SCHOOL EFFECTIVENESS and SCHOOL IMPROVEMENT
An International Journal of Research, Policy and Practice

Routledge
Taylor & Francis Group

School Effectiveness and School Improvement
An International Journal of Research, Policy and Practice

ISSN: 0924-3453 (Print) 1744-5124 (Online) Journal homepage: www.tandfonline.com/journals/nses20

Primary school students' ratings of teaching – do they differentiate between subjects and teachers?

Svenja Rieser & Alexander Naumann

To cite this article: Svenja Rieser & Alexander Naumann (2024) Primary school students' ratings of teaching – do they differentiate between subjects and teachers?, School Effectiveness and School Improvement, 35:4, 486-505, DOI: 10.1080/09243453.2024.2396942

To link to this article: https://doi.org/10.1080/09243453.2024.2396942

Published online: 07 Oct 2024.

SCHOOL EFFECTIVENESS AND SCHOOL IMPROVEMENT
2024, VOL. 35, NO. 4, 486–505
https://doi.org/10.1080/09243453.2024.2396942

Routledge
Taylor & Francis Group

Primary school students' ratings of teaching – do they differentiate between subjects and teachers?

Svenja Rieser[a,b†] and Alexander Naumann[b,c†]

[a]School of Education, University of Wuppertal, Wuppertal, Germany; [b]Center for Research on Education and School Development (IFS), TU Dortmund University, Dortmund, Germany; [c]DIPF | Leibniz Institute for Research and Information in Education, Leibnitz, Germany

ABSTRACT

Our study aims to provide empirical evidence for and against the valid use of primary school students' ratings of three generic dimensions of teaching quality (classroom management, supportive climate, cognitive activation). We examine whether students discriminate between corresponding dimensions in different subjects, taking into account whether these subjects are taught by the same or different teachers. Using data from the German Trends in International Mathematics and Science Study 2015 assessment of 3,853 fourth graders ($M_{age} = 8.8$ years, 195 classes), we conduct multilevel-multigroup confirmatory factor analysis. The results suggest that students only differentiate between subjects when taught by different teachers. The findings also imply that valid inferences from student assessments may be affected by students' experiences with multiple teachers.

ARTICLE HISTORY
Received 22 December 2023
Accepted 22 August 2024

KEYWORDS
Student ratings of teaching; teaching quality; primary school; multilevel CFA

Introduction

When drawing inferences about teaching, student ratings have become a popular method. Relying on student ratings has many advantages. To name but a few, student ratings cost little and are easily collected. Moreover, students share far more experiences with their teachers than with any external observer (Goe et al., 2008; Wagner et al., 2016). However, some serious concerns regarding the validity of students' ratings on teaching have been voiced (e.g., Marder et al., 2021). One major concern addresses the discriminant validity of student ratings. It is questioned whether students, especially younger ones like primary school students, are able to discriminate between different constructs. Often a halo effect is expected (e.g., Benton & Cashin, 2011; Greenwald, 1997). In recent years, several studies have shown that this fear is unfounded, at least for university and college students (e.g., Benton & Cashin, 2011). However, evidence supporting the discriminant validity of younger students' ratings is still scarce.

CONTACT Svenja Rieser svenja.rieser@tu-dortmund.de
†These authors contributed equally to this publication.

Our study aims to add further evidence by investigating discriminant validity of primary school students' ratings of teaching quality. While previous research has focused on highlighting that students can differentiate between different constructs within the same subject, we base our study on a generic framework of teaching quality and investigate whether students are also able to distinguish between corresponding constructs in different subjects. Furthermore, since it can be assumed that the correlations between corresponding constructs in the different subjects depend on whether the subjects are taught by the same or by different teachers, our study will examine this aspect in more detail and provide information on the stability of primary school students' ratings of teaching quality within and between teachers.

Teaching quality

Our study is based on a generic framework of teaching quality by Klieme and colleagues (2009), which comprises three dimensions: cognitive activation, supportive climate, and classroom management. A similar model has also been proposed by Pianta and Hamre (2009). Several previous studies found these dimensions to be positively associated with students' learning outcomes (e.g., Fauth et al., 2014; Lipowsky et al., 2009). We choose this generic framework as it "refers to teaching activities, strategies, or routines that can be enacted across disciplines … . In other words, these practices are key to teaching, regardless of the subject matter under consideration" (Charalambous & Kyriakides, 2017, p. 425).

Cognitive activation comprises teaching practices that aim at fostering a deep understanding of the content to be learned. Cognitively activating teaching explores students' prior knowledge and builds on it. Teachers ask students to explain their ideas, confront them with contradictory facts, or apply their knowledge in new contexts (Lipowsky et al., 2009; Praetorius et al., 2018). Accordingly, a high level of cognitive activation should result in higher student achievement (e.g., Decristan et al., 2015; Fauth et al. 2019; Pianta et al., 2008).

Supportive climate builds on self-determination theory (Deci & Ryan, 1985) and describes the emotional quality of the teacher–student relationship. Teachers can create a supportive classroom climate by providing individual support to students and showing respect for their ideas and needs. Previous research has shown that a supportive classroom climate is positively associated with motivational student outcomes (e.g., Fauth et al., 2019).

Effective classroom management maximises students' time on task by minimising disruption (e.g., Doyle, 2006). Essential elements of effective classroom management include rules and routines that are consistently applied in the classroom, well-structured lessons, and good classroom organization (Praetorius et al., 2018). Research has repeatedly shown that effective classroom management is positively associated with students' academic achievement (e.g., Fauth et al., 2019; Lipowsky et al., 2009).

Previous studies have indicated that secondary school students are able to discriminate between the three dimensions of teaching quality in their ratings within a subject (e.g., Kunter & Baumert, 2006; Wagner et al., 2016, Wallace et al., 2016). However, as our study explores the discriminant validity of younger students' ratings, we will focus on summarizing findings from research in primary school.

Discriminant validity of primary school students' ratings of teaching quality

Overall, there is relatively little research on the discriminant validity of primary school students' ratings of the three dimensions of teaching quality. In 2000, Peterson and

colleagues analysed data from 401 classrooms from kindergarten (57 classes), primary (89 classes), and middle-/high school (255 classes). Using exploratory factor analysis, they found that independent of the students' age cohort, student ratings differentiated between constructs that referred to students' opportunity to learn (cognitive activation) and such constructs that characterized a caring and respectful teacher–student relationship (supportive climate).

In a more recent study, van der Scheer and colleagues (2019) gathered data from 31 Dutch primary school teachers and 675 students (83% fourth-grade students) to assess the construct validity of students' ratings of teaching quality. The student questionnaire contained five scales: classroom climate (supportive climate), goal orientation, clear instruction, challenging students (cognitive activation), and classroom management. Using confirmatory factor analysis (CFA), they found that a five-factor model represented the data adequately. As correlations between the factors were low, a unidimensional model was not plausible. Although the multilevel structure of the data was not considered in these two studies, the results indicate that even primary school students might be able to differentiate dimensions of teaching quality within the same subject.

