ZDM – Mathematics Education
https://doi.org/10.1007/s11858-024-01557-z

ORIGINAL PAPER

Effects of observation mode on ratings of teaching quality in secondary mathematics classrooms

Armin Jentsch[1] · Kirsten Benecke[2] · Sigrid Blömeke[3] · Johannes König[4] · Gabriele Kaiser[2,5]

Accepted: 5 February 2024

Abstract
In educational research, teaching quality is extensively studied because of its role of a mediator between teacher characteristics and student learning. However, empirical evidence on differences between video and live scoring of teaching quality is rare. In the present study, thirty lessons from 15 secondary mathematics classrooms in a German metropolitan area were observed. Lessons were scored both live in the classroom and using video recordings. Live and video scoring was conducted by (different) trained observers. Ratings were obtained with a "hybrid" observational instrument that covers generic and subject-specific characteristics of teaching quality in mathematics classrooms. Generalizability analysis and paired *t* tests were performed to investigate mode effects. The findings showed that in live scoring, classroom management was rated lower, and cognitive activation was rated higher. Rankings of lessons or classrooms were very similar across modes, and reliabilities did not differ to a meaningful extent either, except for classroom management reaching better results for live ratings. This suggests that based on the present findings, classroom observation performed with our hybrid framework of teaching quality generalizes across observation mode only under certain circumstances. Further research is necessary to better understand the relation between observation mode and teaching quality ratings, as well as the impact of the scoring procedures. We discuss the implications of our findings for educational research and practice.

Keywords Teaching quality · Classroom observation · Mode effect · Generalizability analysis · Mathematics instruction

✉ Armin Jentsch
armin.jentsch@ils.uio.no

✉ Gabriele Kaiser
gabriele.kaiser@uni-hamburg.de

Kirsten Benecke
kirsten.benecke@uni-hamburg.de

Sigrid Blömeke
sigrid.blomeke@cemo.uio.no

Johannes König
johannes.koenig@uni-koeln.de

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

[2] Faculty of Education, University of Hamburg, Von-Melle-Park 8, 20146 Hamburg, Germany

[3] Center for Educational Measurement Oslo (CEMO), Gaustadalléen 30d, Oslo 0373, Norway

[4] Faculty of Human Sciences, University of Cologne, Gronewaldstr. 2a, 50931 Cologne, Germany

[5] Faculty of Education, Faculty of Education and Arts, University of Hamburg, Germany and Nord University, Universitetsalléen 11, Bodø 8026, Norway

1 Introduction

Teaching quality has been researched extensively in the past years with a high number of empirical studies in educational sciences and psychology (Bill and Melinda Gates Foundation, 2012; Charalambous & Praetorius, 2018), as it is regarded an important mediator between teacher competence and student learning (Baumert et al., 2010; Blömeke et al., 2022; Nilsen et al., 2018). To better understand how learning develops in the classroom, scholars are concerned with the reliable and valid measurement of teaching quality (e.g., Hill et al., 2012; Jentsch et al., 2020; Ryan et al., 1995). In doing so, Helmke (2012) considers classroom observation as the "gold standard" amongst other ways of capturing teaching quality (e.g., student ratings in large-scale assessment) because of its direct assessment of teaching practices.

However, current research has shown that classroom observation suffers from several methodological issues (e.g., segment length, rater bias, measurement error, stability over time, Leckie & Baird, 2011; Mashburn et al.,

Published online: 13 February 2024

2014; Praetorius et al., 2014; White & Klette, 2023), one of which are mode effects (Casabianca et al., 2013; Jaeger 1993). *Mode effects* are differences in scores that are due to observation mode rather than true variation in the latent construct. They are therefore a potential danger to the validity of the inferences drawn from data (Bell et al., 2012; Kane, 2013). If mode effects occur, live and video ratings might not yield the same findings, although the same frameworks or measures are applied to capture teaching quality.

To the best of our knowledge, only one study has been conducted so far to rigorously analyze the differences between live and video ratings of teaching quality (Casabianca et al., 2013; see also Frederiksen et al., 1992).[1] Casabianca and colleagues (2013) investigated variation across observation modes in 82 US-American classrooms during one year of schooling. Ratings were obtained with the established *Classroom Assessment Scoring System* (CLASS, Pianta et al., 2008), which captures three generic dimensions of teaching quality (classroom organization, emotional support, instructional support). The scholars found that video ratings were slightly higher, which was explained by a time lag of about 100 days between live and video scoring. They concluded that the observed differences were not due to observation modes but to raters' increased experience over time. Additional correlation analysis showed that live and video ratings resulted in similar rankings of classrooms.

In the present study, we apply a design similar to the one by Casabianca et al. (2013) to analyze live and video ratings of teaching quality. However, this study is set in a different educational context (i.e., mathematics classrooms in secondary schools in a German metropolitan area) and draws on a *hybrid* conceptual framework (Charalambous & Praetorius, 2018) which also takes subject-specific characteristics of teaching quality into account. We investigate to what extent differences in observation modes could be associated to how scores are assigned to classrooms or lessons (*absolute* decisions), and how classrooms or lessons are ranked (*relative* decisions), as well as measurement error and generalizability of teaching quality scores (Cronbach et al., 1972).

2 Conceptual framework

2.1 Teaching quality in mathematics classrooms

Following the *TIMSS 1995 Video Study* (Stigler et al., 1999), German educational researchers have developed a generic framework of teaching quality with three basic dimensions (Klieme et al., 2006) which are classroom management, student support, and (potential for) cognitive activation. The three basic dimensions have been shown to positively relate to students' achievement in mathematics classrooms across several studies and various operationalizations (e.g., Baumert et al., 2010; Lipowsky et al., 2009; for an overview see Praetorius et al., 2018). *Classroom management* refers to teachers' procedures and strategies that enable efficient use of time (time on task), as well as behavioral management (Helmke, 2012; Kounin, 1970). *Student support* draws on self-determination theory (Deci & Ryan, 1985) and aims at both motivational and emotional support, as well as individualization and differentiation. *Cognitive activation*, finally, addresses opportunities for "high-order thinking" from a socio-constructivist perspective on teaching and learning (e.g., problem-solving, Mayer, 2004; Shuell, 1993). According to Klieme and Rakoczy (2008), cognitive activation should be operationalized with regard to subject-specific differences to better understand student learning in the corresponding domains (e.g., modelling tasks in mathematics vs. classroom discourse in history vs. text-based instruction in language arts). In a similar vein, scholars in mathematics education have argued that generic operationalizations of the three basic dimensions might not address all the characteristics that are relevant to teaching quality in mathematics classrooms (e.g., general mathematical competencies according to national curricular standards, Blum et al., 2015, or mathematical correctness, Brunner, 2018; see also Learning Mathematics for Teaching Project, 2011).

