Discover Education

Research

Observing effective classroom management in early instruction in primary school: rating instrument construction and its link to teacher knowledge

Johannes König[1] · Nina Glutsch[1] · Jonas Weyers[1] · Gino Casale[3] · Petra Hanke[2] · Chantal Knips[2] · Thorsten Pohl[2] · Tina Waschewski[4] · Michael Becker-Mrotzek[2] · Alfred Schabmann[2] · Birgit Träuble[2]

Received: 9 February 2023 / Accepted: 5 September 2023
Published online: 14 September 2023

Abstract

This study presents a novel standardized rating instrument for observing and measuring effective classroom management (ECM) as part of the teaching and learning environments in primary school. The instrument comprises eight high-inferent items on organizational aspects (lack of disruptions/discipline problems, withitness, effective time use, clear rules, clear routines, appreciation) and instructional aspects (structuring, goal clarity). It was applied in second grade classrooms of German primary school teachers ($n = 35$) providing early reading and writing instruction. Pairs of trained raters (student teachers) observed one to three lessons in each classroom over 3–4 months, reaching acceptable interrater agreement. The instrument showed acceptable internal consistency. Factor analyses revealed structures with good to acceptable fit indices, with confirming the differentiation into organizational and instructional ECM aspects. Correlations between observed ECM and facets of teacher knowledge (that were directly assessed by using paper–pencil tests) provide divergent and convergent validity evidence: Whereas no significant correlations could be found between pedagogical content knowledge for early reading and writing instruction and ECM, findings show significant correlations between general pedagogical knowledge and the ECM. The added value of the study is therefore to provide a novel instrument that can be applied in future empirical research on primary school classroom management.

Keywords Effective classroom management observation · Early reading and writing instruction · Primary school teaching · Rating instrument · Teacher knowledge

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44217-023-00058-7.

✉ Johannes König, johannes.koenig@uni-koeln.de; Nina Glutsch, nina.glutsch@swk.kmk.org; Jonas Weyers, jonas.weyers@uni-koeln.de; Gino Casale, gcasale@uni-wuppertal.de; Petra Hanke, petra.hanke@uni-koeln.de; Chantal Knips, chantal.knips@uni-koeln.de; Thorsten Pohl, thorsten.pohl@uni-koeln.de; Tina Waschewski, tina.waschewski@bildung.bremen.de; Michael Becker-Mrotzek, becker.mrotzek@uni-koeln.de; Alfred Schabmann, alfred.schabmann@uni-koeln.de; Birgit Träuble, b.traeuble@uni-koeln.de | [1]Present Address: Empirical School Research, University of Cologne, Gronewaldstr. 2, 50931 Cologne, Germany. [2]University of Cologne, Cologne, Germany. [3]University Wuppertal, Wuppertal, Germany. [4]Institut Für Qualitätsentwicklung Im Land Bremen, Bremen, Germany.

1 Introduction

Classroom management relates to teacher behavior and action providing learning environments free from interference or distractions as well as ensuring effective time use in classroom teaching and learning [1, 2]. As an educational concept, classroom management serves to describe specific measures taken by the teacher such as providing clear rules, structuring the lesson, or monitoring student behavior [3]. Managing classrooms is a core requirement expected from a professional teacher [4].

For decades, many researchers have contributed to a scientific understanding of classroom management [5, 6]. Several meta-analyses of empirical studies proliferated reliable evidence that classroom management positively affects student cognitive and affective-motivational learning outcomes (e.g., [7–10]). Today, broad agreement exists that effective classroom management (ECM) is a core constituent of teaching practice and a basic dimension of instructional quality [11, 12].

However, the reliable assessment of instructional quality is a complex process [13, 14]. Classroom observations are necessary and lesson evaluation has to be underpinned by scientific concepts. As teachers, their students, and external observers may assess the same lesson differently, their varying perspectives can influence empirical studies' results [15]. While student ratings were frequently used as a measure for ECM (e.g., [15, 16]), ratings by external and trained raters are favored to accesses teachers' instructional quality directly and prevent self-reported bias [13]. This is particularly important with early primary school students as they can be overwhelmed by reading and responding to rating scales in ECM questionnaires.

Different subject-specific (e.g., [17]) and generic instruments (e.g., [18]) for classroom observations were used in various studies. However, only few studies considered generic aspects of ECM that could be applied in early instruction in primary school yet (e.g., [19, 20]), using video recording with time-consuming coding mainly rather than efficient in-vivo ratings. Against this background and to proliferate the development of specific learning environments research instruments, we ask how ECM can be observed and measured reliably and validly by ratings of trained observers in early instruction in primary schools using a novel instrument. The added value of the study is therefore to develop such an instrument that can be applied in future empirical research on primary school classroom teaching and learning in general, including early reading and writing instruction which is the overall focus of the project Professional Knowledge of Teachers, Instructional Quality, and Second Grade Student Progress in Reading and Writing (*Professionelles Wissen von Lehrkräften, Unterrichtsqualität und Lernfortschritte von Schülerinnen und Schülern im basalen Lese- und Schreibunterricht* – WibaLeS) forming the context of present instrument's evolution and development.

2 Theoretical framework

2.1 Effective classroom management

Over decades and following the process–product research paradigm, empirical educational researchers have been investigating student learning outcomes—mostly student achievement—and their link to relevant learning environment characteristics [21, 22]. Numerous studies analyzed classroom instruction components and their possible influence on student learning. Well-known meta-analyses (e.g., [7, 9, 10, 23]) offer systematic and integrating perspectives of many studies' findings, highlighting the most relevant factors for effective teaching. Specifically, Hattie [23] found that relevant factors describing effective instruction are "classroom management" (effect size of $d=0.52$; p. 102), "teacher clarity" ($d=0.75$; p. 126), and "teacher-student relationships" ($d=0.72$; p. 118). According to Hattie's [7] meta-analysis, effect sizes greater than 0.4 are in the "zone of desired effects," indicating high practical relevance regarding the effects of teaching and student outcomes.

To conceptualize ECM, some authors suggested differentiating between organizational and instructional aspects of classroom management [24–26, p. 79]. The former relates to successfully dealing with time management and student behavior, for example, by working with clear rules and routines [6, 26–29], monitoring student behavior [1, 2, 28], and establishing positive teacher–student relationships [30]. The latter comprises pedagogical issues such as teaching methods, structuring lessons, and instructional clarity [1, 25, 31]. The absence of disruptions and students' time on task can be considered outcomes of teachers' ECM [2, 32].