One of the first studies which considered the multilevel structure of primary school students' ratings of teaching quality was conducted by Fauth and colleagues (2014). They analysed data from 1,556 third-grade students in 89 classrooms. Students rated teaching quality following the three-dimensional generic framework described earlier. When comparing different multilevel CFA models (MCFA), the researchers found that a model with three factors at the individual and classroom level fitted the data best. They interpreted these findings as evidence supporting the discriminant validity of primary student evaluations of teaching quality. Stahns and colleagues (2020) provided further support for this interpretation. Analysing data from the German PIRLS 2016 sample ($N = 3{,}797$ students from 192 classrooms) on the three dimensions of teaching quality in German language lessons, they also found that a MCFA model with three factors on both levels fitted the data best.

In summary, there is reason to assume that students from primary school age onwards can differentiate between the dimensions of teaching quality. However, previous studies have only evaluated the discriminant validity of students' ratings of different constructs within the same subject. In the next section we will focus on previous research investigating students' ratings of corresponding constructs in different subjects.

Studies on students' ratings across subjects

While there are some studies showing that even primary school students differentiate between different constructs when assessing teaching quality, there is little research on students' ability to differentiate between the same construct in different subjects. Our review of the existing literature resulted in only three studies, all of which worked with samples from secondary school.

In 2000, Gruehn asked 3,787 German seventh graders from 117 schools to rate characteristics of their mathematics, biology, and physics classes. The items were worded identically for all three subjects. Most of the constructs assessed can be assigned to one of the three dimensions that form the basis of the present study. As part of her research, she tested whether students' ratings could differentiate between the same teaching

characteristic in different subjects using CFA models. For each teaching characteristic (e.g., occurrence of disruptions), she specified three separate models: (a) a model with all items assessing the same teaching characteristic loading on one general factor, (b) a model with three subject-specific factors, and (c) a model with three subject factors and an additional general factor. For most of the 19 teaching characteristics, the three-factor model without a general factor fitted the data best. The correlations between these factors were generally low or medium. Gruehn's results indicate that secondary school students can distinguish between corresponding constructs in different subjects. However, her analysis only focused on the student level. If teaching quality is considered a classroom-level construct (Lüdtke et al., 2009), further studies considering the multilevel structure of students' ratings are needed.

Wagner and colleagues (2013) explored the factorial structure of 6,909 secondary school students' ratings of teaching quality from 280 classes in an MCFA. The students rated their German and their English as a foreign language (EFL) classes, which were taught by different teachers. They found that a model comprising five factors (Motivation, Understandableness, Student Involvement, Structure, and Classroom Management) in each subject on both levels fitted the data best.

Across subjects, Wagner and colleagues (2013) found only low correlations at the student level. At the classroom level, only the latent factors for structure in German and EFL correlated significantly. Thus, secondary school students who are taught in both subjects by different teachers seem to differentiate between corresponding constructs in different subjects.

Praetorius and colleagues (2016) compared students' ratings from 548 ninth graders who were taught by the same teacher in German and EFL. They focused on items that tapped different aspects of classroom management and supportive climate (assessed as motivational support). Using generalizability theory, they showed that most of the variance in the ratings of classroom management (72%) was due to teacher differences while only 7% could be assigned to subject-specific differences. For supportive climate, 37% of the variance was attributable to differences between teachers while 19% was due to subject-specific differences. These results suggest that classroom management is judged consistently across subjects when pupils are taught by the same teacher. Supportive climate, on the other hand, seems to be more variable across subjects.

Overall, these studies provide evidence supporting the assumption that secondary school students can differentiate between corresponding constructs in different subjects. However, the connection between students' ratings of teaching quality across subjects depends on two factors: the construct being considered and whether the class is taught by the same or different teachers. The following paragraphs will elaborate on these points.

Classroom management

Praetorius and colleagues (2016) found that if the same teacher teaches different subjects, classroom management does not vary substantially across subjects. This result is plausible because common operationalizations of classroom management do not include subject-specific characteristics. Research on teacher competence has shown that the quality of classroom management depends on teachers' general pedagogical or psychological knowledge (e.g., Lenske et al., 2016) rather than on subject-specific competencies.

Therefore, it is plausible that students' ratings of classroom management do not vary between subjects if both are taught by the same teacher. However, they can be assumed to differ across subjects if taught by different teachers (Praetorius et al., 2016; Wagner et al., 2013).

Supportive climate

Regarding supportive climate, students' ratings remained relatively stable across subjects when taught by the same teacher, but were more variable than ratings of classroom management. Praetorius and colleagues (2016) used items that focused on fostering students' enthusiasm for a topic and explaining content clearly. Teachers' ability to do so may vary between subjects, depending on their own enthusiasm for a subject or their pedagogical content knowledge (e.g., Kunter et al., 2008). When taught by different teachers, supportive climate can be expected to vary between subjects (e.g., Gruehn, 2000; Wagner et al., 2013). Conceptually, supportive climate comprises individualized support by the teacher but also the emotional quality of the teacher–student relationship (Klieme et al., 2009). No empirical work has been found that focuses on this aspect, but we assume that the emotional relationship between teacher and student is not dependent on the subject that is being taught.

Cognitive activation

It is hard to make assumptions about the relationship between students' ratings of cognitive activation in two different subjects, as only Gruehn's (2000) study assessed any teaching characteristics that could be subsumed under this construct. Her findings suggest that students differentiate cognitive activation in different subjects. However, Gruehn did not consider whether teachers taught only one or multiple subjects, so that teacher and subject are most likely confounded.

However, research on teacher competencies allows for inferences to be made regarding the connection between students' ratings of cognitive activation across subjects. Fauth and colleagues (2019) found a positive correlation between the degree of cognitive activation in classrooms and teachers' pedagogical content knowledge (see also Baumert et al., 2010). As pedagogical content knowledge can vary even within one teacher for different subjects, it is reasonable to expect that students' ratings of cognitive activation will differ between subjects, even if taught by the same teacher.

The inferences drawn above are based on studies from secondary schools. Thus, it remains unclear whether the same conclusions hold for primary school students' ratings. Consequently, we address the following research question in our study: Do primary school students differentiate in their ratings between different subjects and different teachers?

Even though there is only little research addressing this question in secondary school, the results from these studies imply that the connections across subjects may vary, depending on the dimension of teaching quality that is considered and whether the class is taught the subjects by the same or different teachers. Therefore, we integrate this information in our research and formulate the following hypothesis:

(1) For classroom management, we expect the connection across subjects to differ, depending on whether the class is taught these subjects by the same or different

teachers. If taught by the same teacher, we expect to find a stronger connection between students' ratings of classroom management in different subjects than if they are taught by different teachers.

(2) For supportive climate, we expect the connection across subjects to differ, depending on whether the class is taught these subjects by the same or different teachers. If taught by the same teacher, we expect stronger connections between students' ratings of supportive climate across subjects than if they are taught by different teachers.

