frontiers | Frontiers in Education

TYPE Brief Research Report
PUBLISHED 17 October 2024
DOI 10.3389/feduc.2024.1437996

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Catarina Andersson,
Umeå University, Sweden
Wei Shin Leong,
Ministry of Education, Singapore
Seyeoung Chun,
Chungnam National University, Republic of Korea

*CORRESPONDENCE
Kathrin Kohake
✉ Kathrin.kohake@uni-muenster.de

RECEIVED 24 May 2024
ACCEPTED 17 September 2024
PUBLISHED 17 October 2024

CITATION
Kohake K (2024) Systematic observation to measure teaching quality in different contexts: insights from science lessons, physical education lessons, and sports training using the classroom assessment scoring system.
Front. Educ. 9:1437996.
doi: 10.3389/feduc.2024.1437996

Systematic observation to measure teaching quality in different contexts: insights from science lessons, physical education lessons, and sports training using the classroom assessment scoring system

Kathrin Kohake*

Department of Physical Education and Teaching Research, Institute of Sport and Exercise Sciences, University of Münster, Münster, Germany

There is substantial evidence that the quality of classroom interactions is crucial for students' academic and social development. Both subject-specific and generic systematic observation instruments are widely used to assess these interactions. A notable example of a generic observation protocol based on US classroom research is the Classroom Assessment Scoring System (CLASS). This study aims to evaluate the applicability of the CLASS framework for assessing teaching quality in various contexts in Germany, specifically in science classes, physical education classes, and sports training at the elementary school level. In total, 110 video-recorded observation cycles, each lasting 15–20 min, were double-coded by two independent observers. Assessments were conducted across 10 dimensions using a 7-point scale. Results indicate average scores for the dimensions within the Emotional Support domain in the medium to high range, within the Classroom Organization domain in the high range, and within the Instructional Support domain in the low to medium range. CLASS scores varied considerably across different settings, effectively distinguishing between the observed teachers and coaches. The Percentage-Within-One agreement values and Intra-Class-Correlations demonstrate good interrater reliability across all settings. These findings highlight the robustness of the CLASS framework and its transferability to various educational contexts in Germany. This adaptability facilitates future studies on predictive validity and enables cross-country comparisons.

KEYWORDS

systematic observation, interrater agreement, interaction quality, CLASS, interdisciplinarity

1 Introduction

The question of which criteria best characterize "good teaching" has been central to general and subject-specific empirical teaching researchers for years. Researchers largely agree that among other aspects the quality of interactions in classrooms matters and is critical for students' learning and development (Downer et al., 2010; Hattie, 2009). Internationally,

systematic observation instruments are most important in assessing teaching quality. One framework for this purpose is the Teaching-Through-Interactions-Framework (Hamre et al., 2013), operationalized in the CLASS tool (Pianta et al., 2008). It builds on theoretical and empirical work in education and psychology (Hamre et al., 2007). Although the CLASS was developed in US classrooms, a recent meta-analysis showed its international applicability (Hofkens et al., 2023). The CLASS tool has been used to measure teacher-student interactions during varying learning activities with a stronger focus on math and language arts, and less on science lessons. Physical education lessons were typically coded as non-observable time. In Germany, physical education is a compulsory subject throughout all grades and should therefore also be included in pedagogical analysis. There is an ongoing scientific debate as to whether generic or subject-specific quality criteria in physical education best describe effective teaching strategies (Herrmann and Gerlach, 2020; Richartz and Kohake, 2021). Despite physical education classes, half of all children and adolescents in Germany take part in extracurricular sporting activities, which therefore forms another important field of educational research. The usefulness of CLASS for sport-specific contexts has been little studied to date (Kohake et al., 2022; Maier, 2023).

The current study aims to examine the usefulness of the CLASS lens as a generic observation instrument for the assessment of teaching quality in science classes and physical education classes, and coaching quality in sports training with learners of elementary school age in Germany. For this purpose, we conduct CLASS ratings and analyze the data regarding reliability and differentiability within the three settings. Due to the exploratory design, the data collection is limited to small samples, so conclusions about the level of effective teaching/coaching as well as mean comparisons would not be advised due to a lack of representativeness. Nevertheless, the data provides valuable insight into the use of the CLASS in different settings in Germany and adds to the scientific debate regarding its reliable use and its possibility of assessing teaching quality in a differentiated way.