Empirical evidence suggests that generic and subject-specific measures of teaching quality generate moderately correlated, but still unique information about classrooms (Kane & Staiger, 2012). Evaluating this finding, Charalambous and Praetorius (2018) conclude that subject-specific and generic measures together could explain more variance in student learning in mathematics than generic measures alone. Since subject-specificity might be considered a continuum rather than a binary characteristic, they argue that it could be meaningful for scholars to develop *hybrid* frameworks of teaching quality, which take both perspectives into account (i.e., generic and subject-specific, see also Charalambous & Praetorius, 2018).

In the present study, we apply such a hybrid framework (Schlesinger et al., 2018). It draws on the three basic dimensions but adds a fourth dimension (*mathematics educational structuring*, Jentsch et al., 2021; see also Drollinger-Vetter, 2011; Kleickmann et al., 2020) to capture additional characteristics of teaching quality that are relevant to student learning in secondary mathematics classrooms. Mathematics educational structuring refers to teaching practices that provide cognitive or instructional support to students when

[1] The study of Frederiksen et al. (1992) involved only four teachers, which is why we refrain from discussing it in more detail.

building up knowledge (e.g., mathematical correctness, explanations, consolidation). This dimension of teaching quality complements the three basic dimensions by teachers' efforts in adapting cognitive challenges to students' individual characteristics and is closely connected to scaffolding in instruction (e.g., Van de Pol et al., 2010). Jentsch et al. (2021) provide empirical evidence for a four-dimensional structure underlying observer ratings, which corresponds to the three basic dimensions and mathematics educational structuring (see Kleickmann et al., 2020, for a similar finding in science classrooms using student ratings). Furthermore, they find that the latter is also related to mathematics teachers' competence. Blömeke et al. (2022) show that teaching quality as modeled with this framework is connected to student achievement in mathematics.

2.2 Validity arguments for classroom observation

Modern validity theory (Bell et al., 2012; Kane, 2013) states that measures usually serve a specific purpose, and to evaluate the validity of test scores in a meaningful way, such purposes must be considered. We argue alongside Casabianca et al. (2013) that classroom observation can serve at least two different purposes, and these are related to the unit of analysis (e.g., classroom, lesson). Classroom-based (in contrast to lesson-based) conclusions refer to classrooms as the unit of analysis in teaching quality research. They are typically driven by long-term decisions that need to generalize teaching practices over a period involving many lessons, for instance when connecting teaching quality measures to student learning. On the other hand, lesson-based conclusions refer to lessons as the unit of analysis. They are drawn to provide feedback to teachers on a particular topic or classroom setting. As both are widely used in educational research, policy, and practice, we take these two perspectives into account in the present study.

To develop a validity argument specifically suited for measures of teaching quality, Bell et al. (2012) discuss scoring, generalization, extrapolation, and implication as the four main inferences drawn from classroom observation. The *scoring* assumption refers to the appropriateness, accuracy, and consistency of the scoring procedure. *Generalization* means that "the sample of teaching observed is representative of all the instances of teaching to which one wants to generalize" (Bell et al., 2012, p. 67). *Extrapolation* relates teaching quality scores to other meaningful concepts within a theory (e.g., see the offer-use model by Helmke, 2012). Finally, the *implication* inference connects teaching quality scores to decisions that are based on them (e.g., pass/fail grades in practical teacher examinations).

The generalization inference is particularly important, because it refers to the degree to which the observed scores reflect the targeted construct, rather than unintended sources of variation (e.g., mode effects). Towards this end, researchers should provide evidence that inferences drawn from scores observed in a study do not largely depend on the conditions under which it was conducted. Generalizing across observation mode, therefore, makes the claim that the corresponding scores do not differ significantly regarding scoring distribution, ranking of lessons or classrooms, as well as measurement error and reliability.

2.3 Classroom observation mode: live versus video scoring

Observation mode is important to the assessment of teaching quality because of different procedures for data collection in educational research and practice. Due to pragmatic reasons, live scoring is usually performed in educational practice (e.g., school inspection), whereas educational research often applies video scoring. Live scoring entails the advantage of observers being physically in the classrooms, while using video has the benefit that it can be watched many times. Beyond the possibility to obtain multiple ratings (e.g., to decrease measurement error), teachers may find video useful for professional development activities, as they are able to evaluate their performance on their own or with peers (Brunvard, 2010; Sherin & Han, 2004; Van Es & Sherin, 2010). However, capitalizing on these benefits with a framework that was originally developed for live scoring (or vice versa) needs careful evaluation for mode effects, as these might imply increased measurement error or bias.

Casabianca et al. (2013) argue that live and video scoring differ in how raters access information on the lessons they are observing. For instance, during live observation, raters can in principle pay attention to any action of students and teachers, as well as to tasks and material at all times. While this may include ambient audio information, raters' possibilities to capture one-to-one conversations (i.e., teacher/student or student/student) could be limited. On one hand, being able to always observe all students is important to adequately score classroom management because students' individual time-on-task can be taken into account. Ambient audio, on the other hand, might be helpful as contextual information to adequately consider potential disruptions during scoring. Both pieces of information could also help raters to understand to what extent students are cognitively activated in the classroom. For instance, if raters can observe students' reactions to potentially challenging tasks, they might have a chance to capture the amount of productive struggle that students are involved in.