(3) For cognitive activation, we expect similar connections across subjects for classes taught by the same teacher and those taught by different teachers.

Teaching quality is often conceptualized and analysed as a classroom characteristic in terms of a so-called "shared" or "reflective" (Stapleton et al., 2016) construct. Following this conceptualization, we will focus on analysing our data at the classroom level.

Method

Sample

For our analyses, we use data from the German Trends in International Mathematics and Science Study (TIMSS) 2015 sample (Wendt et al., 2018). The sample includes 3,942 fourth graders from 213 classrooms. One class per school participated. The average age of the students was 10.35 years ($SD = 0.51$), and 48.1% were female; 34.6% of them sometimes spoke a language other than German at home, indicating a migrant background.

On average, classes comprised 20.77 students ($SD = 4.02$), of whom 18.54 students ($SD = 5.43$) participated in the assessments. However, in 18 classes the number of participating students was less than 10. To improve the reliability of our analysis at the classroom level (Bliese, 2000) by reducing sampling error, we decided to exclude these classes from our sample. Thus, our final sample consisted of 195 classrooms and 3,853 students: 117 classes (62.2%) were taught mathematics and science by the same teacher, while 71 classes (37.8%) had different teachers in both subjects. For seven classes, the corresponding information was not available. Details on how this information was deduced can be found in the next section.

Measures

The TIMSS 2015 student questionnaires contained items assessing cognitive activation, supportive climate, and classroom management in mathematics and science classes according to Klieme and colleagues' (2009) framework.

Cognitive activation and supportive climate items were adapted from Fauth and colleagues (2014), and answered on a 4-point Likert Scale (1 = *I totally disagree* to 4 = *I fully agree*). Originally, cognitive activation was assessed by seven items (e.g., "Our mathematics teacher wants me to be able to explain my answers"). However, using MCFA, Bellens and colleagues (2019) found that these items do not form a common factor. By conducting an additional multilevel exploratory factor analysis, they were able to show that a model comprising only three items fits the data well enough to be used in further analysis (for more details, see Bellens et al., 2019). Building on these results, we

chose to use the same three items focusing on teacher behaviour in our analysis. Five items focusing on a positive teacher–student relationship (e.g., "Our mathematics teacher believes that I can solve difficult tasks") assessed supportive climate. Classroom management was assessed by five items adapted from Baumert and colleagues (2009). They asked about the frequency of disruptions in class (e.g., "The students do not listen to the teacher") and were answered on a 4-point Likert scale (1 = *in every lesson* to 4 = *never*). The wording of the items was identical in both subjects, differing only in that they referred to either mathematics or science teaching. All items were coded so that higher scores indicated higher quality teaching. Table 1 shows descriptive statistics for all items. For the items referring to mathematics, Bellens and collegues identified the three dimensions in a multilevel CFA model. For the science items, we performed a similar analysis using maximum likelihood estimation and found that a model with three factors at the individual and classroom level fitted the data (root-mean-square error of approximation [RMSEA] = .045; comparative fit index [CFI] = .935; Tucker–Lewis index [TLI] = .919; standardized root-mean-square residual $[SRMR]_{(within)} = .037$; $SRMR_{(between)} = .079$; $\chi^2 = 908.857^*$, $df = 125$, $p = .000$).

To assess whether a class was taught by the same or different teachers in mathematics and science, a variable called TEACHER was calculated. The calculation was based on the TIMSS variable IDSUBJ, which indicates which subjects each teacher in the sample teaches

Table 1. Descriptive statistics for the teaching quality items.

	Construct	Item	*n*	*M*	*SD*	ICC1	ICC2	ω_{Total}
Mathematics	Classroom management	cmm1	3,068	2.72	0.93	.10	.66	
		cmm2	3,018	2.51	0.98	.18	.79	
		cmm3	3,059	2.43	1.05	.20	.82	
		cmm4	3,030	3.08	0.96	.09	.63	
		cmm5	3,065	2.98	1.10	.13	.72	
		CMM	3,076	2.74	0.76	.22	.83	.81
	Supportive climate	scm1	3,059	3.54	0.79	.06	.51	
		scm2	2,978	3.18	0.88	.08	.60	
		scm3	3,040	3.58	0.73	.06	.53	
		scm4	2,949	3.33	0.80	.07	.56	
		scm5	2,973	3.17	0.90	.04	.42	
		SCM	3,073	3.36	0.59	.10	.66	.76
	Cognitive activation	cam1	3,012	3.34	0.87	.05	.48	
		cam	3,020	3.29	0.89	.06	.51	
		cam	3,024	3.32	0.80	.03	.38	
		CAM	3,099	3.32	0.64	.06	.54	.60
Science	Classroom management	cms1	3,029	2.88	0.93	.16	.77	
		cms2	2,937	2.65	1.00	.21	.81	
		cms3	3,022	2.61	1.04	.21	.82	
		cms4	2,998	3.08	0.89	.11	.67	
		cms5	3,020	3.00	1.09	.14	.74	
		CMS	3,058	2.84	0.83	.23	.84	.88
	Supportive climate	scs1	3,039	3.52	0.80	.05	.46	
		scs2	2,953	3.18	0.88	.07	.56	
		scs3	3,017	3.52	0.75	.06	.50	
		scs4	2,915	3.24	0.85	.07	.55	
		scs5	2,959	3.20	0.87	.03	.36	
		SCS	3,051	3.33	0.62	.08	.60	.81
	Cognitive activation	cas1	2,991	3.33	0.87	.04	.43	
		cas2	2,963	3.39	0.83	.03	.33	
		cas3	2,961	3.32	0.82	.03	.38	
		CAS	3,050	3.35	0.66	.04	.44	.68

Note: The precise wording for all items is presented in Appendix 1. ICC = intraclass correlation.

to the participating class. By combining this information for all teachers, it was possible to infer whether a class was taught in both subjects by the same teacher (TEACHER = 0) or by different teachers (TEACHER = 1).

Data analysis

To investigate our hypotheses, we applied multigroup multilevel confirmatory factor analysis (MG-MCFA). In all MG-MCFA models, we specified the factorial structure at the classroom level with each item loading on one of the three dimensions of teaching quality following the aforementioned findings of Bellens and colleagues (2019). At the student level, however, we specified a saturated model as none of our hypotheses targeted student-level relationships.

As our main hypotheses target the classroom-level between-subject correlations of each teaching quality dimension, we first checked each teaching quality dimension separately for measurement invariance (MI; e.g., Bollen, 1989; Little, 1997). MI depends on whether the model parameters are invariant between groups, in our case classes taught by the same or different teachers in both subjects. Cheung and Rensvold (2002) distinguish four levels of MI, with each higher level inheriting the features of the previous level: (1) configural invariance, where the same number of factors is associated with the same set of items in each group; (2) metric invariance, which additionally requires factor loadings to be equal across groups; (3) scalar invariance, meaning that item intercepts are also constant across groups; and finally (4) full invariance, which holds if the item-specific residuals are equal across groups. The comparison of correlations across groups requires at least metric MI. Metric MI ensures that a construct has been measured on the same scale, thus allowing for the comparison of covariances across groups (Little, 1997).