On an aggregated level, ECM aspects have been subsumed into one basic dimension of instructional quality. Following the Third International Mathematics and Science Study (TIMSS), several approaches emerged on basic instructional quality dimensions (e.g., [18, 33]) claiming that these approaches are generic and can be observed in any teaching–learning situation but with different quality manifestations. Besides students' cognitive activation and constructive support, ECM is usually conceptualized as one of three instructional quality dimensions that are considered crucial for high-quality teaching "across education systems, school types, grade levels, and school subjects" [22, p. 193].

2.2 Live classroom observation of ECM

When assessing and measuring ECM as part of learning environments and instructional quality, various methodological approaches are possible. Ratings can be done live and on-site in schools or via video-based lessons or sequences. Further, observers may vary: lesson observations rating can occur through teachers themselves, their students, or external observers. However, external observers are considered being more reliable in their ratings and less subjective in their judgements as they are usually specifically trained and observe different lessons of different teachers [13]. When rating lessons on-site, observers have to decide about complex ad-hoc situations. Here, difficulties due to events taking place in a very fast manner and sometimes simultaneously with instant judgements and lessons that cannot be repeated [34]. However, there are several advantages of live observations over video-taped lessons, such as being in the middle of the events and seeing the whole classroom instead of only what camera angles show.

When developing observational instruments, differences are made between low- or high-inferent items (e.g., [19]). The latter demand complex and fast situational assessments for lessons. Raters have to make decisions for each segment—for example, identifying the quality manifestation, how teachers behave, and how students act or react to teachers' expectations and instructions. High-inferent items often need a lot more interpretation as they relate to "deep structures" during a lesson, such as to students' profoundness of thinking, whereas lower inferent items refer to so-called "surface structures", such as how often an event related to an indicator takes place [31, 35, p. 35]. Rater training should therefore aim at a mutual understanding of indicators' meaning and quality manifestations. The newly developed instrument of the present study contains mostly high-inferent items and requires the assessment of deep structures related to ECM.

2.3 Correlations between effective classroom management and teacher knowledge

Developing a novel rating instrument includes determining its adequacy based on testing for validity evidence. *Validity* refers to how plausible the interpretation of a test score is supported by evidence and theory for the proposed use of a test [36]. In case of in-vivo observations of ECM practices, the observed scores are used to draw inferences about teachers' actual performance and teacher-student interaction in the classroom. Therefore, those interpretations can only be made with plausibility if it is shown that the observational instrument's results are psychometrically sound with theoretical and empirical knowledge about classroom teachers' ECM. That means, empirical relationships to other variables that are related in a plausible way to the interest construct are a central source [37].

As teachers' professional knowledge is supposed to affect their performance [38], the two facets of teachers' professional knowledge—general pedagogical knowledge (GPK) and pedagogical content knowledge (PCK)—are included in correlational analyses with observed instructional quality. Teachers' GPK is a strong predictor of instructional quality and studies showed significant positive correlations between GPK and ECM aspects (e.g., [39–41]). Additionally, ECM is related to student achievement [7] with GPK affecting it and thus linked to student performance [42, 43]. PCK, on the contrary, is a stronger predictor for subject-specific instructional quality, in particular for cognitive activation [44]. Therefore, in relating PCK and GPK to ECM observational and rating data, hypotheses on convergent (GPK) and divergent (PCK) validity evidence can be tested as shown below.

3 The present study

3.1 Rating instrument construction

This study focuses on the construction and implementation of an observation instrument for assessing ECM at the beginning of primary school—particularly applicable to early instruction in primary school, therefore including early reading and writing instruction, but not restricted to language teaching. First, we were seeking for an existing instrument with few, high-inferent items especially applicable in early instruction in primary school. In the English-speaking world, the Classroom Assessment Scoring System (CLASS) [18] is frequently applied for this purpose, however, drawing on a different model of instructional quality than German-language studies [12]. This impeded us from implementing the CLASS to the context of our project WibaLeS: In our project we aimed to assess the quality of early reading and writing instruction comprehensively. To do so, we adopted a model of (measuring) instructional quality along the three basic dimensions (1) effective classroom management, (2) constructive support, and (3) cognitive activation (e.g., [11, 12]). This model is well established in Germany as the country of our study (e.g., [15, 17, 25, 41, 42, 44]). Therefore, the present observation instrument was developed to assess ECM as a *generic* basic instructional quality dimension in this project–besides and conceptually demarcated from the *subject-specific* basic dimensions of constructive support and cognitive activation (both of which are not part of the present study, but still make up an important component of our project). In addition, the CLASS instrument includes three complex 7-point scales to capture classroom organization, each of which combines indicators of different aspects of classroom management. For our measurement approach, however, our goal was a larger number of differentiated items, each with a focus on one aspect of classroom management.

Within the German-speaking area, the observation instrument of the so-called PERLE (*Persönlichkeits- und Lernentwicklung an staatlichen und privaten Grundschulen*) video study should be highlighted (e.g., [45]), which is used to rate videos of early mathematics and language instruction retrospectively. This instrument was conceptually close to the theoretical framework of our project. In addition, it used a similar rating approach with a larger number of categories each targeting a specific aspect of classroom management (e.g., rule clarity). However, the complex and effortful video recording was neither feasible nor necessary in our project context. In contrast, the instrument of the present study aims to enable in-vivo classroom observation and is intended to include a comparably compact selection of categories that consider relevant aspects of ECM in primary school teaching situations. We planned for using a relatively small number of observable indicators to facilitate in-vivo observation, but enough indicators to make a reliable measurement (e.g., [35, 37]). Its implementation in classroom environment is far less effortful compared with video recordings (e.g., concerning data protection requirements in Germany and management of collected data). At least in German schools, access to the field for empirical investigation of instructional quality sometimes might be denied if video recordings are selected as the means of data collection, whereas in vivo ratings have the advantage to open up the field for empirical studies. This was one of the main reasons why our project team decided to develop a novel instrument capturing ECM.

Consequently, for instrument development, ECM as an important generic facet of learning environments, was differentiated into several aspects. ECM results in quality and quantity of actual learning time provided for students (time-on-task) [2, 32]. Due to its complexity, ECM comprises various aspects such as teacher efficacy when dealing with classroom disruptions or disciplinary conflicts [16, 31] and involves the degree of alertness and attentiveness of teachers to student behavior. ECM features structured and well-organized lessons with clear rules and routines [2, 11, 16, 46]. This is particularly relevant in early instruction in primary school, since students still have to get used to typical social interaction of classroom learning (e.g., raising one's hand in order to participate in whole-group discussion, following the teacher's instruction for transition from single to group work, etc.).