During video observation, raters' attention is necessarily drawn to what the cameras have captured. This might increase the amount of standardization because raters

receive similar information at the same time, and therefore lead to higher reliability in scores. What is more, the perceived audio information is different from the live scoring. In most settings teachers are equipped with additional microphones, which ensures that teachers' voices are always heard (Casabianca et al., 2013). This could have an impact on raters' capabilities of scoring how teachers support individual students (e.g., feedback, scaffolding) because these practices usually occur during one-to-one conversations or group work. Thus, raters might benefit from the additional audio information that is accessible to them during video scoring.

2.4 Research questions

Mode effects could lead to different score interpretations on the same construct and following Bell et al. (2012), are therefore a danger to validity. The goal of our study is to investigate to what extent our teaching quality framework is dependent on whether live or video scoring is applied. Given the hybrid nature of our framework, it is also of interest whether mode differences are more likely to occur with generic or subject-specific dimensions of teaching quality. As standardized classroom observations can be used for both absolute and relative decisions, we analyze differences in teaching quality mean scores, as well as differences in rank orders for lessons and classrooms (i.e., correlation analysis). This is done to explore the degree to which teaching quality scores are associated with observation mode. We address the following research questions (Casabianca et al., 2013), focusing on differences in mode effects between generic and subject-specific teaching quality dimensions:

1. Do raters use the scale of our observational instrument differently across modes? Are there differences in scale use between generic and subject-specific dimensions?
2. To what extent do live and video scores rank lessons or classrooms differently? Are these rankings different for generic and subject-specific dimensions?
3. To what extent do sources of variance (i.e., classrooms, lessons, segments, and raters) compare between scoring modes? Are there differences in variance decompositions between generic and subject-specific dimensions?
4. What are the implications for measurement error and reliability of live and video scores regarding classroom-based as well as lesson-based decisions? Are these implications different for generic and subject-specific dimensions?

3 Methods

3.1 Participants

Data were collected from a subsample of the *Teacher Education and Development Study–Instruct* (TEDS-Instruct). Both TEDS-Instruct and the present study took place in secondary school mathematics classrooms in a German metropolitan area, years 7–10. A convenience (i.e., non-random) subsample of the TEDS-Instruct participants took part in this follow-up study. We observed and video-recorded two lessons of 90 min in every classroom between December 2016 and May 2017, usually within two weeks' time. Fifteen licensed mathematics teachers participated, eight of which were female and seven were male. The teachers' age median was 36 years ($min = 28$, $max = 71$) and they had been teaching for six years on average ($min = 0.5$, $max = 30$).

3.2 Observational instrument

The observational instrument was developed within TEDS-Instruct and consists of 21 high-inference items (see Table 1). It captures three basic dimensions (classroom management, student support, cognitive activation, Praetorius et al., 2018) and mathematics educational structuring, covering more subject-specific characteristics of teaching quality (Jentsch et al., 2021). Raters assign scores on a four-point scale (from 1: very low teaching quality, through 4: very high teaching quality). Classroom management is assessed with three items (e.g., time on task, Cronbach's $\alpha = 0.87$). Student support is captured with four items (e.g., dealing with heterogeneity, $\alpha = 0.73$). Cognitive activation is measured with seven items (e.g., challenging questions, $\alpha = 0.80$). Finally, mathematics educational structuring is also captured with seven items (e.g., mathematical correctness, $\alpha = 0.81$). Additional information on the development of the observational instrument can be found in Schlesinger et al. (2018) and Jentsch et al. (2021).

3.3 Scoring procedure

Lesson scoring was conducted by six extensively trained raters. All of them were student teachers or PhD students in a mathematics education program and had obtained at least a Bachelor's degree. The training took 30–40 h and consisted of both live and video scoring, peer discussions, as well as more theoretical work involving the manual for the observational instrument and additional literature. In a pilot study high rater reliability for all items of the observational instrument was reached ($ICC > 0.80$). Scoring was performed four times per lesson (approx. every 22 min, see Mashburn et al., 2014, for a discussion of the potential

Table 1 Scoring distribution by item, dimension and mode (percentages)

Item	Live				Video			
	1	2	3	4	1	2	3	4
Classroom Management								
Use of time	0.0	3.3	12.5	84.2	1.3	1.3	5.4	92.1
Disturbances	0.0	6.3	33.3	60.4	0.0	0.4	15.0	84.6
Atmosphere	0.0	6.7	25.4	67.9	0.0	2.5	12.9	84.6
Student Support								
Individual	19.6	44.6	30.8	5.0	34.2	33.3	26.3	6.3
Heterogeneity	82.9	15.8	1.3	0.0	71.7	23.3	5.0	0.0
Self-directed	67.5	25.4	7.1	0.0	67.1	27.9	5.0	0.0
Collaboration	41.3	32.1	23.8	2.9	50.4	35.0	10.8	3.8
Cognitive Activation								
Challenge	0.0	25.0	66.3	8.8	5.4	37.5	53.8	3.3
Methods	0.0	20.4	63.3	16.3	2.1	27.9	52.9	17.1
Representations	3.8	31.7	58.3	6.3	8.2	19.8	69.8	2.2
Practice	2.0	50.7	46.7	0.7	3.3	59.2	36.8	0.7
Examples	0.0	6.8	80.1	13.1	3.0	13.8	64.7	18.5
Relevance	12.5	59.2	25.0	3.3	30.0	52.9	16.3	0.8
Depth	2.5	47.9	44.2	5.4	12.9	50.0	35.8	1.3
Mathematics Educational Structuring								
Structure	0.4	17.5	47.1	35.0	1.3	11.7	30.0	57.1
Feedback	0.0	7.6	56.3	36.1	0.8	17.1	36.7	45.4
Co-construction	0.8	23.1	68.1	7.9	1.7	25.1	61.3	2.1
Recalling	6.5	39.2	34.9	19.4	3.8	30.1	58.1	8.1
Errors	0.0	10.5	67.0	22.5	3.6	22.7	60.5	13.2
Correctness	0.0	2.1	17.9	80.0	0.0	2.1	16.3	81.7
Explanations	0.4	11.6	49.6	38.4	0.8	11.3	35.8	52.1

benefits for reliability and validity), and we ensured that live and video scoring took place at the same time points within a lesson. All lesson segments were double-coded (i.e., two independent scores are available for every segment). In addition, raters were allowed to change their scores based on the manual for the observational instrument and peer discussion after the lesson had ended.