For the analysis of MI, we first divided our sample into four groups depending on the subject and the TEACHER variable: (1) mathematics & same teacher, (2) science & same teacher, (3) mathematics & different teacher, and (4) science & different teacher (see Figure 1).

We then checked for classroom-level MI across these four groups in a MG-MCFA model. In a first step (a) we tested for MI across subjects within each of the two TEACHER groups, and in a second step (b) across the two TEACHER groups given the previously detected

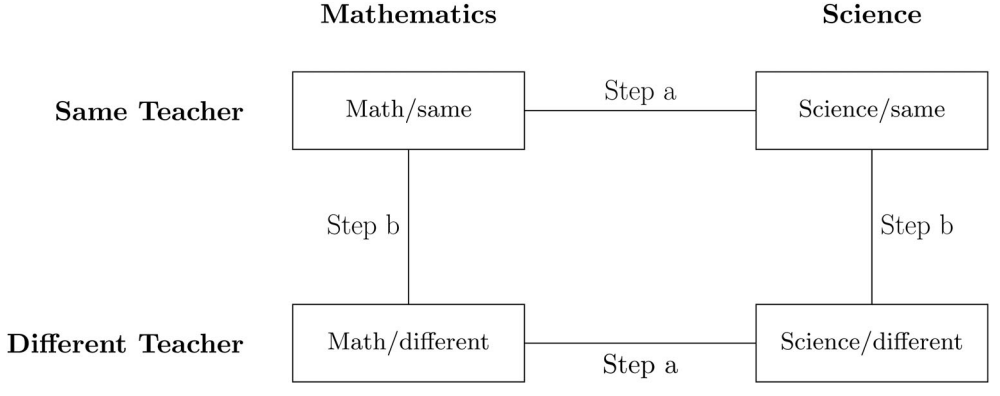

Figure 1. Procedure for the assessment of measurement invariance across subject and TEACHER.

level of MI across subjects (see Figure 1). The level of MI was determined via model comparisons, choosing the most restrictive model that did not have a meaningfully worse fit to the data than the previous, less restrictive model (Millsap, 2011).

Having secured at least metric MI across the TEACHER groups, we addressed our main hypotheses. For comparing the classroom-level between-subject correlations of each teaching quality dimension across the two TEACHER groups, we also applied MG-MCFA models. Yet this time, the multigroup analyses targeted the two TEACHER groups only; that is, we did not treat the subjects as groups in the multigroup part of the MG-MCFA models. Instead, each TEACHER groups' measurement model now accounted for the fact that teaching quality was measured in two subjects. Hence, the MG-MCFA models now included two classroom-level factors corresponding to one dimension of teaching quality in each of the two subjects (Figure 2). If a model indicated high classroom-level factor correlations, we additionally tested whether a unidimensional model fit the data better, thus testing whether students still differentiate between the subjects.

Correlation coefficients usually do not follow a normal distribution. Thus, to test the hypotheses of whether the correlations of the teaching quality dimensions differ between the two teacher groups, we first perform a Fisher's *z* transformation for each correlation parameter (Fisher, 1915).[1] Fisher's *z* transformation performs an asymptotic normalization, whereby the distribution of the correlation coefficients is approximately converted into a normal distribution. We then use the 95% Bayesian credible intervals (BCI) to test whether the difference between the Fisher *z*-transformed correlations is significantly different from zero. If the 95% BCI of the difference does not comprise zero, we conclude that correlations differ between the two teacher groups.

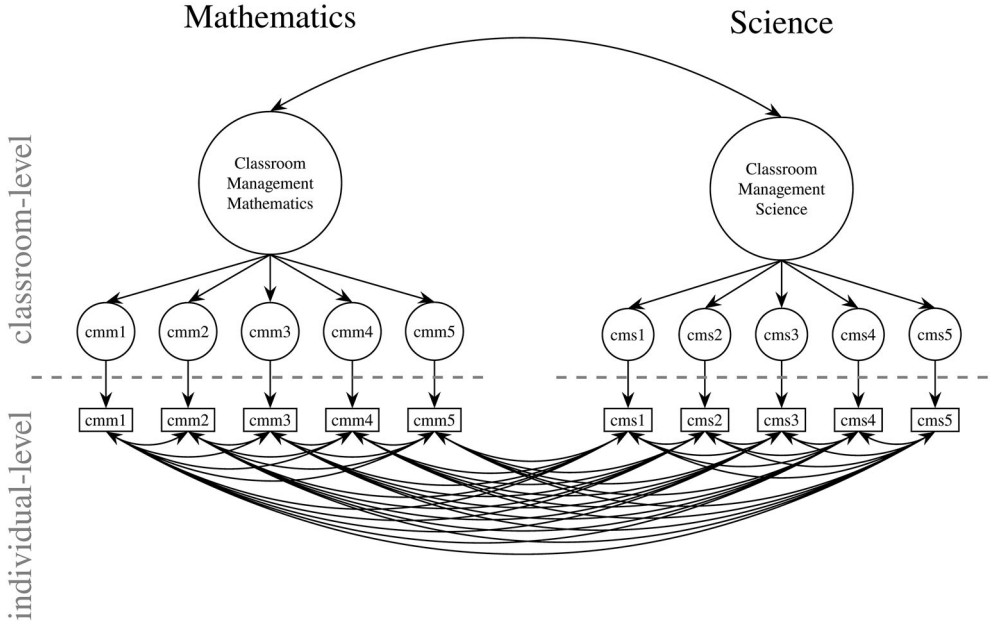

Figure 2. Exemplary multilevel factor model for classroom management.

Estimation was carried out using Hamiltonian Monte Carlo in Stan (Stan Development Team, 2023) within a Bayesian framework (e.g., Gelman et al., 2013). To deal with missing data, we applied a Bayesian full information approach. As we had no reliable prior information available, in all models, we assumed weakly informative priors only following the principles by Gelman and colleagues (2017). More precisely, we specified Lewandowski–Kurowicka–Joe (LKJ) distributions (Lewandowski et al., 2009) as priors for the Cholesky factors of correlation matrices with shape parameter $\eta = 1$ and vague half-Cauchy priors with location = 0 and scale = 2.5 for variables' standard deviation parameters, as well as wide normal distributions with mean $\mu = 1$ and standard deviation $\sigma = 1$ as priors for intercept and loading parameters, respectively. We ran four Markov chains with 10,000 iterations each per model, using the initial 5,000 iterations as adaptation phase and burn-in. Convergence was checked via visual inspection of trace plots and the Gelman–Rubin R statistic (Gelman et al., 2013). Analysis results were only interpreted if Gelman–Rubin R was below 1.05 and the effective sample size was greater than 1,000 per parameter.