When developing the rating instrument, we used and combined existing classroom management categories from several empirical studies [2, 11, 12, 17, 33, 47] and adapted and enhanced them using additional indicators. In particular, we sought to select categories that seemed appropriate for early instruction in primary school. These included a number of categories already used for the similar instrument in the PERLE project, for example, global indicators, such as effective time use or (lack of) disruptions, which generally reflected successful classroom management, and organizational aspects of classroom management (e.g., withitness, establishing clear rules). However, we additionally included instructional aspects, that is, structuring and goal clarity (e.g., [24]), taking into account that a clear lesson structure and transparent communication of learning objectives can play an important role in establishing a learning environment. Moreover, a separate category focusing on clear routines was included, which may be regarded as a key requirement of primary school instruction, since first and second grade students are not yet familiar with instructional routines [45]. In addition, teachers' appreciation—an aspect that affects both social climate and classroom management—was considered. This also takes into account the assumption that the relationship between teacher and students affects classroom interactions, especially with younger children.

In general, the item development accounted for primary school teaching and learning, explicitly including early instruction where the observational instrument was to be applied in our study. Finally, the instrument comprised the following eight high-inferent items:

- lack of classroom disruptions and discipline problems (DIS),
- withitness (WIT),
- effective time use (ETU),
- clear rules (CRU),
- clear routines (CRO),
- appreciation (APP),
- structuring (STR), and
- goal clarity (GCL).

Each item comprises four categories, which are described qualitatively (1 = low quality/low manifestation of instructional quality; 4 = high quality/high manifestation of instructional quality) (see, for further details, Additional file 1: Appendix Table S1).

3.2 Research questions

The present study serves to investigate whether the novel rating instrument captures the ECM quality. We therefore address the following research questions:

1.1. Does the novel rating instrument provide a reliable measurement of ECM as one basic instructional quality dimension in early instruction in German primary schools?

We assume that our rating instrument is appropriate to depict ECM by conducting confirmatory factor analyses (CFA) and testing for different model structures. Drawing on classroom management as a basic instructional quality dimension, we first investigated whether all items used for the rating instrument represent aspects of one latent construct. Second, we tested for a two-factor model that differentiates between organizational and instructional classroom management aspects (Table 1). The former is related to the "social dimension" and "time-related dimension" of teaching [48, p. 467] and encompasses the items, "lack of classroom disruptions and disciplinary problems, withitness, effective time use, clear rules, clear routines, [and] appreciation." The remaining items—structuring and goal clarity—are conceptualized as instructional classroom management aspects as they depict the "content dimension" of teaching [48, p. 469]. However, these are not necessarily subject-specific, but focus on instructional teaching content processes from a general standpoint [49]. So, these items are generic instruction aspects featured in subject-specific instruction rather than subject instructions reflecting particular disciplines, that is, in our case, the specific aspect of early reading and writing.

2.2. Can validity evidence be provided for ECM based on relations to teachers' professional knowledge?

Table 1 Measures of interrater reliability of rating items capturing effective classroom management

Aspects of classroom management	Item	Plenary phase				Working phase			
		n	Kappa	*ICC*	*PA* (%)	*n*	Kappa	*ICC*	*PA* (%)
Organizational aspects									
Lack of disruptions	DIS	83	0.716	0.717	88	85	0.677	0.702	84
Withitness	WIT	83	−0.025	−0.019	95	83	0.534	0.675	82
Effective time use	ETU	83	0.366	0.419	81	81	0.361	0.313	79
Clear rules	CRU	83	0.399	0.404	76	85	0.627	0.631	84
Clear routines	CLO	82	0.161	0.166	81	84	0.258	0.319	76
Appreciation	APP	82	0.345	0.351	87	82	0.308	0.314	95
Instructional aspects									
Structure	STR	82	0.182	0.183	92	84	0.325	0.358	87
Goal clarity	GCL	83	0.659	0.662	93	84	0.464	0.540	81

Statistics are based on a concurrent analysis of the three measurement time points. *ICC* Intraclass correlation, one-way random *ICC* (1,1), *PA* Percent agreement

The present study aims to provide validity evidence based on relations to other constructs. As instructional quality of teachers (i.e., performance) is related to their professional knowledge [38, 42, 44], we hypothesize that high-quality manifestations of ECM as an instructional quality dimension irrespective of specific school subjects are significantly positively related to teachers' GPK (convergent evidence), but only weakly related to their PCK of early reading and writing instruction (divergent evidence) due to its subject-specific character. Because previous investigations on teachers' GPK and PCK came to conclude the two knowledge facets are substantially inter-correlated, accounting for both facets using different hypotheses related to convergent and divergent evidence can strengthen validity evidence for ECM based on relations to teachers' professional knowledge. Moreover, such empirical evidence could highlight the specific need for teachers' GPK *besides* PCK as part of the professional teacher's knowledge base [40, 41, 43]. We consider this examination also an important issue with regards to our conceptualizing of instructional aspects of classroom management, which, as mentioned above (see the outline of our first research question), focuses on instructional teaching content processes from a general standpoint rather than from a subject-specific perspective.

4 Material and methods

4.1 Participants and procedure

The data derives from the project WibaLeS that focuses on early reading and writing instruction, in which primary school teachers in the administrative district of Cologne, Germany participated. Conducted between fall 2020 and spring 2021, the COVID-19 pandemic affected data collection procedures, causing partial school closures and reduced class sizes. Since schools were subjected to strict rules for external visitors, some teachers dropped out during our study, causing a decrease in the number of participating teachers from 35 at the time of the first classroom observation to 29 in the second and 23 in the third.

An online questionnaire for assessing teachers' GPK and PCK and gathering demographic information was completed by 30 teachers who, on average, were 41 years old ($SD = 8.99$) and had 12.5 years of teaching experience ($SD = 7.42$). The sample consisted of 26 women (90%) and four men, a distribution typical for the profession of primary teachers in Germany. Except for one person who had originally studied to become a teacher for upper secondary school (academic track), all had been trained as primary school teachers (non-academic track).

All teachers participated voluntarily and were recruited first via email invitation sent to all primary schools in greater Cologne region, Germany, and second via central information events for principals in the administrative district. Teachers got reimbursed with a financial compensation of 50 Euros. Keeping data privacy was in accordance with the researchers' university's provisions and commissioner for data protection. The study was carried out in compliance with regulations by the German Research Foundation (*Deutsche Forschungsgemeinschaft*, DFG).