2 Theoretical background

2.1 Quality characteristics of teaching

When stating that the quality classroom experiences matters, high quality is often referred to as the effectiveness with which students achieve developmental and academic learning goals (Praetorius and Gräsel, 2021). While teacher quality refers to certain characteristics of the teacher (e.g., experience, education, background, beliefs, knowledge), teaching quality refers to the behavioral element and thus the quality of the teacher's instructions. Research shows that classroom interactions are relevant to student success (Hofkens et al., 2023). Today, there is an increasing focus on the so-called "three basic dimensions" of teaching quality across subjects (Klieme et al., 2009): Cognitive Activation, Classroom Management, and Supportive Climate. The triangulated structure is well aligned with international research: Hamre et al. (2007) were able to show that a 3-domain structure of interaction quality including Emotional Support (Supportive Climate), Classroom Organization (Classroom Management), and Instructional Support (Cognitive Activation) was generalizable from preschool to fifth grade.

According to a recent review by Herrmann and Gerlach (2020), the increasing focus on the three basic dimensions/domains can also be seen in sports didactic research. According to their findings, generic rather than subject-specific quality criteria play a dominant role in physical education research (Herrmann and Gerlach, 2020). Following the argumentation of Richartz and Kohake (2021), socio-emotional support and classroom management are hardly subject specific. The theoretical foundations of these dimensions, such as attachment theory (Bowlby, 1991) and self-determination theory (Ryan and Deci, 2017), as well as clarity of rules, discipline, and use of time (Evertson and Emmer, 2000; Kounin, 1976), are *per se* already non-subject-specific criteria. In other words, the underlying concepts for effective teaching at a deeper level remain the same across subjects. Against the background of research on effective coaching, Richartz et al. (2021) show that these criteria can also be applied to sports training. In agreement with Herrmann and Gerlach, subject-specific adaptations in the area of instructional support/cognitive activation domain are necessary for sport contexts, although the authors argue differently about the content of these adjustments. The greatest need for change arises from the fact that the deeper concept of higher-order thinking skills is only relevant to some learning processes in sport. Movement learning and tactical learning require additional concepts that do not apply to other subjects (for a detailed discussion see Richartz and Kohake, 2021). This needs to be taken into account when using the CLASS.

2.2 Systematic observation in educational research

The primary aim of systematic observation is usually to examine the relationships between teaching and learning (research purpose) (Charalambous and Praetorius, 2022). Other purposes include improving teaching by providing feedback to teachers based on the observations (formative evaluation) or even making decisions about hiring, promotion, and dismissal (summative evaluation) (Charalambous and Praetorius, 2022). "The real potential of classroom observations is their usefulness for diagnosis and development of instructional practice" (Kane and Staiger, 2012, p. 15). However, deep structures of teaching (Kunter and Ewald, 2016) are difficult to measure because they are not directly observable.

The need for subject-specific vs. transferability of generic quality measures is widely debated. Comparing different approaches is difficult due to the lack of categorical clarity: despite differences in the hierarchical order (domains/dimensions/subdimensions) of frameworks, similar names of quality criteria often contain different concepts and ideas (Charalambous and Praetorius, 2022). The advantages of generic measures therefore include a common language and structure across subjects, which allows comparisons across subjects, settings, and countries, and increases opportunities for interdisciplinarity and collaboration. In contrast, it can be argued that generic instruments lack the necessary subject-specificity, create measurement problems when it comes to context-specific operationalization, and make it difficult to provide specific feedback to teachers (Charalambous and Praetorius, 2022). However, rather than comparing the two approaches, as argued earlier, there are areas of teaching quality that can be assumed to be generic anyway, and therefore particularly well captured by generic instruments that have

already been proven to work across the board. The CLASS, as one such a generic instrument, has already been used successfully in different teaching-learning scenarios, subjects, and countries (Hofkens et al., 2023).

Furthermore, the utilization of generic observation instruments is also justified by the absence of any identified disadvantages in terms of predictive validity when compared to subject-specific instruments. The prognostic validity of various subject-specific and generic frameworks was the subject of extensive examination in the Measures-of-Effective-Teaching study. A total of 1,300 teachers from the USA, teaching Mathematics and English in grades 4 to 8, were observed using three subject-specific and two generic instruments (FFT; CLASS; PLATO; MQI, UTOP). The authors demonstrated correlations between observations and student performance on achievement tests (including mathematics, reading, and English) for all of the five instruments. In conclusion, the authors posit that: "Currently, we see little reason to choose different classroom observation instruments in math and ELA (English Language Arts). The generic instruments, designed for use across subjects, appear to be just as correlated to student achievement gains in math and ELA as the subject-specific instruments overall "(Kane and Staiger, 2012, p. 57). The benefits of observations with a common language and structure for cross-context comparisons can therefore be leveraged effetively with the help of CLASS.