Absolute model fit was checked using posterior predictive *p* values (PPP), based on the discrepancy in the posterior log-likelihood of the actual and replicated data. Accordingly, PPP values of .50 indicate ideal model fit, with model fit decreasing as the PPP deviates from .50 (Meng, 1994). PPP values below .10 indicate poor model fit and can therefore be considered unacceptable, as can values above .50, which may indicate overfitting (e.g., Muthén & Asparouhov, 2012).

For model comparison, we employed leave-one-out cross-validation (LOO) using the LOO package (Vehtari et al., 2022). Cross-validation allows selecting the model with highest predictive performance from a set of models based on the Bayesian LOO estimate of the expected log pointwise predictive density (elpd_loo). Model differences in elpd_loo (i.e., elpd_diff) values less than 4 can be seen as small, while values greater than 4 are substantial. In the results section, we report elpd_diff together with the LOO information criterion (LOOIC), which corresponds to -2*elpd_loo and whose interpretation is similar to other information criteria; that is, lower values are better.

Results

Descriptives

The descriptive data for all items and scales that were used in the analyses are presented in Table 1. At the student level, the students reported to experience a rather high degree of cognitive activation ($M_{maths} = 3.32$; $SD_{maths} = 0.64$; $M_{science} = 3.35$; $SD_{science} = 0.66$) and supportive climate ($M_{maths} = 3.36$; $SD_{maths} = 0.59$; $M_{science} = 3.33$; $SD_{science} = 0.62$) in both subjects. Classroom management was rated less positively in both subjects ($M_{maths} = 2.74$; $SD_{maths} = 0.76$; $M_{science} = 2.84$; $SD_{science} = 0.83$). McDonalds Omega indicates at least acceptable internal consistency for all scales (Table 1: $\omega \geq .60$).

For cognitive activation and supportive climate, only little variance was detected at the classroom level (intraclass correlation [ICC]1 between .03 and .08), resulting in low ICC2 values ranging from .38 to .60. Classroom management showed higher ICC1 (from .09 to .21) and ICC2 (from .63 to .82) values.

Measurement invariance

Measurement invariance across subjects

Before testing our main hypotheses, we had to check whether the assumption of at least metric MI across groups in our sample holds. For all tested models, PPP values ranged between PPP = .18 and PPP = .33, thus indicating acceptable but not perfect fit. However, model comparison via LOOIC and elpd_diff revealed more pronounced differences in model performance. Table 2 shows the results of these analyses.

For classroom management, the model assuming metric MI provided the lowest LOOIC value (LOOIC = 71,261.70). Only the configural model offered similar predictive performance following elpd_diff (elpd_diff = −3.45), while both the scalar (elpd_diff = −26.93) and the full MI model (elpd_diff = −25.52) performed worse. As the metric model is more parsimonious than the configural model, for classroom management, we thus assumed metric invariance across subjects.

At a first glance, the metric model also showed the best predictive performance for supportive climate (LOOIC = 62,525.27). However, the LOOIC and the elpd_diff for the full model only deviated marginally from the metric model (LOOIC = 62,525.77, elp_diff = −0.25). Consequently, as the full model is more parsimonious than the metric model, we assumed full MI for supportive climate across subjects (see Table 2).

The results for the MI analyses across subjects for cognitive activation were similar to those for classroom management. The LOOIC was smallest for the metric (LOOIC = 40,956.24) and the configural model (LOOIC = 40,956.24). However, the eldp_loo indicated the best performance for the metric model, which is also more parsimonious than the configural model. Thus, we assumed metric invariance for cognitive activation across subjects (see Table 2).

Measurement invariance across TEACHER

Next, we checked for MI across the TEACHER groups (same & different). Again, all PPP values were larger than the cut-off value of PPP = .10 but smaller than PPP < .50, indicating at least acceptable model fit for all dimensions of teaching quality and all levels of MI. The differences between the models in terms of predictive performance according to LOOIC and elpd_diff were again more pronounced (see Table 3).

Table 2. Measurement invariance across subjects.

	MI type	PPP	LOOIC	elpd_diff (*SE*)
Classroom management	Configural	.33	71,268.61	−3.45 (2.45)
	Metric	**.25**	**71,261.70**	**0**
	Scalar	.26	71,315.56	−26.93 (9.63)
	Full	.23	71,312.74	−25.52 (9.69)
Supportive climate	Configural	.23	62,530.54	−2.63 (0.97)
	Metric	.20	62,525.27	0
	Scalar	.19	62,527.74	−1.23 (5.98)
	Full	**.18**	**62,525.77**	**−0.25 (6.11)**
Cognitive activation	Configural	.31	40,956.24	−0.85 (0.43)
	Metric	**.28**	**40,956.24**	**0**
	Scalar	.27	40,979.35	−11.55 (6.18)
	Full	.25	40,977.14	−10.45 (6.58)

Note: No restrictions were introduced for the TEACHER groups. The type of invariance that can be assumed is marked by bold type. MI = measurement invariance; PPP = posterior predictive *p* values; LOOIC = leave-one-out cross-validation information criterion; elpd_diff = difference in expected log pointwise predictive density.

Table 3. Measurement Invariance across TEACHER groups.

	MI type	PPP	LOOIC	elpd_diff (*SE*)
Classroom management	Configural	.25	71,261.70	−12.67 (4.80)
	Metric	.22	71,257.14	−10.38 (4.79)
	Scalar	.22	71,236.37	0
	Full	**.20**	**71,236.65**	**−0.14 (1.34)**
Supportive climate	Configural	.18	62,526.47	0
	Metric	**.15**	**62,528.32**	**−1.00 (2.7)**
	Scalar	.16	62,556.59	−15.1 (8.00)
	Full	.15	62,556.60	−15.1 (8.00)
Cognitive activation	Configural	.28	40,956.24	0
	Metric	**.26**	**40,956.48**	**−0.12 (0.44)**
	Scalar	.26	40,974.39	−9.08 (5.91)
	Full	.24	40,972.89	−8.33 (6.01)

Note: For cognitive activation and classroom management, MI between subjects is assumed to be metric, for supportive climate, full invariance between subject is assumed. The type of invariance that can be assumed between TEACHER groups is marked by bold type. MI = measurement invariance; PPP = posterior predictive *p* values; LOOIC = leave-one-out cross-validation information criterion; elpd_diff = difference in expected log pointwise predictive density.

For classroom management, LOOIC and elpd_diff favour the scalar invariance model (LOOIC = 71,236.37; elpd_diff = 0). However, predictive performance of the more parsimonious full MI model is not substantially worse (LOOIC = 71,236.65; elpd_diff = −0.14). Therefore, we assume full MI across TEACHER groups for classroom management (see Table 3).

For supportive climate, we assume metric invariance. Although LOOIC and elpd_diff favour configural invariance (LOOIC = 62,526.47; elpd_diff = 0), these indices only deviate slightly from those of the metric model (LOOIC = 62,528.32; elpd_diff = −1.00) so that we chose the more parsimonious metric model (see Table 3).

