacher needs to take in the classroom has the potential to affect scores, or, if such decisions vary systematically between subjects, explain subject-specificity in measurement (e.g., how frequently topics are changed, the amount of collaborative learning in classrooms, the extent to which digital technologies are being used). If lessons were coded using more meaningful segments rather than equal-interval segments, and segments were coded with information about the didactical nature of the included instruction (e.g., topic, grouping format), scores may provide more detailed information about the nature of instruction (White et al., 2022).

5. Conclusions

Scholars have recently argued that research on teaching quality should consider subject-specificity to a larger degree. However, there is little consensus to what this means exactly, as subject-specificity may equally refer to constructs, measures, scores, and more. This limits the extent to which empirical work can contribute to our understanding of subject-specificity. We suggest one way to do so by assuming that subject-specificity is essentially a characteristic of a construct. Then, we can explore differences in the measurement of teaching quality across subjects by claiming functional and operational equivalence (Fischer et al., 2025). Our results indicate that scores in Instructional Scaffolding show the largest variation between subjects, suggesting different patterns of teaching practices in math and language arts classrooms. With respect to future research, we recommend researchers to be more explicit about their assumptions on subject-specificity, for example, by organizing them along the terms of functional, operational, and measurement equivalence.

CRediT authorship contribution statement

Armin Jentsch: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Mark C. White:** Writing – review & editing, Validation, Methodology, Conceptualization. **Kirsti Klette:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

None.

Acknowledgements

This work was funded by the Research Council of Norway under research grant #300791. An earlier version of this paper was presented at the AERA conference in 2024.