2.3 Classroom Assessment Scoring System (CLASS)

Teachers' and students' interactions in the classroom can be measured by the Classroom Assessment Scoring System (CLASS), which is founded on the developmental theory of learning (Pianta and Hamre, 2009). The CLASS is a subject-independent instrument based on extensive research on effective teaching practices in teaching-learning scenarios in the USA. However, the developers of the CLASS have not yet employed it in physical education lessons. Different specifications of the CLASS for different age groups exist (Infant, Toddler, Pre-K-3, Upper Elementary, Secondary). Across age groups, interaction quality is assessed in three overarching domains: Emotional Support, Classroom Organization, and Instructional Support.

The **Emotional Support** domain focuses on the teacher's ability to create a positive and supportive learning environment. The theoretical and empirical basis for this approach is rooted in attachment theory (Ainsworth, 2003; Bowlby, 1991) as well as self-determination theory (Ryan and Deci, 2017). This domain encompasses the overall atmosphere of the classroom, including the teacher's ability to foster a supportive, respectful, and warm environment for students. **Classroom Organization** refers to the structure and management of the learning environment. The concept is grounded in research on self-regulation and classroom management (Evertson and Emmer, 2000; Kounin, 1976). The domain assesses the efficacy with which teachers establish and maintain routines, procedures, and expectations, with the objective of optimizing instructional time and minimizing disruptions. **Instructional Support** assesses the extent to which teachers are able to foster cognitive and academic engagement among their students. This domain is concerned with the quality of instruction, encompassing the clarity of explanations, the depth of conceptual development, and the effectiveness of feedback provided to students.

In regard to the structural validity of the CLASS, a recent meta-analysis based on 26 studies found support for a two-factor as well as a three-factor structure for the K-3 level (Li et al., 2020). However, individual studies often report a superior fit of the three-factor model (e.g., Hafen et al., 2015; Kohake et al., 2022; Pakarinen et al., 2010), which is referred to as the Teaching-Through-Interactions-Framework (TTIF; Hamre et al., 2013). In sum, the results consistently indicate that the interpretation of a global CLASS score would be an inappropriate interpretation of the data and that the results should be considered at either the domain or dimension level.

Studies outside the US have linked CLASS scores to a number of key variables including students' empathy, disruptive or problematic behavior, self-regulation, attention and impulse control, executive functions, learning motivation, and engagement (for a comprehensive overview see Hofkens et al., 2023). In studies conducted in the United States, results indicate that the dimensions of Emotional Support and Classroom Organization are predictive of self-regulatory and social outcomes, while the dimension of Instructional Support is related to language and literacy skills in pre-kindergarten through third-grade classrooms (Pianta et al., 2014). Given the dearth of research on the CLASS framework in the context of sports, its predictive validity remains to be demonstrated.

This study employs an interdisciplinary approach to analyze teaching quality, with a particular focus on the three well-funded areas of Emotional Support, Classroom Organization, and Instructional Support. Systematic observation represents a valuable methodology for assessment and serves as an initial point of departure for further development. To this end, the three areas of science classes, physical education, and extracurricular sports are examined in more detail. The objective is to ascertain the reliability of the CLASS in these three contexts and its suitability for differentiating the quality of interaction between different teachers. This may serve as a foundation for future studies on predictive validity.

3 Methods

3.1 Participants

The study included convenient samples from three different projects, where systematic observations were conducted by the same research team: $N = 24$ teachers ($M = 42{,}2$ years; $SD = 7{,}4$; 88% female) of science classes in grades 3 and 4, $N = 5$ physical education teachers ($M = 37{,}4$ years; $SD = 9{,}8$; 20% female) of students in grades 3 to 6 (Maier, 2023), and $N = 26$ sports coaches ($M = 38{,}0$ years; $SD = 12{,}7$; 54% female) of gymnastics (50%), rhythmic gymnastics (15%), judo (15%), and handball (20%) with children at the age of 8–12 (which corresponds to approx. 3rd to 6th grade). All participants as well as the children's parents voluntarily consented to participate in the study.