All lessons were scored under both observation modes with different raters. Otherwise, procedures were the same for live and video scoring. This means that raters were not allowed to stop the videos during scoring, nor to move around in the classroom to increase the amount of standardization across observation modes. Due to practicalities, however, it was not possible to assign raters randomly to classrooms, lessons, or observation modes (as e.g., in a random block design). This resulted in an uneven distribution of raters across modes, with four raters being assigned more frequently to live scoring, and two raters working mainly with video scoring.

Video scoring was performed within two weeks after the corresponding live observations had taken place to minimize rater drift (Casabianca et al., 2013). To this end, two cameras and a teacher microphone (lavalier) were used. A static camera was set on the class using a wide angle, and one camera followed the teacher.

3.4 Statistical analysis

Statistical analysis was performed with IBM SPSS 26 and consisted of three steps. First, we examined the scoring distributions for all items with respect to observation mode. Second, mean differences as well as bivariate correlations across modes and teaching quality dimensions were estimated, which involved both lesson-level and classroom-level scores. Mean differences were investigated for statistical significance with paired *t* tests.

As reported above, raters were not distributed evenly across modes. In the case of rater effects, mean differences could occur across modes that are in fact due to raters using the scales differently. Adjusting for these effects is therefore crucial in the present study. To do so, we estimated separate mixed models for every teaching quality dimension with fixed effects for raters and observation mode, as well as random classroom effects.

Following Casabianca and colleagues (2013), we also looked at time trends in the scoring of teaching quality. Raters might change how they assign scores to lesson segments over time, and this could be a confounder when investigating mode effects. We estimated linear mixed models involving fixed effects for observation mode, time (months) and

the interaction between mode and time, as well as random effects for classrooms.

In addition, mode-specific *Generalizability* and *Decision studies* (G and D studies, Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991) are conducted to provide an in-depth analysis of the dependability of the corresponding scores. G Theory is an approach to decompose the observed variability in scores with respect to study conditions by performing analysis of variance (Brennan, 2001). The resulting variance decomposition provides insights on potential sources of measurement error, as it allows for distinguishing between wanted (e.g., differences in teaching quality between lessons or classrooms) and unwanted variability in scores (e.g., rater bias). A D study (Shavelson & Webb, 1991) is an exploratory simulation study based on the results of a G study. It aims at estimating how the present study conditions affect measurement error and reliability, similarly to the Spearman-Brown formula (Cronbach et al., 1972). In this study, we explore for each mode how teaching quality varies with respect to classrooms, lessons, segments, and raters. We estimate measurement error as well as reliabilities for live and video ratings and discuss how these could be improved under different study conditions.

We estimated random effects for classrooms, lessons, segments, and raters, as well as interactions between classrooms and raters. Regarding D studies, we investigated the potential to decrease measurement error by varying the number of observed lessons (2, 4) and segments (2, 4, 8). As only a negligible amount of variance was due to rater effects, we refrained from also conducting D studies that varied regarding the number of raters.

4 Results

4.1 Scoring distributions, time trends, and bivariate correlations

Table 1 provides the scoring distributions across modes as well as generic and subject-specific teaching quality dimensions.[2] We see that raters do not make use of the full rating scale in its breadth, but this appears to be similar across modes. For classroom management a ceiling-effect can be observed, as the lowest score is almost never applied, while the highest score is given most often. The highest score has been assigned even more frequently to video-recorded lesson segments, yielding percentage differences of 12–24% across modes for items assessing classroom management. For student support raters mostly apply lower scores in both observation modes, and the highest score is rarely used instead. When scoring cognitive activation, raters seem to make a wider use of the four-point-scale during video scoring, as the ratings are more evenly distributed than in the live scoring (percentage differences for the lowest score 1–17%). For most items assessing cognitive activation lower scores are obtained, too. We do, however, not identify a clear picture regarding mathematics educational structuring. Several items have a similar distribution between modes (e.g., correctness, co-construction), whereas for others the scores vary to a larger extent (e.g., structure, explanations).

4.1.1 Mean differences

Table 2 shows descriptive statistics and correlations across modes after item scores were aggregated to dimensions and then to the lesson level. The reported mean differences in Table 2 for classroom management (live vs. video: M_{diff} = -0.18, SE = 0.06, $t(14)$ = -2.92, p = .011, Cohen's d = -0.76) and cognitive activation (M_{diff} = 0.17, SE = 0.04, $t(14)$ = 3.97, p = .001, d = 1.02) are statistically significant, while those for student support (M_{diff} = 0.07, SE = 0.08, $t(14)$ = 0.83, p = .418, d = 0.22) as well as mathematics educational structuring (M_{diff} = -0.02, SE = 0.06, $t(14)$ = -0.36, p = .725, d = -0.09) are not. According to Cohen's classification (Cohen, 1992) the effects are moderate to large. Adjusting for rater differences across modes yields similar results (mode effect live vs. video for classroom management: M_{diff} = -0.11, SE = 0.04, p = .002, student support: M_{diff} = 0.09, SE = 0.05, p = .070, cognitive activation: M_{diff} = 0.16, SE = 0.05, p < .001, mathematics educational structuring: M_{diff} = 0.02, SE = 0.03, p = .411). This shows that for video-recorded lessons, higher scores in classroom management and lower scores in cognitive activation are assigned, which suggests mixed results regarding generic versus subject-specific dimensions of teaching quality.

4.1.2 Time trends

Figures 1 and 2 show time trends in scoring the four teaching quality dimensions across observation modes. We see that particularly the subject-specific dimensions evolve differently over time, with slightly lower scores in the third month for video-recorded lessons. For classroom management we observe more variation over time in the live ratings. However, after having adjusted the reported mean differences for time trends we obtain similar findings with an additional small effect for mathematics educational structuring (mode effect live vs. video for classroom management:

[2] Note that we do not employ statistical inference on the item level (i.e., regarding the raw scoring distributions). The reason for this is that scores are not used on the item level, but we aggregate them to form teaching quality dimensions. In addition, our previous research has shown that items should be considered as fixed rather than as random effects (Jentsch et al., 2020), which in G Theory are usually treated by averaging over them (e.g., Shavelson & Webb, 1991).