The same level of invariance across TEACHER groups holds for cognitive activation. Even though the configural model shows the best predictive performance (LOOIC = 40,956.24; elpd_diff = 0), the more parsimonious metric model performs equally well (LOOIC = 40,956.48; elpd_diff = −0.12; see Table 3).

Correlations across subjects

After at least metric invariance was established between the TEACHER groups for each dimension of teaching quality, we proceeded with the analysis for our main research question. We assessed the six correlations across subjects, one for each dimension of teaching quality in each TEACHER group using MCFA models (see Figure 2 for an example).

Within the different-teachers group, the correlation across subjects for classroom management was high and statistically meaningful (r_{CM} = .75 [0.59, 0.85], Table 4). However, compared to a unidimensional model, elpd_diff favoured the original two-dimensional model with one classroom management factor in each subject (elpd_diff = −4.1). Thus, although the correlation was high and statistically meaningful, students of the different-teachers group differentiated between classroom management in science and mathematics (see Table 5). For cognitive activation and supportive climate, the correlations across subjects were rather low and not statistically meaningful (r_{CA} = .28 [−0.33, 0.75]; r_{SC} = .14 [−0.23, 0.47]; see Table 4).

Within the same-teacher group, we found very high and statistically meaningful correlations for each teaching quality dimension across subjects (r_{CA} = .94 [0.80, 1.00], r_{SC} = .97

Table 4. Comparison of the correlations across subjects for students taught either by the same or different teachers in both subjects.

	Correlations across subjects		Comparison of the *z*-transformed correlations	
	Same teacher	Different teachers	Difference of the *z*-transformed correlations	Credible intervals
Classroom Management	1.00* [0.99, 1.00]	.75* [0.59, 0.85]	2.38*	[0.64, 3.94]
Supportive Climate	.97* [0.92, 1.00]	.14 [−0.23, 0.47]	2.22*	[0.58, 3.64]
Cognitive Activation	.94* [0.80, 1.00]	.28 [−0.33, 0.75]	1.63*	[0.54, 3.19]

*Statistically meaningful with corresponding Bayesian credible intervals (BCI) not comprising zero.

Table 5. Comparing two-factor models (one dimension for each subject) to one-factor models (one dimension across both subjects) in both TEACHER groups.

		Different teachers			Same teacher		
		LOOIC 2dim	LOOIC 1dim	elpd_diff (*SE*)	LOOIC 2dim	LOOIC 1dim	elpd_diff (*SE*)
Science	Classroom management	24,883.6	24,891.7	−4.1 (3.5)	35,862.5	35,862.0	−0.2 (0.4)
	Supportive climate	Not tested			24,309.4	24,309.3	−0.1 (0.3)
	Cognitive activation	Not tested			42,271.7	42,272.6	−0.5 (0.4)

Note: LOOIC = leave-one-out cross-validation information criterion; elpd_diff = difference in expected log pointwise predictive density.

[0.92, 1.00], $r_{CM} = 1.00$ [0.99, 1.00]; see Table 4). Corresponding model comparisons revealed a comparable fit to the unidimensional models. As the unidimensional models are more parsimonious, we assume that students within the same-teacher group did not differentiate between the two subjects in their ratings of the three teaching quality dimensions (see Table 5).

The comparison of the respective correlations between the same- and different-teachers groups, using the Fisher *z*-transformed correlations, revealed meaningful differences for the correlation of each dimension between the TEACHER groups (classroom management: $\Delta z = 2.38$ [0.64, 3.94]; supportive climate: $\Delta z = 2.22$ [0.58, 3.64]; cognitive activation: $\Delta z = 1.63$ [0.54, 3.19]).

Discussion

The aim of our study was adding to the literature on teaching quality by investigating discriminant validity of primary school students' ratings. While previous research focused on students' ability of differentiating between relevant dimensions of teaching quality within one subject, we investigated whether students are also capable of distinguishing corresponding dimensions of teaching quality across different subjects, thereby considering whether these subjects were taught by the same or by different teachers.

Our analyses focus on the classroom level because aggregated student ratings of teaching quality are often used in the context of school evaluation or the monitoring of education systems. In this context, however, the low ICC1 and ICC2 values in our data are striking. Aligning with conventional interpretations, low ICC1 values indicate that ratings are inconsistent within classes, while low ICC2 values indicate a reliability

problem for the class mean rating (Lüdtke et al., 2009; Marsh et al., 2012). However, it is important to note that the ICC2 depends on the magnitude of the ICC1 and the number of raters. Recent discussions offer a nuanced view, suggesting the wording of items affects ICC1 values, with lower values for items addressing individuals rather than groups (Jaekel et al., 2022). Additionally, items with different referents show differentiated relationships with achievement and motivation (Flunger et al., 2024; Göllner et al., 2020; Jaekel et al., 2022; Rieser & Decristan, 2023). These findings suggest that the individual experience of teaching quality can make its own contribution to the analysis of teaching and therefore merits further attention (see also Rieser & Decristan, 2023; Igler et al., 2019; Willems, 2022).

Measurement invariance

Measurement invariance across subjects

With respect to MI across subjects, we found that metric invariance holds for cognitive activation and classroom management, while full invariance holds for supportive climate. At first, this finding may appear counter-intuitive as the three teaching quality dimensions are conceptualized as generic; that is, one might have expected full MI across subjects for all dimensions. However, researchers quite recently have raised serious doubts about the generic nature of teaching quality dimensions. Especially Praetorius et al. (2020), in their synthesis of the conceptualization of teaching quality in different subjects, have pointed out that while such a generic framework may fit most subjects at a rather general level, teaching quality dimensions actually require subject-specific modifications as their manifestation varies depending on the subject (see also Charalambous & Praetorius, 2020). Thus, MI across subjects needs not necessarily be full for all teaching quality dimensions due to differences in the meaning of specific items in the context of different subjects. From this perspective, our findings are in line with Praetorius and colleagues' (2020) argumentation and, consequently, underline the necessity of further theoretical discussion and empirical research on the subject specificity of teaching quality ratings.

Measurement invariance across TEACHER groups

In a next step, we tested for MI across the two TEACHER groups, that is, the degree of MI depending on whether mathematics and sciences classes were taught by the same or different teachers. Results indicate that metric invariance across groups holds for cognitive activation and supportive climate, whereas full invariance can be assumed for classroom management ratings.

While the latter finding is consistent with our expectations, results for cognitive activation and supportive climate are rather puzzling. Although this finding does not compromise the further analyses regarding our main hypotheses, technically, results imply that only associations are comparable across the two TEACHER groups, yet not the absolute level of cognitive activation or supportive climate, respectively. If this is indeed the case, the comparability of teaching quality ratings even within a single subject has to be seriously questioned.