3.2 Measures

The Classroom Assessment Scoring System K-3 (CLASS; Pianta et al., 2008) was used to observe and evaluate the quality of pedagogical interactions between teachers and students/coaches and

athletes. The CLASS K-3 employs a hierarchical structure comprising three overarching domains, which are themselves subdivided into dimensions, indicators, and behavioral markers. The CLASS is designed to facilitate in-depth evaluations of teacher-student interactions through the use of a high-inference measurement approach. The assessments are assigned on a 7-point scale to 10 dimensions: (a) Positive Climate, (b) Negative Climate, (c) Teacher Sensitivity, (d) Regard for Student Perspectives, (e) Behavior Management, (f) Productivity, (g) Instructional Learning Formats, (h) Concept Development, (i) Quality of Feedback, and (j) Language Modeling (Pianta et al., 2008; see Table 1). Low scores indicate low interaction quality, whereas high scores indicate a high level of quality, with the exception of the Negative Climate dimension, where low values are desirable. A manual provides detailed descriptions of low-, medium-, and high-quality interactions for each indicator. Raters participated in a two-day training session for observers, during which they were instructed on the dimensions of the CLASS and the methodology for observing them using short video clips. With the assistance of observation sheets and an observation manual, longer video sequences were subsequently coded, and the results were compared with written master-code justifications. Following the training, the raters completed a reliability test comprising five video sequences to show a minimum of 80% consistency with the master codes. As the reliability test included video footage from classrooms in the United States, further calibration sessions were conducted with video examples from German classrooms as well as German sports training sessions before the final ratings were assigned. The feasibility of applying the CLASS to extracurricular sporting contexts was previously demonstrated in an earlier study (Kohake et al., 2022). In accordance with the methodology employed in the aforementioned study, the dimensions of Concept Development and Language Modeling were excluded from the analysis of sports training and physical education classes. The Concept Development dimension is based on the concept of higher-order thinking skills. In contrast to other academic disciplines, the focus in physical education and sports training is on motor and tactical learning processes. The extant research demonstrates that higher-order thinking is an ineffective approach to these learning processes (Hossner and Künzell, 2022). Therefore, measuring teaching quality based on this dimension is an unjustified approach for sports contexts. Similarly, the Language Modeling dimension, which is primarily concerned with language development through intensive classroom conversations and discussions, is not a focus in sports contexts. In order to adequately address the subject-specific characteristics of physical education and sports training, these dimensions were excluded. More detailed explanations can also be found elsewhere (Kohake et al., 2022; Richartz and Kohake, 2021).

3.3 Procedure

One lesson/training session of each participant was recorded from which two 15–20-min sequences were coded with the CLASS. This resulted in a total of 110 observation cycles (2 cycles for each of the 55 participants). The coding was randomly assigned to two of the three potential raters. All segments were subject to independent double-coding in the 8/10 dimensions.

3.4 Analysis

The analyses were conducted using R. There was no missing data that could have affected the results. Descriptive statistics for the subsamples included ranges of variation, mean scores, and standard deviations for each CLASS dimension. The scores assigned by the independent raters were averaged for the observation cycles and for the teachers and coaches. To assess interrater reliability, both Percentage-Within-One (PWO), and Intra-Class-Correlations (ICC; two-way mixed effect with unadjusted estimates) were calculated. PWO-values above 0.8 were considered good (Pianta et al., 2008). An ICC greater than 0.7 was deemed acceptable, above 0.8 was good. The PWO is a relatively broad measure; therefore, the ICCs can provide further insight and critical information. It is important to note, however, the ICC is heavily dependent on sample size and variance. Consequently, lower values can be expected with the small sample size in this study.

4 Results

4.1 Descriptive statistics

Individual CLASS scores per teacher and coach widely range between dimensions from 1.3 (Quality of Feedback) to 7 (Positive Climate, Teacher Sensitivity, Behavior Management, Productivity). Across settings, the greatest range of scores was observed for Quality of Feedback (min = 1.25; max = 5.25) followed by Regard for Student Perspectives (min = 2.50; max = 5.75) and Productivity (min = 3.75; max = 7.00). The smallest range of variation was observed for the Negative Climate dimension (min = 1; max = 2.75). The overall mean scores for teachers' and coaches' per lesson/session ranged between $M = 1.18$ (Negative Climate) and 6.28 (Behavior Management).