Table 2 Descriptives and bivariate correlations by dimension and mode, lesson-level. ote. Correlations were obtained from 1,000 Bootstrap samples (Efron & Tibshirani, 1993). CM = Classroom Management, SS = Student Support, CA = Cognitive Activation, MS = Mathematics Educational Structuring

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Live								
1. CM	--							
2. SS	−0.080	--						
3. CA	0.514	0.326	--					
4. MS	0.417	0.273	0.807	--				
Video								
5. CM	0.625	−0.429	0.268	0.162	--			
6. SS	−0.308	0.454	−0.133	0.025	−0.271	--		
7. CA	0.535	0.197	0.781	0.730	0.218	−0.100	--	
8. MS	0.276	0.186	0.522	0.725	0.127	0.371	0.627	--
M	3.67	1.68	2.68	3.18	3.85	1.61	2.51	3.19
SD	0.35	0.32	0.32	0.38	0.21	0.34	0.30	0.39
min	3.00	1.22	2.13	2.43	3.17	1.09	1.94	2.39
max	4.00	2.44	3.28	3.89	4.00	2.34	3.14	3.65

Note. Correlations were obtained from 1,000 Bootstrap samples (Efron & Tibshirani, 1993). CM = Classroom Management, SS = Student Support, CA = Cognitive Activation, MS = Mathematics Educational Structuring

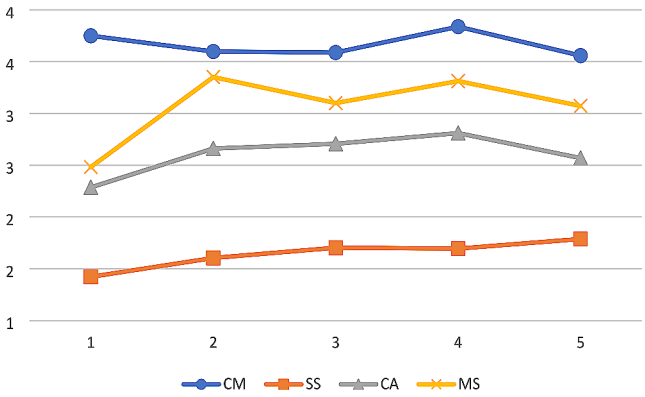

Fig. 1 Time trends for live observation by dimension. *Note*. The x-axis represents the month of the lesson scoring and the y-axis is the average score across lessons. CM = Classroom Management (top line), SS = Student Support (bottom line), CA = Cognitive Activation, MS = Mathematics Educational Structuring

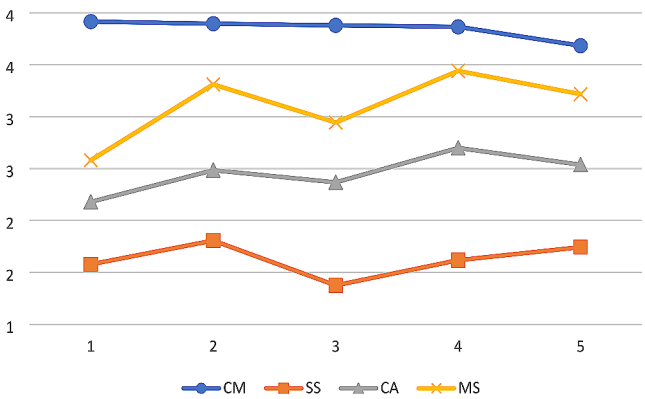

Fig. 2 Time trends for video observation by dimension. *Note*. The x-axis represents the month of the lesson scoring and the y-axis is the average score across lessons. CM = Classroom Management (top line), SS = Student Support (bottom line), CA = Cognitive Activation, MS = Mathematics Educational Structuring

M_{diff} = -0.28, SE = 0.04, p < .001, student support: M_{diff} = 0.07, SE = 0.06, p = .107, cognitive activation: M_{diff} = 0.25, SE = 0.04, p < .001, mathematics educational structuring: M_{diff} = 0.08, SE = 0.03, p = .015).

4.1.3 Correlations

Estimating bivariate correlations for teaching quality dimensions both across and within modes (see Table 2), we see that cognitive activation (r = .78) and mathematics educational structuring (r = .73) reach values across modes close to what is usually considered an acceptable reliability in the social sciences. The association between live and video scores in classroom management is slightly lower (r = .63), and for student support even more so (r = .45). However, as will be revealed in the D studies, these correlations are close to the estimated reliabilities for the corresponding teaching quality dimensions. This suggests that, live and video scoring results in similar lesson rankings after having adjusted for measurement error (i.e., disattenuated correlations are 0.83 for classroom management, 0.76 for student support, 1.00 for cognitive activation, and 0.85 for mathematics educational structuring).[3]

Table 2 also provides correlations within modes, which reveal further differences between live and video ratings. First, classroom management is associated with cognitive activation and mathematics educational structuring at a moderate or large effect sizes in the live ratings. At the

[3] Because we reported only the correlations estimated on the lesson level here, we re-calculated Table 2 on the classroom level, i.e. with corrected standard errors. The differences were negligible, which is probably due to the low amount of lesson variance within classrooms (see Table 3).

same time, there is no correlation to student support. For video-recorded lessons, the association between classroom management and any other dimension is of similar size. Second, student support and cognitive activation are moderately related in the live setting, but not at all in video observation. This suggests that teaching quality dimensions correlate differently across modes and need further investigation in future research.