Substantially, results suggest that students taught mathematics and science by the same teacher seem to understand the cognitive activation and supportive climate items differently than students taught by different teachers. We expect one possible explanation

in the organization of teaching in German primary schools. In Germany, many students are taught only by very few teachers in the course of their years in primary school. In such cases, oftentimes only one or two teachers cover all subjects in class. Accordingly, we assume that primary school students who have such limited experience with different teachers may not be as good at rating their teaching and may be less able to differentiate subjects and teachers, especially compared to more experienced students of the same age.

Correlation across subjects

Same teacher

As the level of invariance was sufficient, we finally focused on our main research question and investigated the correlations of the dimensions of teaching quality in different subjects in the two TEACHER groups. Correlation suggests that students do not differentiate between subjects in their teaching quality ratings when the same teacher teaches mathematics and science. This finding is supported by MCFA comparing one- and two-factor models for mathematics and science for each teaching quality dimension, favouring one-factor models across subjects within the same-teacher group.

For classroom management, this finding is in line with previous research (Praetorius et al., 2016). In contrast, previous research indicated at least some degree of differentiation when it comes to supportive climate (Praetorius et al., 2016). However, a closer look at the items used in the TIMSS 2015 assessment may at least partly explain why we did not find such differentiation in our study. In the study by Praetorius et al. (2016), items mainly focused on the teachers' enthusiasm and support for learning. In contrast, TIMSS 2015 items focused more on the social-emotional relationship between teachers and students. While teachers' enthusiasm and support may very well vary depending on the subject that is being taught, we expect this to be less likely for the quality of the teacher–student relationship, resulting in the inconsistent findings.

This consideration highlights an open issue within the present and other studies. Following the reasoning of Praetorius et al. (2018), each generic dimension of teaching quality consists of several subdimensions. In our data, each generic dimension was represented by only a single subdimension. Therefore, if different or more subdimensions were assessed, different results might be obtained.

For cognitive activation, we would have expected an even stronger differentiation between subjects, as pedagogical content knowledge, which is predictive of cognitive activation, may vary from subject to subject even within the same teacher. However, the items assessing cognitive activation in our study targeted rather general teaching practices (i.e., asking for explanations). Such practices might be employed by a teacher as a general teaching strategy, no matter which subject they teach at the moment, resulting in a high between-subject correlation. Consequently, we suggest further systematic research evaluating different ways of operationalization of cognitive activation to substantiate this thought.

Different teachers

In contrast to the same-teacher group, students taught by different teachers differentiated between each dimension of teaching quality in both subjects. For cognitive activation and supportive climate, the correlation across subjects was rather small and not statistically meaningful, whereas the correlation of classroom management across subjects was high

and statistically meaningful. Still, the latent correlation was not perfect, and additional MCFA favoured a two-factor model over a single factor for both subjects. We assume this high correlation to originate in the strong focus on students' behaviour in items assessing classroom management. Student behaviour is likely to be similar independent of the teacher or subject. This assumption is also supported by previous research pointing out that items referring to student behaviour show different connections with achievement and student characteristics than items addressing teacher behaviour (Göllner et al., 2020).

Comparison of the correlation between the TEACHER groups

The comparison of the correlation across the TEACHER groups only supported our hypotheses for classroom management and supportive climate but not for cognitive activation. As we had expected, the connections across subjects between the ratings of classroom management and supportive climate were stronger in the same-teacher group than in the different-teachers group. For cognitive activation, we found the same result although we had expected to find similar connections in both TEACHER groups. Possible reasons for the strong connection between the ratings of cognitive activation across subjects in the same-teacher group have already been discussed above.

Overall, our results create a mixed picture regarding the discriminate validity of students' ratings. When taught by different teachers, students seem to be able to differentiate between the dimensions of teaching quality in the different subjects. However, it must be remembered that in the different-teachers group teacher and subject are confounded. Thus, it cannot be said for certain whether the differentiation is caused by the different subjects or by the different teachers. When considering the results from the same-teacher group, that is, that students do not seem to differentiate between the teaching quality dimensions in two subjects if these are taught by the same teacher, it seems plausible that the differentiation in the different-teachers group is based on the different teachers rather than on the different subjects. At least for supportive climate and classroom management, these results are in line with theoretical considerations. Further research is needed regarding cognitive activation where different operationalizations are considered (e.g., Rieser & Decristan, 2023).

Note

1. Fisher's *z* transformation is not defined for values of 1. We have therefore used 0.999999 as the plug-in value if a correlation of 1 would have been included in the calculation.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Svenja Rieser is a member of the research staff at the Center for Research on Education and School Development (IFS) at the TU Dortmund University. Her research focuses on the assessment of teaching quality, particularly cognitive activation, and the promotion of self-regulated learning and the validity of student surveys.

Alexander Naumann is Substitute Professor for Educational Data Science at the Center for Research on Education and School Development (IFS) at the TU Dortmund University and a member of the research staff at the Department of Teacher and Teaching Quality (TTQ) at DIPF | Leibniz-Institute for Research and Information in Education, Frankfurt am Main, Germany. His primary research interests include validity of assessments, teaching quality, and longitudinal as well as multilevel latent variable models.

References

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., Krauss, S., Kunter, M., Löwen, K., Neubrand, M., & Tsai, Y.-M. (2009). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente* [Professional competence of teachers, cognitively activating mathematics instruction, and the development of students' mathematical literacy: Documentation of survey instruments (COACTIV)]. Max-Planck-Institut für Bildungsforschung.

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

Bellens, K., Van Damme, J., Van Den Noortgate, W., Wendt, H., & Nilsen, T. (2019). Instructional quality: Catalyst or pitfall in educational systems' aim for high achievement and equity? An answer based on multilevel SEM analyses of TIMSS 2015 data in Flanders (Belgium), Germany, and Norway. *Large-scale Assessments in Education*, *7*, Article 1. https://doi.org/10.1186/s40536-019-0069-2

Benton, S. L., & Cashin, W. E. (2011). *Student ratings of teaching: A summary of the research and literature* (IDEA Paper No. 50). https://ideacontent.blob.core.windows.net/content/sites/2/2020/01/PaperIDEA_50.pdf

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons. https://doi.org/10.1002/9781118619179

Charalambous, C. Y., & Kyriakides, E. (2017). Working at the nexus of generic and content-specific teaching practices: An exploratory study based on TIMSS secondary analyses. *The Elementary School Journal*, *117*(3), 423–454. https://doi.org/10.1086/690221

Charalambous, C. Y., & Praetorius, A.-K. (2020). Creating a forum for researching teaching and its quality more synergistically. *Studies in Educational Evaluation*, *67*, Article 100894. https://doi.org/10.1016/j.stueduc.2020.100894

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

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

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., Hondrich, A. L., Rieser, S., Hertel, S., & Hardy, I. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal*, *52*(6), 1133–1159. https://doi.org/10.3102/0002831215596412

Doyle, W. (2006). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 97–125). Lawrence Erlbaum Associates.