4.2 Tests measuring teacher knowledge

Teachers' GPK was measured using the standardized test developed in TEDS-M [50]. The test items address knowledge on generic teaching dimensions: structuring, motivation, classroom management, adaptivity, and assessment. Moreover, they can be divided into three categories regarding the quality of cognitive demands: when responding to test items, teachers are required to *recall* knowledge elements; *understand/analyze* concepts and situations based on GPK; *generate* options for action using their GPK. The test design can thus be described by a 3 × 4 matrix (cognitive demands x knowledge dimensions). To avoid cognitive overload, a short form of the original TEDS-M instrument was implemented, which included 15 complex test items (five multiple choice items, 10 open-response items) resulting in 60 dichotomous test items for scaling analysis. It took approximately 20 min to complete. PCK was measured using an instrument developed by our research team [51]. It comprises acquisition of reading and writing, including child development and diagnostics, instruction, including curricular aspects, and teaching strategies (29 items: 26 multiple choice items, three open-response items).

Open-response items of both tests were coded by two experienced coders, respectively, based on comprehensive coding manuals. Since the GPK test is a frequently used and established instrument, double coding was only necessary for a subsample of 10 cases. As the PCK test is less frequently used, the entire sample of 30 cases was double coded. Average interrater reliability was good for both instruments (GPK: $M_{\text{Kappa}} = 0.74$, $SD_{\text{Kappa}} = 0.27$; PCK: $M_{\text{Kappa}} = 0.80$, $SD_{\text{Kappa}} = 0.16$).

Scaling analysis based on item response theory (IRT) was conducted using the software package *Conquest* [52] and revealed acceptable reliability values for both GPK (*WLE* = 0.676, *EAP* = 0.724) and PCK (*WLE* = 0.58, *EAP* = 0.62). The weighted mean square of items did not exceed the recommended range (0.80–1.20) [53]. For the subsequent analysis, weighted likelihood estimates (*WLE*) [54] were used as ability parameters. Additionally, the GPK test was used to construct subscales, which was possible for items measuring GPK of motivating students (12 items, α = 0.66), structuring lessons (18 items, α = 0.59), and adaptivity (12 items, α = 0.53). These subscales will be used for providing further validity evidence related to our second research question.

4.3 Rating instrument

Observation and rating occurred directly on-site in schools by a team of two trained observers that observed in different constellations. Each rater had a coding sheet with items and a 4-point rating scale (see Additional file 1: Appendix Table S1). Raters had to check off one of the possible four quality manifestations. They were advised to rate the stricter category if they were unsure between two scale points. At the end of the coding sheet, raters were asked to evaluate the rating difficulty on a 4-point Likert-scale. In 60% of all ratings, observers provided such an evaluation. 90.4% found the rating "not at all difficult" or "rather not difficult." Considering that raters evaluated the rating difficulty when facing challenges in particular, the rating procedure is interpreted as predominantly feasible.

Teachers had not been asked to prepare specific lessons. However, the rating took place in lessons in which the teacher was in charge of the whole class. The rating took place twice during a typical lesson of 45 min, mainly for two reasons: Ratings should be processed as close to the observed situation as possible and because every lesson, despite structural variation, started basically with a plenary phase followed by a working phase, two ratings were carried out to capture ECM comprehensively. First, the plenary phase had to be rated after about 15 to 20 min of the lesson, depending on the individual duration of this phase. Even if the introductory plenary phase was shorter (e.g., 5 min only), this requirement assured that rating was not delayed too much, for example, until the end of the lesson. The working phase succeeding the plenary phase was rated separately after again 15–20 min (again to prevent rating delays) using the same items (except for group focus due to its specific nature related to the plenary phase) before the lesson was ended with a feedback phase, which was not rated separately. There were no deviating lesson patterns, so the rating of the two phases was applicable to every lesson without any difficulties.

Teacher education students who were at least in the fourth semester of their bachelor studies were recruited to be trained as reliable observers. Altogether, eight raters were trained at three time points for several hours in the weeks before data collection started. They had to read introductory theoretical texts beforehand and the concept of ECM was explained at the beginning of training. Their attention was also called to their implicit beliefs about "good teachers" and rater bias. Next, the instrument and rating manual were explained, after which they had to rate several example lessons on the basis of video-taped lessons. Ratings were subsequently discussed and compared to reach a mutual understanding of the instrument's conceptual underpinning. Interrater reliability was checked by the project team, and critical items were discussed again. In some cases, items had to be adjusted to be more concise.

Interrater reliability was investigated for the plenary and working phases using Cohen's Kappa, intraclass correlation coefficients (*ICC*), and percent agreement (Table 1). All observations for each item were included in a concurrent analysis of interrater reliability. When considering percent agreement values, all items show high or very high agreement by raters. However, it should be noted that Kappa values and *ICC* show partly low values especially for the items WIT, STR, and CRO (Table 1). This is due to high agreements on the highest category on the one hand, and a low base rate of lower categories combined with relatively low agreement on the other hand.

4.4 Analyses

For answering our first research questions, CFAs were conducted to provide validity evidence based on internal structure. The different factor models were compared considering the following indices of model fit: comparative fit index (*CFI*), root mean square error of approximation (*RMSEA*), standardized root mean square residual (*SRMR*), and the results of the Chi^2 goodness of fit test (Chi^2/df). For the comparison of models, internal scale consistency (Cronbach's α) was taken into account.

For answering our second research question, to provide divergent and convergent validity evidence [55], the relationship between ECM and professional knowledge was investigated using Spearman's rank correlation coefficient because of the small sample size and the distribution of rating scores being left-skewed. To investigate whether the correlation between observed ECM and GPK differs significantly from that between the former and PCK, *z*-tests for comparing correlations from dependent samples were applied [56]. Due to small teacher sample size, we use the 10% significance level to reduce the risk to miss statistically significant correlations.

In the course of conducting the study and our analyses, it became apparent that the sample studied was a positive selection in terms of teaching quality and teachers' professional knowledge. This circumstance of limited variance may bias statistical findings. Therefore, to explore the potential of the instrument, separate generalizability studies were conducted for all indicators. This allowed us to identify those indicators for which variance was attributable to differences between the teachers. Assuming that these indicators have particular diagnostic value in our sample, a scale was constructed using these indicators and linked to the knowledge tests. It should be noted that this procedure is exploratory and significance testing using this scale should be interpreted with caution. A short description of the generalizability study approach can be found in the Additional file 1: Appendix.