APPENDIX A. SCORING DISTRIBUTIONS FOR ALL PLATO ELEMENTS BY SUBJECT

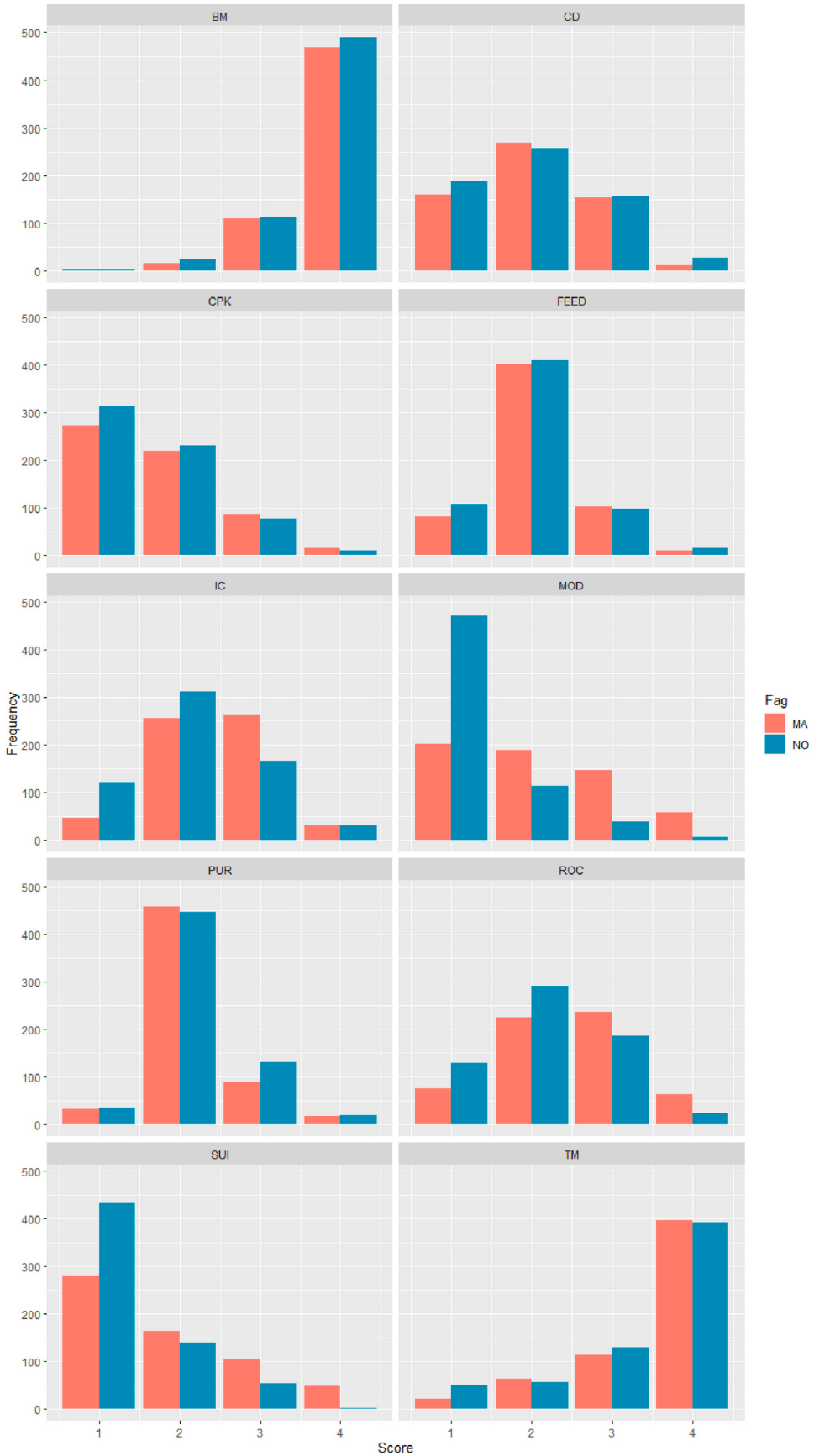


Fig. A. Scoring distributions for all PLATO elements by subject (segment level). We see that for most PLATO elements, distributions are similar across subjects. However, for several elements, the percentage of the second-to-highest score (e.g., Intellectual Challenge) or the lowest score (e.g., Modeling, Strategy Use) varies considerably between mathematics and language arts lessons.

PUR = Purpose, IC = Intellectual Challenge, ROC = Representation of Content, CPK = Connections to Prior Knowledge, MOD = Modelling, SUI = Strategy Use, FEED = Feedback, CD = Classroom Discourse, BM = Behavior Management, TM = Time Management.

Data availability

Data will be made available on request.

References

Blömeke, S., Jentsch, A., König, J., & Kaiser, G. (2022). Opening up the black box: Teacher competence, instructional quality, and students' learning progression. *Learning and Instruction, 79*, Article 101600. https://doi.org/10.1016/j.learninstruc.2022.101600

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

Casabianca, J. M., McCaffrey, D., Gitomer, D., Bell, C., Hamre, B., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757–783. https://doi.org/10.1177/0013164413486987

Charalambous, C., & Praetorius, A.-K. (2018). Studying mathematics instruction through different lenses: Setting the ground for understanding instructional quality more comprehensively. *ZDM Mathematics Education, 50*(3), 355–366. https://doi.org/10.1007/s11858-018-0914-8

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

Cohen, J., Ruzek, E., & Sandilos, L. (2018). Does teaching quality cross subjects? Exploring consistency in elementary teacher practice across subjects. *AERA Open, 4*(3), 1–16. https://doi.org/10.1177/2332858418794492

Cronbach, L. J., Glaser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley.

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*. https://doi.org/10.1177/10693971241312459. online first.

Grossman, P., Cohen, 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. https://doi.org/10.3102/0013189X14544542

Grossman, P., & Stodolsky, S. (1995). Content as context: The role of school subjects in secondary school teaching. *Educational Researcher, 24*(8), 5–23. https://doi.org/10.3102/0013189X024008005

Jentsch, A., & Kirchhoff, P. (2024). Was bedeutet Fachspezifität? Überlegungen und Beispiele zu fächerübergreifender sowie fachdidaktisch konkretisierter Unterrichtsqualitätsforschung [What is subject-specificity? Considerations on and examples of instructional quality from a subject-specific and an interdisciplinary research perspective]. In V. Lohe, A. Lindl, & P. Kirchhoff (Eds.), *Unterrichtsqualität in schulischen Fremdsprachen. Theoretische Ansätze und empirische Ergebnisse aus den Fachdidaktiken* (pp. 39–55). Waxmann. https://doi.org/10.31244/9783830999201.02.