When the three contexts are considered separately, the greatest ranges for CLASS scores (greater than or equal to 3 points) were observed for sports coaches in the dimensions of Positive Climate, Quality of Feedback, and Behavior Management, followed by physical education teachers in Regard for Student Perspectives and Productivity, and science teachers in Positive Climate. However, the relatively small and disparate sample sizes for the subsets likely affect these results.

The highest scores within settings were observed in the Classroom Organization domain, with $M = 6.44$ in Productivity for science teachers, $M = 6.15$ for Behavior Management for Physical Educational teachers, and $M = 6.22$ for Behavior Management for sports coaches. With the exception of the Negative Climate dimension, in which mean scores are below 1.3 for all settings, teachers and coaches scored lowest in Quality of Feedback ($M_{science} = 3.00$; $M_{PE} = 2.25$; $M_{training} = 3.64$). For more detailed results see Figure 1.

4.2 Rater reliability

The overall mean Percentage-Within-One (PWO) Agreement for the two independent raters was 94%. Upon examination of the three settings separately, it was found that PWO consistently exceeded 90%, with rates of 97% for science classes, 91% for physical education classes, and 91% for sports coaching. A more detailed analysis of the individual dimensions also reveals values that consistently exceed

TABLE 1 Domains and dimensions of the classroom assessment scoring system K-3 [see also Downer et al., 2024; for more details: Pianta et al., 2008].

Domain	Dimension	Description
Emotional Support	Positive Climate (PC)	Evaluates the overall warmth and positivity of the classroom environment fostering a sense of safety and belonging for students. Encompasses respect and sensitivity toward students' emotions and the degree to which they show enthusiasm and enjoyment.
	Negative Climate (NC)	Focuses on the occurrence and management of negative behaviors such as hostility, irritability, or frustration within the classroom. Considers the frequency, quality, and intensity of expressed negativity by teachers and peers.
	Teacher Sensitivity (TS)	Assesses the teachers' awareness and responsiveness to students' individual academic and emotional needs and interests. Teachers who score high in this dimension demonstrate empathy, understanding, and attentiveness to students' cues, adjusting their interactions accordingly to support each student's learning and well-being.
	Regard for Student Perspectives (RSP)	Encompasses how teachers acknowledge, value, and incorporate students' perspectives, ideas, and contributions into classroom activities and discussions. Teachers scoring high in this dimension show flexibility and encourage students' autonomy.
Classroom Organization	Behavior Management (BM)	Assesses the teachers' ability to establish and maintain a structured classroom environment conducive to learning. It evaluates the effectiveness of behavior management strategies promoting positive student behavior and minimizing disruptions by presenting clear behavioral expectations
	Productivity (PD)	Measures the efficiency and effectiveness of instructional time utilization within the classroom. It evaluates the extent to which teachers maximize learning opportunities. Not related to the quality of instructions.
	Instructional Learning Formats (ILF)	Focuses on the variety of instructional approaches and formats used by the teacher to facilitate student learning. The extent to which teachers provide interesting activities, groupings, and materials to promote active engagement and participation among students.
Instructional Support	Concept Development (CD)	Evaluates the teacher's efforts to foster higher-order thinking skills versus focusing on rote and fact-based learning. Assesses the depth of teachers' explanations, demonstrations, and discussions aimed at promoting conceptual understanding and critical thinking.
	Quality of Feedback (QF)	Captures teachers' provision of feedback regarding its specificity, accuracy, and usefulness for expanding students' learning and understanding. It evaluates the extent to which feedback is timely, informative, and tailored to individual student needs, promoting reflection, learning, and growth.
	Language Modeling (LM)	Considers the quality and amount of teachers' use and facilitation of rich and varied language during individual, small-group, and large-group interactions to support students' language development and academic discourse. It includes self and parallel talk, open-ended questions, repetition, expansion, and the use of advanced language.

90%. Nevertheless, when examining the various dimensions within the different settings, the PWO for the dimensions Regard for Student Perspectives (85%), Instructional Learning Formats (88%), and Quality of Feedback (88%) are slightly lower but still well above the targeted threshold of 80% (see Table 2). Only the PWO for the dimension Teacher Sensitivity of physical education teachers falls below the acceptable range (60%).