Table 2 Descriptive item statistics

Observation 1	Plenary phase (*n* = 33)			Working phase (*n* = 34)		
	M (*SD*)	Min	Max	*M* (*SD*)	Min	Max
DIS	3.64 (0.46)	3.00	4.00	3.57 (0.46)	3.00	4.00
WIT	3.94 (0.17)	3.50	4.00	3.62 (0.57)	2.00	4.00
ETU	3.73 (0.47)	2.00	4.00	3.78 (0.35)	3.00	4.00
CRU	3.68 (0.39)	3.00	4.00	3.71 (0.43)	3.00	4.00
CRO	3.89 (0.27)	3.00	4.00	3.87 (0.26)	3.00	4.00
APP	3.89 (0.27)	3.00	4.00	3.90 (0.27)	3.00	4.00
STR	4.00 (0.00)	4.00	4.00	3.85 (0.32)	3.00	4.00
GCL	3.89 (0.30)	3.00	4.00	3.75 (0.35)	3.00	4.00
Observation 2	Plenary phase (*n* = 29)			Working phase (*n* = 29)		
	M (*SD*)	Min	Max	*M* (*SD*)	Min	Max
DIS	3.71 (0.43)	3.00	4.00	3.43 (0.44)	3.00	4.00
WIT	4.00 (0.00)	4.00	4.00	3.88 (0.26)	3.00	4.00
ETU	3.88 (0.29)	3.00	4.00	3.81 (0.34)	3.00	4.00
CRU	3.72 (0.39)	3.00	4.00	3.60 (0.41)	3.00	4.00
CRO	3.81 (0.28)	3.00	4.00	3.76 (0.35)	3.00	4.00
APP	3.83 (0.31)	3.00	4.00	3.98 (0.09)	3.50	4.00
STR	3.88 (0.26)	3.00	4.00	3.91 (0.23)	3.00	4.00
GCL	3.84 (0.33)	3.00	4.00	3.81 (0.39)	3.00	4.00
Observation 3	Plenary phase (*n* = 22)			Working phase (*n* = 22)		
	M (*SD*)	Min	Max	*M* (*SD*)	Min	Max
DIS	3.80 (0.37)	3.00	4.00	3.52 (0.57)	2.00	4.00
WIT	4.00 (0.00)	4.00	4.00	3.79 (0.52)	2.50	4.00
ETU	3.82 (0.29)	3.00	4.00	3.74 (0.40)	3.00	4.00
CRU	3.80 (0.33)	3.00	4.00	3.72 (0.45)	3.00	4.00
CRO	3.91 (0.20)	3.50	4.00	3.76 (0.42)	2.50	4.00
APP	3.95 (0.15)	3.50	4.00	4.00 (0.00)	4.00	4.00
STR	3.95 (0.15)	3.50	4.00	3.89 (0.34)	2.50	4.00
GCL	3.91 (0.25)	3.00	4.00	3.70 (0.49)	2.00	4.00

Table 3 Results of confirmatory factor analyses (working and plenary phases)

Item	Model 1	Model 2	
	Classroom management (all items)	Organizational aspects	Instructional aspects
DIS	0.945	0.968	
WIT	0.422	0.407	
ETU	0.336	0.327	
CRU	0.650	0.645	
CRO	0.489	0.477	
APP	0.135	0.131	
STR	0.349		0.781
GCL	0.244		0.529
Fit-indices			
Chi^2 (*p*-value)	54.941 (< 0.001)	30.636 (< 0.05)	
Degrees of freedom	20	19	
RMSEA	0.101	0.060	
SRMR	0.074	0.057	
CFI	0.848	0.949	

Table 4 Means and standard deviations for observed effective classroom management and teacher knowledge

	n	*M*	*SD*	M_{Points} (*SD*)
Class observation scales				
(1) Classroom management (all items)	36	3.82	0.11	–
(2.1) Factor 1: Organizational aspects	36	3.79	0.14	–
(2.2) Factor 2: Instructional aspects	36	3.87	0.13	–
(3) Selected indicators	36	3.78	0.14	–
Professional knowledge tests				
GPK	30	0.94	0.60	41.97 (5.54)
PCK	30	1.13	0.77	20.17 (3.31)

Mean scores of class observations represent aggregated means across all three observation time points

5 Results

5.1 Descriptive statistics

Table 2 shows the means and standard deviations of each item separately for phases and observational time points. High-quality ratings with upper categories three to four were common with observers and thus, ceiling effects occurred. The items WIT and STR in the plenary phase and APP in the working phase were particularly highly rated.

5.2 Confirmatory factor analyses and internal consistency

To provide validity evidence based on internal structure, we conducted *CFA* (*n* = 170) using the mean aggregated rating values of both raters (Table 3). Two models were tested according to our hypotheses: a one-factorial model with all items loading on one factor of ECM aspects and a second model with two theoretically assumed factors—one

Table 5 Rank correlations between knowledge and observed effective classroom management

		PCK	(1) CM (all)	(2.1) F1: Organization	(2.2) F2: Instruction	(3) Selected indicators
GPK	General Pedagogical Knowledge	0.501**	0.385*	0.292	0.334$^{\#}$	0.434*
PCK	Pedagogical Content Knowledge		0.103	0.002	0.098	0.089
CM (all)	Classroom management (all items)			0.961**	0.567**	0.960**
F1	Factor 1: Organization				0.366*	0.921**
F2	Factor 2: Instruction					0.519**

$^{\#}p<0.10$; $*p<0.05$; $**p<0.01$. Values indicate Spearman's rank correlation coefficient. Correlations between classroom management and knowledge: $n=30$. Correlations within classroom management scales: $n=36$

Table 6 Rank correlations between GPK subscales and observed effective classroom management

		GPK-Str	GPK-Ada	(1) CM (all)	(2.1) F1: Organization	(2.2) F2: Instruction	(3) Selected indicators
GPK-Mot	GPK subscale: motivation	0.159	−0.057	0.340$^{\#}$	0.333$^{\#}$	0.018	0.263
GPK-Str	GPK subscale: structuring		0.233	0.347$^{\#}$	0.258	0.373*	0.409*
GPK-Ada	GPK subscale: adaptivity			−0.223	−0.299	0.104	−0.190
CM (all)	Classroom management (all items)				0.961**	0.567**	0.960**
F1	Factor 1: organization					0.366*	0.921**
F2	Factor 2: instruction						0.519**

$^{\#}p<0.10$; $*p<0.05$; $**p<0.01$. Values indicate Spearman's rank correlation coefficient. Correlations between classroom management and knowledge: $n=30$. Correlations within classroom management scales: $n=36$

organizational and one instructional factor. Model 1 shows poor model fit, which is due to low factor loadings for GCL and APP. By contrast, model 2 has an acceptable fit with factor loadings ranging from reasonable to high.