Fauth, B., Decristan, J., Decker, A.-T., Büttner, G., Hardy, I., Klieme, E., & Kunter, M. (2019). The effects of teacher competence on student outcomes in elementary science education: The mediating

role of teaching quality. *Teaching and Teacher Education*, *86*, Article 102882. https://doi.org/10.1016/j.tate.2019.102882

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507–521. https://doi.org/10.2307/2331838

Flunger, B., Verdonschot, A., Zitzmann, S., Hornstra, L., & van Gog, T. (2024). A Bayesian approach to students' perceptions of teachers' autonomy support. *Learning and Instruction*, *91*, Article 101873. https://doi.org/10.1016/j.learninstruc.2023.101873

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press/Chapman & Hall.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), Article 555. https://doi.org/10.3390/e19100555

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. National Comprehensive Center for Teacher Quality. https://gtlcenter.org/sites/default/files/docs/EvaluatingTeachEffectiveness.pdf

Göllner, R., Fauth, B., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Do student ratings of classroom management really tell us more about teachers or about classroom composition? *Zeitschrift für Pädagogik. Beiheft*, *66*, 156–172. https://doi.org/10.3262/ZPB2001156

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, *52*(11), 1182–1186. https://doi.org/10.1037/0003-066X.52.11.1182

Gruehn, S. (2000). *Unterricht und schulisches Lernen: Schüler als Quellen der Unterrichtsbeschreibung* [Teaching and school learning: Students' ratings as a source to describe teaching]. Waxmann.

Igler, J., Ohle-Peters, A., & McElvany, N. (2019). Mit den Augen eines Grundschulkindes [Through the eyes of a primary school child]. *Zeitschrift für Pädagogische Psychologie*, *33*(3–4), 191–205. https://doi.org/10.1024/1010-0652/a000243

Jaekel, A.-K., Wagner, W., Trautwein, U., & Göllner, R. (2022). "The teacher motivates us – or me?" – The role of the addressee in student ratings of teacher support. *Contemporary Educational Psychology*, *71*, Article 102120. https://doi.org/10.1016/j.cedpsych.2022.102120

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

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7

Kunter, M., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction*, *18*(5), 468–482. https://doi.org/10.1016/j.learninstruc.2008.06.008

Lenske, G., Wagner, W., Wirth, J., Thillmann, H., Cauet, E., Liepertz, S., & Leutner, D. (2016). Die Bedeutung des pädagogisch-psychologischen Wissens für die Qualität der Klassenführung und den Lernzuwachs der Schüler/innen im Physikunterricht [The role of pedagogical/psychological knowledge for classroom management and learning gains in physics lessons]. *Zeitschrift für Erziehungswissenschaft*, *19*(1), 211–233. https://doi.org/10.1007/s11618-015-0659-x

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

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

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*(1), 53–76. https://doi.org/10.1207/s15327906mbr3201_3

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, *34*(2), 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Marder, J., Göllner, R., Wagner, W., & Fauth, B. (2021). Ask me, I (Dis)agree! Acquiescence in student ratings of teaching quality in German vocational schools. *Studies in Educational Evaluation*, *68*, Article 100937. https://doi.org/10.1016/j.stueduc.2020.100937

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*(2), 106–124. https://doi.org/10.1080/00461520.2012.670488

Meng, X.-L. (1994). Posterior predictive *p*-values. *Annals of Statistics*, *22*(3), 1142–1160.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Muthén, B., & Asparouhov, A. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education*, *14*(2), 135–153. https://doi.org/10.1023/A:1008102519702

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*(2), 365–397. https://doi.org/10.3102/0002831207308230

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119. https://doi.org/10.3102/0013189X09332374

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM Mathematics Education*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A.-K., Klieme, E., Kleickmann, T., Brunner, E., Lindmeier, A., Taut, S., & Charalambous, C. Y. (2020). Towards developing a theory of generic teaching quality: Origin, current status, and necessary next steps regarding the three Basic Dimensions model. *Zeitschrift für Pädagogik*, *Beiheft*, *66*, 15–36. https://doi.org/10.3262/ZPB2001015

Praetorius, A.-K., Vieluf, S., Saß, S., Bernholt, A., & Klieme, E. (2016). The same in German as in English? Investigating the subject-specifity of teaching quality. *Zeitschrift für Erziehungswissenschaft*, *19*(1), 191–209. https://doi.org/10.1007/s11618-015-0660-4

Rieser, S., & Decristan, J. (2023). Kognitive Aktivierung in Befragungen von Schülerinnen und Schülern: Unterscheidung zwischen dem Potential zur kognitiven Aktivierung und der individuellen kognitiven Aktivierung [Cognitive activation in student questionnaires: Distinguishing between the potential for cognitive activation and individual cognitive activation]. *Zeitschrift Für Pädagogische Psychologie*. Advance online publication. https://doi.org/10.1024/1010-0652/a000359

Stahns, R., Rieser, S., & Hußmann, A. (2020). Können Viertklässlerinnen und Viertklässler Unterrichtsqualität valide einschätzen? Ergebnisse zum Fach Deutsch [Are fourth grade students able to rate instructional quality validly? Results from German Language classes]. *Unterrichtswissenschaft*, *48*(4), 663–682. https://doi.org/10.1007/s42010-020-00084-6

Stan Development Team. (2023). *RStan: The R interface to Stan* (R package Version 2.32.3). http://mc-stan.org/

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520. https://doi.org/10.3102/1076998616646200

van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*, *30*(1), 30–50. https://doi.org/10.1080/09243453.2018.1539015

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P., Paananen, T., & Gelman, A. (2022). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models* (R package Version 2.5.1). https://mc-stan.org/loo/

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, *28*, 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, *108*(5), 705–721. https://doi.org/10.1037/edu0000075

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal*, *53*(6), 1834–1868. https://doi.org/10.3102/0002831216671864

Wendt, H., Bos, W., Köller, O., Selter, C., & Schwippert, K. (2018). *Trends in International Mathematics and Science Study 2015 (TIMSS 2015)* (Version 1) [Data set]. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.5159/IQB_TIMSS_2015_v1

Willems, A. S. (2022). Individuelle Schüler*innenprofile des situationalen und dispositionalen Interesses und ihre Bedeutung für die Wahrnehmung der Unterrichtsqualität im Fach Mathematik [Students' individual profiles of situational and dispositional interest and their relevance for students' perceptions of teaching quality in mathematics education]. *Zeitschrift für Erziehungswissenschaft*, *25*(2), 377–404. https://doi.org/10.1007/s11618-022-01094-z

Appendix 1. Items on teaching quality

Cognitive activation

Our mathematics /science teacher …

asks me what I have and what I haven't understood.
asks us what we already know about a new topic.
wants me to explain my answers.

Supportive climate

Our mathematics/science teacher …

is nice to me even when I make a mistake.
cares about me.
tells me how to do better when I make a mistake.
likes me.
believes that I can solve difficult tasks.

Classroom management

How often does this happen in your mathematics/science class?

The students don't listen to the teacher.
The class is noisy and unsettled.
Our teacher has to wait for a long time until all students are quiet.
The students can't work well.
The students only start work long after the lesson has begun.