4.2 Generalizability and decision studies

Table 3 provides the results of G studies with random effects for classrooms, lessons, segments, and raters. The variance decompositions explain more than 85% of the total variance in teaching quality dimensions, which shows that are large share of variability in scores is due to the investigated sources of variation. In accordance with the mixed models presented above, we find only a small amount of variance that is due to rater effects (0–7%, main and interaction effects summed up). Large mode differences occur for classroom management, such that variation between classrooms explains twice the amount of variance (live vs. video: 47% vs. 23%). In contrast, variation between lesson segments within classrooms explains less variance in classroom management in the live ratings than in the video ratings (37% vs. 59%). For student support, the variance decomposition yields very similar results across modes. Scoring the cognitive activation and mathematics educational structuring results in differences for between-lesson and within-lesson components across observation modes, with the first being larger in live observations (11% vs. 4%, 15% vs. 2%), and the latter yielding higher percentages for video recordings (32% vs. 51%, 18% vs. 22%).

4.2.1 D studies

Table 4 presents D studies for various research designs, including both lesson-based and classroom-based decisions. We see that for live ratings under the present study conditions (two lessons per classroom, four segments per lesson), classroom management, cognitive activation, and mathematics educational structuring reach reliabilities close to or larger than 0.80, which are usually considered acceptable. For video scoring and lesson-based decisions, the reliabilities are lower, and no acceptable results are obtained for classroom management. For student support we do not obtain acceptable reliabilities either, and at least four lessons (live ratings, classroom-based decisions) or eight segments (video ratings, lesson-based decisions) would be necessary to reach values above 0.70. Summing up, this suggests that with the conditions applied in the present study, live observations of classroom management and subject-specific characteristics of teaching quality yield good reliabilities for both classroom-based and lesson-based decisions. For the video ratings, this only holds true for cognitive activation and mathematics educational structuring (cognitive activation even questionable for lesson-based decisions).

5 Discussion

In the present study, we analyzed live and video ratings with a hybrid framework of teaching quality, involving both generic and subject-specific instructional practices (Schlesinger et al., 2018). We investigated how absolute (scores raters assign to classrooms or lessons) and relative decisions (rankings of classrooms and lessons) could be influenced by observation mode, as well as measurement

Table 3 Variance decomposition of scores by dimension and mode (percentages of total variability in parentheses)

Source	Live				Video			
	CM	SS	CA	MS	CM	SS	CA	MS
Classroom	0.107	0.033	0.069	0.111	0.025	0.044	0.051	0.146
	(46.8)	(17.1)	(45.8)	(54.9)	(23.4)	(16.7)	(35.5)	(69.6)
Lesson	0.002	0.006	0.016	0.031	0.000	0.018	0.006	0.005
	(0.8)	(3.4)	(10.9)	(15.4)	(0.0)	(6.7)	(4.1)	(2.3)
Segment	0.085	0.135	0.048	0.037	0.064	0.191	0.073	0.047
	(37.3)	(69.9)	(31.8)	(18.3)	(59.1)	(72.4)	(51.1)	(22.2)
Rater	0.003	0.000	0.000	0.000	0.001	0.000	0.001	0.003
	(1.3)	(0.0)	(0.2)	(0.0)	(0.8)	(0.0)	(0.6)	(1.3)
Classroom x Rater	0.000	0.005	0.005	0.014	0.005	0.003	0.002	0.002
	(0.0)	(2.8)	(3.3)	(6.7)	(4.7)	(1.0)	(1.6)	(1.0)
Residual	0.032	0.013	0.012	0.009	0.013	0.008	0.010	0.008
	(13.8)	(6.7)	(8.0)	(4.7)	(12.0)	(3.2)	(7.1)	(3.7)
Total	0.228	0.192	0.151	0.203	0.108	0.264	0.143	0.210
	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)

Note. CM = Classroom Management, SS = Student Support, CA = Cognitive Activation, MS = Mathematics Educational Structuring

Table 4 Decision studies by dimension and mode for various numbers of lessons and segments

	Live				Video			
	CM	SS	CA	MS	CM	SS	CA	MS
Classroom-based decisions								
Two lessons per classroom and four segments per lesson (original design)								
SEM	0.130	0.152	0.134	0.167	0.114	0.176	0.122	0.110
Reliability	0.863	0.591	0.791	0.798	0.667	0.589	0.773	0.926
Two lessons per classroom and eight segments per lesson								
SEM	0.098	0.123	0.121	0.159	0.088	0.151	0.098	0.091
Reliability	0.917	0.684	0.825	0.815	0.764	0.660	0.839	0.947
Four lessons per classroom and two segments per lesson								
SEM	0.126	0.155	0.118	0.141	0.114	0.187	0.114	0.100
Reliability	0.866	0.574	0.830	0.845	0.667	0.556	0.790	0.933
Four lessons per classroom and four segments per lesson								
SEM	0.095	0.118	0.105	0.130	0.089	0.134	0.089	0.084
Reliability	0.921	0.708	0.868	0.864	0.764	0.589	0.867	0.954
Lesson-based decisions								
Four segments per lesson (original design)								
SEM	0.164	0.195	0.126	0.130	0.145	0.224	0.145	0.122
Reliability	0.803	0.509	0.841	0.892	0.551	0.552	0.728	0.909
Eight segments per lesson								
SEM	0.118	0.143	0.097	0.106	0.109	0.160	0.106	0.094
Reliability	0.886	0.659	0.901	0.922	0.682	0.706	0.833	0.945

Note. CM = Classroom Management, SS = Student Support, CA = Cognitive Activation, MS = Mathematics Educational Structuring, SEM = Standard Error of Measurement

error and generalizability. Every lesson in our study was rated using both observation modes (i.e., live and video scoring).

Classroom management scored lower in live ratings, and video ratings resulted in unacceptable reliability on this dimension because of large segment variability within lessons. Increasing the number of observed segments per lesson could still improve the reliability. Cognitive activation scored higher in live ratings, but the results in terms of reliability were very similar across modes. For student support we did not find any mean differences between live and video ratings, and variation in reliabilities was negligible, too. We had similar results for mathematics educational structuring, with larger variation between lessons for live ratings. This is an unexpected result, as we assumed larger measurement error for student support and mathematics educational structuring in live scoring. The reason for this may be that raters should be able to assess the interactions between teachers and students more accurately during video scoring because of the teacher microphone (i.e., discussions can be heard and scored accordingly, which may not be the case for live ratings). Further research is necessary to shed light on how observation mode affects the assessment of scaffolding and supportive teaching practices.