The scales' internal consistency was examined using Cronbach's alpha (see Additional file 1: Appendix Tables S2 and S3). For each observational time point, working and plenary phases were analyzed separately with each scale consisting of eight items. They were also tested as one scale with 16 items. Internal consistency was sufficient for the 16-item-scale at all three measurement time points ($\alpha=0.69$-0.79). Upon analyzing each phase separately, internal consistency values were still acceptable but lower due to the lower number of items. Internal consistency was higher for the working phase ($\alpha=0.68$-0.86) than the plenary phase ($\alpha=0.50$-0.60). Regarding item discrimination, no item was systematically conspicuous regarding low discrimination at all measurement time points.

5.3 Correlations to teachers' professional knowledge

Table 4 shows means and standard deviations for the ECM scales and knowledge tests. M_{Points} represents the points scored in the test (GPK: Maximum = 60; PCK: Maximum = 29). Since Cronbach's alpha was at least acceptable for the different ECM scales investigated in the factors analyses, we included all scales when we investigated the relationship to teachers' professional knowledge.

The analyses in this section are based on scales which were constructed as follows— (1) Classroom management with all items (DIS, WIT, ETU, CRU, CRO, APP, GCL, STR); (2.1) Factor 1 (Organizational aspects: DIS, WIT, ETU, CRU, CRO, APP); and (2.2) Factor 2 (Instructional aspects: STR, GCL). As described above, given that ceiling effects of this measure might lead to underestimated correlations, generalizability study for each indicator was conducted to investigate which indicators were relevant at the teacher level (i.e., showed a sufficient proportion of variance between teachers). Using this explorative approach, we included another scale called "selected indicators" encompassing DIS, CRU, CRO, APP, and GCL (3).

As depicted in Table 5, GPK is related to ECM. This is especially true for the (1) classroom management scale including all items as well as the (3) scale containing selected indicators only. For these, the correlation with GPK is significant. By contrast, there is no significant relationship with PCK. Even though this pattern of findings corresponds to our hypothesis, it is important to note that only the difference between the correlations of (3) classroom management with selected items only and GPK versus PCK is significant ($z = 1.66$; $p = 0.048$). Given that the scale (3) was constructed on the basis of an exploratory analysis, this result, despite its plausibility, should be interpreted with caution.

When the GPK test is differentiated into a selection of reliable subscales, "structuring lessons" as a knowledge dimension shows significant relationships to observed ECM (Table 6). Whereas GPK of "motivating students" is significantly correlated on the 10% significance level, GPK of "adaptivity" does not significantly correlate with the observed ECM.

6 Conclusion and discussion

6.1 Main research findings

In this study, a newly developed standardized rating instrument for observing and measuring ECM in primary school—exemplified in early reading and writing instruction—was presented and tested for validity evidence. The generic instrument was applied in second-grade classrooms of 35 primary school teachers. Their instructional quality was assessed by trained raters. We investigated reliability and the dimensionality of the observer ratings and were able to show that ECM as a generic basic instructional quality dimension in German primary schools can be reliably and validly measured.

With respect to the factor structure, contrary to our expectations, the one-factor model showed a poor fit. Nevertheless, the two-factor model suggesting a theory-based differentiation into organizational and instructional dimensions showed an acceptable or even good fit. Internal consistency values were sufficient and higher for working phases as for plenary phases. Since in early reading and writing lessons, working phases are longer than plenary phases, rating might have been facilitated, thus strengthening ratings' reliability. Still, that every lesson contained this partial structure of having an introductory plenary phase followed by a working phase (even if there was substantial variation of the plenary phase's length and the working phase later on might be followed by another phase, such as discussion or transfer phase), might indicate that only little variation was found in the formal structure of the rated lessons. This kind of monotony of teaching methods in Germany primary schools could be critically regarded as a possible detriment of teaching quality.

In our second research question, we considered the validity evidence necessary for ECM based on relations to teachers' professional knowledge. In our sample, both GPK and PCK were positively related to the observed ECM, but when using inferential statistics, no significant correlations could be found between PCK and ECM, whereas, significant correlations existed between GPK and the ECM scale including all items as well as the scale with selected indicators. A differentiation of GPK into subscales with sufficient reliability (motivating students, structuring, adaptivity) showed that in particular GPK of structuring lessons correlated significantly with the observed ECM. Since structuring of lessons has long been related to ECM in the literature (e.g., [25, 49]), this finding can be interpreted as further evidence on validity of the novel ECM observation instrument. Moreover, motivating students might be seen being partly related to ECM as well [16], as significant correlations between GPK of student motivation and the observed ECM were just missed ($p < 0.10$).

6.2 Implications for measuring effective classroom management

The present study showed the development and validation of a new generic rating instrument being implemented in early primary school classrooms. The instrument is designated to be applicable to early reading and writing, but also to other subjects (e.g., mathematics) and grades, both at the primary school and beginning secondary school levels. However, the application of this instrument in these contexts should be investigated in future studies. The results of factor and correlational analyses suggest that rating values can be validly interpreted as teachers' ECM. The two-factor model provided evidence that ECM can better be depicted as a multifaceted construct and divided into organizational and instructional aspects, supporting theoretical assumptions (e.g., [24–26]. Correlations with teachers' GPK evidenced that the instrument covers generic facets of ECM [39, 43]. In contrast, a link to PCK as subject-specific teacher knowledge could not be found, which corresponds well with other studies [41, 44].

As shown by additional analyses using Generalizability Theory (see Additional file 1: Appendix), there was no main effect for a measurement point. However, as the interaction of teachers and measurement points are significant, one might conclude teachers differed in their ECM over the three observation time points. Possible explanations are that preparations might have been different, partial school closures could have played an important role, and due to restricted lessons, methods and instruction might have differed especially during the third observation. In general, stability of observational time points plays an important role when assessing teachers' instructional quality. Their performance may vary especially during the first two school years of primary school as they adapt to students' needs more fluidly than in secondary school, where routines and rules have already been established and transitions between phases are well-practiced [45].

6.3 Limitations and implications for future research

Several limitations to our instrument and study should be mentioned. On the methodological level, the exploratory approach to select indicators based on the generalizability studies should be highlighted. Also, the high number of tested correlations could be associated with an increased risk of alpha error. However, it should be emphasized that the pattern of correlations is consistent across the included ECM scales.

Concerning the instrument, ceiling effects showed that the definition of the four possible coding categories of the single item (see Additional file 1: Appendix, Table S1, as an example) could be more precise. The meaning of frequency indicators and terms such as "always", "rarely", and "sometimes" need to be explained more accurately in the rating manual. Frequency and level of, for example, disruptive classroom behavior should not be confounded [19]. Possibly, further items can be developed and added to the instrument, including negative item statements for the purpose of increasing variance.