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), 10. https://doi.org/10.5617/adno.4729

Kyriakides, L., Creemers, B., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: The contribution of the dynamic model. *ZDM Mathematics Education*. https://doi.org/10.1007/s11858-018-0919-3

Litke, E., Boston, M., & Walkowiak, T. A. (2021). Affordances and constraints of mathematics-specific observation frameworks and general elements of teaching quality. *Studies In Educational Evaluation, 68*, Article 100956. https://doi.org/10.1016/j.stueduc.2020.100956

Magnusson, C. G., Luoto, J. M., & Blikstad-Balas, M. (2023). Developing teachers' literacy scaffolding practices—successes and challenges in a video-based longitudinal professional development intervention. *Teaching and Teacher Education, 133*, Article 104274. https://doi.org/10.1016/j.tate.2023.104274

Praetorius, A. K., Jentsch, A., Luoto, J., Keller, S. D., & Fauth, B. (2025). Context in research on teaching quality: Bringing a complex idea into the spotlight. *School Effectiveness and School Improvement, 36*(2), 263–296. https://doi.org/10.1080/09243453.2025.2482577

Praetorius, A.-K., Vieluf, S., Saß, S., & Bernholt, A. (2015). The same in German as in English? Investigating the subject-specificity of teaching quality. *Zeitschrift für Erziehungswissenschaft, 19*, 191–209. https://doi.org/10.1007/s11618-015-0660-4

Schlesinger, L., & Jentsch, A. (2016). Theorerical and methodological challenges in measuring instructional quality related in mathematics education. *ZDM Mathematics Education, 48*(1–2), 29–40. https://doi.org/10.1007/s11858-016-0765-0

Senden, B., Teig, N., & Nilsen, T. (2023). Studying the comparability of student perceptions of teaching quality across 38 countries. *International Journal of Educational Research Open, 5*, Article 100309. https://doi.org/10.1016/j.ijedro.2023.100309

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. SAGE Publications. https://doi.org/10.1002/9781118445112.stat00068

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

van de Vijver, F. J. R., Jude, N., & Kuger, S. (2019). Challenges in international large-scale educational surveys. In *The SAGE handbook of comparative studies in education* (pp. 83–99). SAGE Publications Ltd. https://doi.org/10.4135/9781526470379.

Wemmer-Rogh, W., Grob, U., Charalambous, C. Y., & Praetorius, A.-K. (2024). Measurement invariance between subjects: What can we learn about subject-related differences in teaching quality? *ZDM Mathematics Education, 56*, 831–844. https://doi.org/10.1007/s11858-024-01622-7

White, M. (2025). A peculiarity in psychological measurement practices. *Psychological Methods*. https://doi.org/10.1037/met0000731

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*, 1–25. https://doi.org/10.1080/0969594X.2024.2368252

White, M., & Ronfeldt, M. (2024). Monitoring rater quality in observational systems: Issues due to unreliable estimates of rater quality. *Educational Assessment, 29*(2), 124–146. https://doi.org/10.1080/10627197.2024.2354311

Update

Teaching and Teacher Education
Volume 169, Issue , January 2026, Page

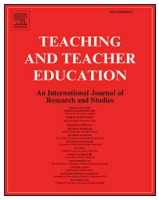
DOI: https://doi.org/10.1016/j.tate.2025.105299

Teaching and Teacher Education 169 (2026) 105299

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate



Corrigendum

Corrigendum to "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 (2026) 105291]

Armin Jentsch [a,*], Mark C. White [b], Kirsti Klette [b]

[a] University of Cologne, Germany
[b] University of Oslo, Norway

The authors regret the misidentification of affiliation in the original article. Mark C. White and Kirsti Klette are not affiliated with University of Cologne.

The authors would like to apologise for any inconvenience caused.

DOI of original article: https://doi.org/10.1016/j.tate.2025.105291.

* Corresponding author. University of Cologne, Faculty of Human Sciences, Gronewaldstr. 2, Cologne, 50931. Germany.
E-mail address: ajentsc2@uni-koeln.de (A. Jentsch).

https://doi.org/10.1016/j.tate.2025.105299

Available online 7 November 2025