Although some effect sizes are large, we should acknowledge that mean differences were presented in the original metric and therefore account for a quarter of a scale point at maximum (1 through 4), which is only slightly more than the estimated standard error (see Table 4). In a less homogenous sample, therefore, variability across modes would likely be less prominent. Therefore, we conclude that differences in mean values in classroom management could be due to differences in volume levels of teachers' and students' voices across modes. By listening to recordings from the teachers' microphones, raters might have difficulties to judge the volume level in the classroom (i.e., students' voices could be perceived quieter than they actually were), causing bias at that end. Consequently, raters could perceive disturbances as less problematic during video scoring. At the same time, video scoring of cognitive activation could be more difficult for the raters, as it might be unclear what students are working on, gestures and small movements may not be clearly visible in the video. During live observation, it is more likely that raters assess student discourse or problem-solving activities more accurately.

Overall, we found that live and video rating led to a very similar ranking of lessons and classrooms regarding teaching quality. Regarding the intercorrelations within modes (see Table 2), differences in how classroom management is associated with cognitive activation and mathematics educational structuring could be explained by measurement error. We found poor reliability for video ratings of classroom management, which leads to underestimating correlations with other variables. However, this phenomenon does not explain how classroom management is associated to student support, where the correlation is negative in the video

setting and virtually zero in the live setting. Regarding the associations between student support and cognitive activation, we believe tasks might be perceived as less cognitively activating if the level of student support is high. This indicates that students were less involved in higher-order thinking. This claim is supported by a recent study making use of the same data (Benecke & Kaiser, 2023), as teachers provide more content-related than just strategic help to students. Again, raters could rather perceive differences in student support during live scoring because of the teacher microphone.

The study by Casabianca et al. (2013) was the only one so far to explore differences with respect to observation mode in teaching quality, and the findings in terms of ranking classrooms and lessons were very similar to those in the present study. However, in contrast to Casabianca et al. (2013) we did not find that mean differences across modes could be explained by time trends regarding the scoring procedure. They remained statistically significant for classroom management and cognitive activation after adjusting for rater and time differences, even though the latter were marginal in this study. Future studies could explore further aspects of the study design that might influence observer ratings in the classroom (e.g., by exploring the dependability of different kinds of measures on observation mode and by comparing raters with varying amounts of experience or content knowledge).

5.1 Limitations

This study was set in a particular Western European context (i.e., mathematics classrooms in secondary schools in a German metropolitan area), which has probably shaped our view on teaching and learning accordingly. We acknowledge that our data stem from a convenience sample that is likely to represent a positive selection of German mathematics teachers, because they volunteered to participate in our study. Therefore, it might be worthwhile to replicate our results with a random sample of larger size.

Another limitation is that we did not explore the full potential of the video scoring in our study. Raters were asked not to stop the videos during scoring, and neither were they allowed to watch videos more than once. We took this decision to increase standardization across scoring procedures and to capitalize on the different types of information that are available to raters during live or video scoring, respectively. However, we understand that scholars often drop these restrictions when they use video scoring, and future research projects could take this procedure into account by employing a three-arm design (e.g., live vs. restricted video vs. unrestricted video scoring). In doing so, it would be possible to explore if being able to stop or rewind the video comes with additional benefits for reliability.

Finally, we could not assign raters randomly to classrooms or modes for pragmatic reasons (e.g., as in a random block design). This resulted in an uneven distribution of raters across modes, but statistical analysis adjusted for rater main effects. Our procedure has also the limitation of allowing raters to change their scores after the lesson has ended based on the manual of the observational instrument. This might result in bias if some observers change their scores more often than others. However, König (2015) argues that this approach can also lead to higher reliability if observers score more closely to the manual.

5.2 Conclusions

Mode effects are a potential danger to validity in studies using classroom observation, because they can affect the scoring procedure as well as the conclusions drawn from scores. In the present study, we compared live and video scoring of teaching quality in German secondary mathematics classrooms regarding absolute and relative decisions (i.e., scoring distributions and rankings of classrooms or lessons). Although relative decisions were only marginally affected, our findings suggest that the extent to which observation mode influences the precision of the scoring procedure is dependent on the teaching quality dimension, rather than the degree of subject-specificity: Given our hybrid framework, live scoring of classroom management and cognitive activation should be preferred over video scoring, particularly for lesson-based decisions. Vice versa, scoring teachers' cognitive and instructional support to students (i.e., mathematics educational structuring) benefits from videotaping lessons, which is likely due to better audio capture. Special attention should be paid to within-lesson variance in future studies, which may affect the validity of the conclusions drawn from scores if long-term decisions are made. We therefore recommend that both researchers and practitioners discuss carefully which conclusions they wish to draw from data, and choose frameworks, instruments as well as observation mode accordingly.

Acknowledgements The study was funded by the German Federal Ministry of Education and Research, grant numbers 01PK15006A and 01PK15006B. We thank Dr. Mark C. White for his helpful comments on an earlier version of the manuscript.

Funding Open access funding provided by University of Oslo (incl Oslo University Hospital)

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and sudent progress. *American Educational Research Journal*, *47*(1), 133–180. https://doi.org/10.3102/0002831209345157.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. Educational Assessment, 17(2–3), 62–87. https://doi.org/10.1080/10627197.2012.715014.

Benecke, K., & Kaiser, G. (2023). Teachers' approaches to handling student errors in mathematics classes. *Asian Journal for Mathematics Education*, *2*(2), 161–182. https://doi.org/10.1177/27527263231184642.

Bill and Melinda Gates Foundation (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf.

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*, 101600. https://doi.org/10.1016/j.learninstruc.2022.101600.

Blum, W., Vogel, S., Drüke-Noe, C., & Roppelt, A. (Eds.). (2015). *Bildungsstandards aktuell: Mathematik in Der Sekundarstufe II*. Schroedel.

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

Brunner, E. (2018). Qualität Von Mathematikunterricht: Eine Frage Der Perspektive. *Journal für Mathematik-Didaktik*, *39*, 257–284. https://doi.org/10.1007/s13138-017-0122-z.