The sample was probably a positive selection of teachers. Due to the outbreak of COVID-19 during the study, only few and very committed teachers participated. The instrument should thus be tested again based on a larger and more heterogeneous teacher sample. Its efficiency might be underestimated due to the highly selective sample. It can thus be assumed that a more heterogeneous sample would produce a rank order, which is more stable across measurement time points.

Although the lessons were observed by two raters, we could not specifically control for them as they were not systematically arranged into pairs. However, this is an important aspect as in our study, they had to observe lessons with deep structure and rate high-inferent items. Studies show that observer errors can constitute about 40% of variance [13]. Rater bias (e.g., halo and leniency effects) and errors could be methodically addressed by decision studies and should be considered in future research. Because of ceiling effects, in future studies teachers should be explicitly encouraged to teach as usual as possible, so that daily teaching can be observed. Although there is potential for further development of the instrument in terms of controlling for observer errors, the results of our study show that the rating instrument can assess instructional quality in terms of ECM, with the potential to be implemented in future studies accounting for other contexts, subjects, and grades in primary school.

Author contributions JK—conceptualization, investigation, writing—original draft, supervision, resources, project administration. NG—conceptualization, data curation, methodology, formal analysis, writing—review and editing. JW—data curation, methodology, formal analysis, writing—review and editing. GC—investigation, formal analysis, writing—review and editing. PH—validation, supervision, resources, project administration. CK—validation. TP—validation, supervision, resources, project administration. TW—validation, writing—review and editing. MB-M—validation, supervision. AS—validation, supervision. BT—validation, supervision.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was suppported by German Research Foundation (Deutsche Forschungsgemeinschaft, DFG, Grant KO3947/9–1).

Data availability The dataset generated during and analyzed during the current study are not publicly available due data protection obligation the authors agreed on when conducting their study, but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate For conducting the present study in Germany, an explicit ethics statement from a committee was not officially required. As researchers of an empirical project that was funded by German Research Foundation, we as authors comprehensively oblige to adhere to standards for good research practice.

Consent for publication Informed consent was obtained from all individual participants included in the study.

Competing interests There are no competing interests.

References

1. Doyle W. Ecological approaches to classroom management. In: Evertson CM, Weinstein CS, editors. Handbook of classroom management: Research, practice, and contemporary issues. Mahwah, NJ: Erlbaum; 2006. p. 97–125.
2. Kounin JS. Discipline and group management in classrooms. Oxford: Holt; 1970.
3. Emmer ET, Stough LM. Classroom management: a critical part of educational psychology, with implications for teacher education. Educ Psychol. 2001;36(2):103–12.
4. NBPTS [National Board for Professional Teaching Standards]. What teachers should know and be able to do. 2016
5. Evertson CM, Weinstein CS. Handbook of classroom management: research, practice, and contemporary issues. Mahwah, NJ: Erlbaum; 2006.
6. Emmer ET, Sabornie EJ. Introduction to the second edition. In: Evertson CM, Weinstein CS, editors. Handbook of classroom management: research, practice, and contemporary issues. 2nd ed. Mahwah: Erlbaum; 2015. p. 3–12.
7. Hattie J. Visible learning for teachers: maximizing impact on learning. London: Routledge; 2012.
8. Korpershoek H, Harms T, de Boer H, van Kuijk M, Doolaard S. A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. Rev Educ Res. 2016;86(3):643–80.
9. Seidel T, Shavelson RJ. Teaching effectiveness research in the past decade. Theory and research design in disentangling meta-analysis results. Rev Educ Res. 2007;77(4):454–99.
10. Wang MC, Haertel GD, Walberg HJ. Toward a knowledge base for school learning. Rev Educ Res. 1993;63(3):249–94.
11. Helmke A, Brühwiler C. Unterrichtsqualität. In: Rost DH, Sparfeldt JR, Buch SR, editors. Handwörterbuch pädagogische psychologie. 5th ed. Weinheim: Beltz; 2018. p. 860–9.
12. Praetorius A-K, Klieme E, Herbert B, Pinger P. Generic dimensions of teaching quality: the German framework of three basic dimensions. ZDM. 2018;50(3):407–26.
13. Praetorius A-K, Lenske G, Helmke A. Observer ratings of instructional quality: do they fulfill what they promise? Learn Instr. 2012;22(6):387–400. https://doi.org/10.1016/j.learninstruc.2012.03.002.
14. Strong M, Gargani J, Hacifazlioğlu Ö. Do we know a successful teacher when we see one? Experiments in the identification of effective teachers. J Teach Educ. 2011;62(4):367–82.
15. Fauth B, Decristan J, Rieser S, Klieme E, Büttner G. Student ratings of teaching quality in primary school. Dimensions and prediction of student outcomes. Learn Instr. 2014;29:1–9.
16. Kunter M, Baumert J, Köller O. Effective classroom management and the development of subject-related interest. Learn Instr. 2007;17(5):494–509.
17. Schlesinger L, Jentsch A, Kaiser G, König J, Blömeke S. Subject-specific characteristics of instructional quality in mathematics education. ZDM Math Educ. 2018. https://doi.org/10.1007/s11858-018-0917-5.
18. Pianta RC, Hamre BK. Conceptualization, measurement, and improvement of classroom processes: standardized observation can leverage capacity. Educ Res. 2009;38(2):109–19.
19. Lotz M, Gabriel K, Lipowsky F. Niedrig und hoch inferente Verfahren der Unterrichtsbeobachtung. Analysen zu deren gegenseitiger Validierung. Zeitschrift für Pädagogik. 2013;59(3):357–80.
20. Taut S, Rakoczy K. Observing instructional quality in the context of school evaluation. Learn Instr. 2016;46:45–60.
21. Gage NL, Needels MC. Process-product research on teaching: A review of criticisms. Elem Sch J. 1989;89(3):253–300.
22. Praetorius A-K, Vieluf S, Saß S, Bernholt A, Klieme E. The same in German as in English? Investigating the subject-specificity of teaching quality. Z Erzieh. 2015;19(1):191–209.
23. Hattie J. Visible learning: a synthesis of over 800 meta-analyses relating to achievement. London: Routledge; 2009.
24. Gilberts GH, Lignugaris-Kraft B. Classroom management and instruction competencies for preparing elementary and special education teachers. Teach Teach Educ. 1997;13(6):597–610.
25. Clausen M, Reusser K, Klieme E. Unterrichtsqualität auf der Basis hochinferenter Unterrichtsbeurteilungen. Ein Vergleich zwischen Deutschland und der deutschsprachigen Schweiz. Unterrichtswissenschaft. 2003;31(2):122–41.
26. Gettinger M, Kohler K. Process-outcome approaches to classroom management and effective teaching. In: Evertson CM, Weinstein CS, editors. Handbook of classroom management: Research, practice, and contemporary issues. Mahwah: Erlbaum; 2006. p. 73–96.
27. Carton A, Fruchart E. Sources of effects of the level of experience in primary school stress, coping strategies, emotional experience: teachers in France. Educ Rev. 2014;66(2):245–62.
28. Emmer ET, Gerwels MC. Classroom management in middle and high school classrooms. In: Evertson CM, Weinstein CS, editors. Handbook of classroom management: research, practice, and contemporary issues. Mahwah: Erlbaum; 2006. p. 407–37.
29. Marzano RJ, Marzano JS. The key to classroom management. Educ Leadersh. 2003;61(1):6–8.
30. Grossman P, McDonald M. Back to the future: directions for research in teaching and teacher education. Am Educ Res J. 2008;45(1):184–205.
31. Brophy J. Observational research on generic aspects of classroom teaching. In: Alexander PA, Winne PH, editors. Handbook of educational psychology. 2nd ed. Mahwah: Erlbaum; 2006. p. 755–80.
32. Ratcliff NJ, Jones CR, Costner RH, Savage-Davis E, Sheehan H, Hunt GH. Teacher classroom management behaviors and student time-on-task: implications for teacher education. Action Teach Educ. 2010;32(4):38–51. https://doi.org/10.1080/01626620.2010.549714.
33. Klieme E, Schümer G, Knoll S. Mathematikunterricht in der Sekundarstufe I. "Aufgabenkultur" und Unterrichtsgestaltung. In: Klieme E, Baumert J, editors. TIMSS – Impulse für Schule und Unterricht. BMBF: Bonn; 2001. p. 43–57.

34. Casabianca JM, Mccaffrey D, Gitomer D, Bell C, Hamre B, Pianta RC. Effect of observation mode on measures of secondary mathematics teaching. Educ Psychol Measur. 2013;73(5):757–83.
35. Schlesinger L, Jentsch A. Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. ZDM. 2016;48(1–2):29–40.
36. Kane MT. Validation as a pragmatic, scientific activity. J Educ Meas. 2013;50(1):115–22.
37. Bell CA, Gitomer DH, McCaffrey DF, Hamre BK, Pianta RC, Qi Y. An argument approach to observation protocol validity. Educ Assess. 2012;17(2–3):62–87. https://doi.org/10.1080/10627197.2012.715014.
38. Blömeke S, Gustafsson J-E, Shavelson RJ. Beyond dichotomies: competence viewed as a continuum. Zeitschrift für Psychologie. 2015;223(1):3–13.
39. König J, Kramer C. Teacher professional knowledge and classroom management: on the relation of general pedagogical knowledge (GPK) and classroom management expertise (CME). ZDM. 2016;48(1):139–51.
40. König J, Pflanzl B. Is teacher knowledge associated with performance? On the relationship between teachers' general pedagogical knowledge and instructional quality. Eur J Teach Educ. 2016;39(4):419–36.
41. Voss T, Kunter M, Seiz J, Hoehne V, Baumert J. Die Bedeutung des pädagogisch-psychologischen Wissens von angehenden Lehrkräften für die Unterrichtsqualität. Zeitschrift für Pädagogik. 2014;60:184–201.
42. König J, Blömeke S, Jentsch A, Schlesinger L, Felske C, Musekamp F, Kaiser G. The links between pedagogical competence, instructional quality, and mathematics achievement in the lower secondary classroom. Educ Stud Math. 2021;107:189–212.
43. Lenske G, Wagner W, Wirth J, Thillmann H, Cauet E, Liepertz S, Leutner D. Die Bedeutung des pädagogisch-psychologischen Wissens für die Qualität der Klassenführung und den Lernzuwachs der Schüler/innen im Physikunterricht. Z Erzieh. 2016;19(1):211–33.
44. Baumert J, Kunter M, Blum W, Brunner M, Voss T, Jordan A. Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. Am Educ Res J. 2010;47:133–80.
45. Gabriel-Busse K, Jentsch A, Lipowsky F. Prozess- und strukturorientierte Klassenführungsmaßnahmen von Lehrpersonen im Anfangsunterricht – Ergebnisse zur zeitlichen Stabilität von Beobachterratings innerhalb und zwischen 90-minütigen Unterrichtseinheiten. Z Bild. 2021. https://doi.org/10.1007/s35834-021-00325-3.
46. Lipowsky F, Rakoczy K, Pauli C, Drollinger-Vetter B, Klieme E, Reusser K. Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean theorem. Learn Instr. 2009;19:527–37.
47. Seidel T, Rimmele R, Prenzel M. Clarity and coherence of lesson goals as a scaffold for student learning. Learn Instr. 2005;15:539–56. https://doi.org/10.1016/j.learninstruc.2005.08.004.
48. König J, Krepf M, Bremerich-Vos A, Buchholtz C. Meeting cognitive demands of lesson planning: introducing the CODE-PLAN model to describe and analyze teachers' planning competence. Teach Educ Q. 2021;56(4):466–87.
49. Krepf M, König J. Structuring lessons as an aspect of preservice teachers' planning competence: a scaling-up analysis. Z Erzieh. 2022;25(4):917–46.
50. König J, Blömeke S, Paine L, Schmidt WH, Hsieh F-J. General pedagogical knowledge of future middle school teachers: on the complex ecology of teacher education in the United States, Germany, and Taiwan. J Teach Educ. 2011;62(2):188–201. https://doi.org/10.1177/0022487110388664.
51. König J, Hanke P, Glutsch N, Jäger-Biela D, Pohl T, Becker-Mrotzek M, Schabmann A, Waschewski T. Teachers' professional knowledge for teaching early literacy: Conceptualization, measurement, and validation. Educ Assess Eval Account. 2022;34:483–507. https://doi.org/10.1007/s11092-022-09393-z
52. Wu ML, Adams RJ, Wilson MR. ConQuest: Multi-aspect test software [computer program]. Camberwell, Australia: ACER; 1997.
53. Bond TG, Fox CM. Applying the Rasch Model. Fundamental measurement in the human sciences. 3rd ed. New York: Routledge; 2015.
54. Warm TA. Weighted likelihood estimation of ability in item response theory. Psychometrika. 1989;54(3):427–50.
55. American Educational Research Association (AERA) American Psychological Association (APA), & National Council on Measurement in Education (NCME). Standards for educational and psychological testing. 6th ed. Washington, DC: American Educational Research Association; 2014.
56. Meng X, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. Psychol Bull. 1992;111(1):172–5. https://doi.org/10.1037/0033-2909.111.1.172.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.