Brunvard, S. (2010). Best practices for producing video content for teacher education. *Contemporary Issues in Technology and Teacher Education*, *10*, 247–256. https://doi.org/10.1007/978-1-4757-3456-0.

Casabianca, J. M., Mccaffrey, D., Gitomer, D., & Bell, 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 instructional quality in mathematics through different lenses: In search of Common Ground. *Zdm*, *50*, 535–553. https://doi.org/10.1007/s11858-018-0914-8.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155.

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.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior. Perspectives in social psychology*. Plenum.

Drollinger-Vetter, B. (2011). *Verstehenselemente Und Strukturelle Klarheit. Fachdidaktische Qualität Der Anleitung Von Mathematischen Verstehensprozessen Im Unterricht*. Waxmann.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Frederiksen, J. R., Sipusic, M., Gamoran, M., & Wolfe, E. W. (1992). *Video portfolio assessment: A study for the National Board for Professional Teaching standards*. Educational Testing Service.

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, evaluation und Verbesserung Des Unterrichts*. Klett-Kallmeyer.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*, 56–64. https://doi.org/10.3102/0013189X12437203.

Jaeger, R. M. (1993). Live vs. Memorex: Psychometric and practical issues in the collection of data on teachers' performances in the classroom. *Paper presented at the meeting of the American Educational Research Association, Atlanta, GA* (ERIC Document Reproduction Service No. ED360325). Retrieved from http://www.eric.ed.gov/PDFS/ED360325.pdf.

Jentsch, A., Casale, G., Schlesinger, L., Kaiser, G., König, J., & Blömeke, S. (2020). Variabilität und generalisierbarkeit von ratings zur Qualität Von Mathematikunterricht zwischen und innerhalb Von Unterrichtsstunden. *Unterrichtswissenschaft*, *48*(2), 179–197. https://doi.org/10.1007/s42010-019-00061-8.

Jentsch, A., Schlesinger, L., Heinrichs, H., Kaiser, G., König, J., & Blömeke, S. (2021). Erfassung Der Fachspezifischen Qualität Von Mathematikunterricht: Faktorenstruktur Und Zusammenhänge Zur Professionellen Kompetenz Von Mathematiklehrpersonen. *Journal für Mathematik-Didaktik*, *42*, 97–121. https://doi.org/10.1007/s13138-020-00168-x.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation.

Kleickmann, T., Steffensky, M., & Praetorius, A. K. (2020). Quality of teaching in science education: More than three basic dimensions? A.-K. Praetorius, J. Grünkorn, E. Klieme (Eds.), *Zeitschrift für Pädagogik. 66. Beiheft. Empirische Forschung zu Unterrichtsqualität: Theoretische Grundfragen und quantitative Modellierungen* (pp. 37–53) Beltz Juventa. https://doi.org/10.25656/01:25862.

Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung Und Fachdidaktik. Outcome-orientierte Messung Und Prozessqualität Des Unterrichts. *Zeitschrift für Pädagogik*, *54*, 222–237. https://doi.org/10.25656/01:4348.

Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit Von Mathematikunterricht: Theoretische Grundlagen und ausgewählte Ergebnisse Des Projekts Pythagoras. In M. Prenzel, & L. Allolio-Näcke (Eds.), *Untersuchungen Zur Bildungsqualität Von Schule. Abschlussbericht Des DFG-Schwerpunktprogramms* (pp. 127–146). Waxmann.

König, J. (2015). Measuring classroom management expertise (CME) of teachers: A video based assessment approach and statistical results. *Cogent Education*, *2*(1), 991178. https://doi.org/10.1080/2331186X.2014.991178.

Kounin, J. S. (1970). *Discipline and group management in classrooms*. Holt, Rinehart & Winston.

Learning Mathematics for Teaching Project. (2011). Measuring the Mathematical Quality of instruction. Learning Mathematics for

Teaching Project. *Journal of Mathematics Teacher Education*, *14*(1), 25–47. https://doi.org/10.1007/s10857-010-9140-1.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multi-level analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*, 399–418. https://doi.org/10.1111/j.1745-3984.2011.00152.x.

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

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement*, *74*(3), 400–422. https://doi.org/10.1177/0013164413515882.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist*, *59*(1), 14–19. https://doi.org/10.1037/0003-066X.59.1.14.

Nilsen, T., Scherer, R., & Blömeke, S. (2018). The relation of science teachers' quality and instruction to student motivation and achievement in the 4th and 8th grade: A nordic perspective. (Ed.), *Northern lights on TIMSS ansd PISA 2018* (pp. 61–94). Nordic Council of Ministers. The Nordic Council of Ministers.

Pianta, R. C., Paro, L., K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual K-3*. Paul H. Brookes Publishing Co.

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

Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three Basic dimensions. *Zdm*, *50*, 407–426. https://doi.org/10.1007/s11858-018-0918-4.

Ryan, A. M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T., & McCormick, S. (1995). Direct, indirect, and controlled observation and rating accuracy. *Journal of Applied Psychology*, *6*, 664–670. https://doi.org/10.1037/0021-9010.80.6.664.

Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *Zdm*, *50*, 475–491. https://doi.org/10.1007/s11858-018-0917-5.

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

Sherin, M., & Han, S. Y. (2004). Teacher learning in the context of a video club. *Teaching and Teacher Education*, *20*, 163–183. https://doi.org/10.1016/j.tate.2003.08.001.

Shuell, T. J. (1993). Toward an integrated theory of teaching and learning. *Educational Psychologist*, *28*, 291–311. https://doi.org/10.1207/s15326985ep2804_1.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS Videotape Classroom Study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States*. U.S. Department of Education, National Center for Education Statistics.

Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*, 271–296 (2010). https://doi.org/10.1007/s10648-010-9127-6.

Van Es, E. A., & Sherin, M. G. (2010). The influence of video clubs on teachers' thinking and practice. *Journal of Mathematics Teacher Education*, *13*, 155–176. https://doi.org/10.1007/s10857-009-9130-3.

White, M. C., & Klette, K. (2023). What's in a score? Problematizing interpretations of observation scores. *Studies in Educational Evaluation*. https://doi.org/10.1016/j.stueduc.2023.101238.