The overall Intra-Class-Correlation for the two independent observers was 0.90. The ICC for the eight dimensions across groups exhibited variability, with values ranging from 0.46 (Instructional Learning Formats) to 0.73 (Productivity). Considering the three groups separately, ICCs varied between 0.45 (ILF) and 0.73 (Negative Climate) for science classes, between 0.44 (Teacher Sensitivity) and 0.94 (Behavior Management) for physical education classes, and between 0.18 (Instructional Learning Formats) and 0.74 (Positive Climate) for sports training (see Table 2). Consequently, some values fall below the acceptable range. In particular, the ICCs for Negative Climate, Teacher Sensitivity, Regard for Student Perspectives, and Instructional Learning Formats are less robust than would be desirable.

5 Discussion

A substantial body of research underscores the pivotal role of interactions on student achievement. However, the majority of research has focused on English Language Arts and Mathematics lessons, with comparatively less attention directed toward science lessons and other subjects. In the present study, the generic quality criteria of good teaching in different settings were examined through the Teaching-Through-Interactions Framework, operationalized in the CLASS tool (Pianta et al., 2008). CLASS observations were conducted in three different contexts with similar age groups: science classes, physical education classes, and sports training. While a substantial corpus of data on CLASS exists, primarily derived from the preschool domain, its systematic application to sports has thus far been limited to our prior research endeavors. One of the primary advantages of the CLASS lies in its generic nature, which allows for a unified language that enables the comparison of the implementation of deep pedagogical concepts (as opposed to surface features of teaching) construct equivalent across diverse settings.

As the present study employs a convenience sampling strategy, resulting in a relatively small sample size, particularly in the context of physical education, the descriptive data should be interpreted with great caution. Notwithstanding these constraints, the data demonstrate a notable degree of variability within the observed dimensions, with values spanning a range of 1.3 to 7 points across the 7-point rating scale. It is noteworthy that the participants voluntarily consented to have their teaching/training recorded and evaluated using the CLASS dimensions, which may explain the scarcity of observations falling within the "low range." Nevertheless, it could be anticipated that a greater degree of homogeneity would be evident in a sample of teachers who had volunteered to be observed. But despite the

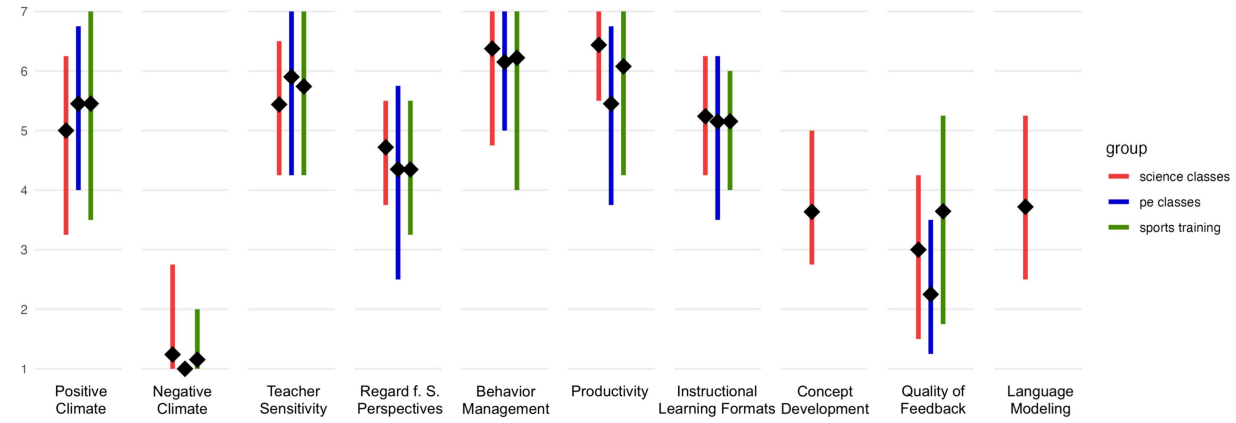

FIGURE 1
CLASS-mean scores and ranges for three subsamples.

TABLE 2 Interrater-reliability for CLASS observations.

	ICC				PWO			
	Overall	Science classes	PE classes	Sports training	Overall	Science classes	PE classes	Sports training
PC	0.71	0.66	0.73	0.74	91.8	91.7	90.0	92.3
NC	0.55	0.73	NA	0.27	98.2	100	100	96.2
TS	0.58	0.69	0.44	0.52	91.8	97.9	60.0	92.3
RSP	0.59	0.67	0.76	0.46	91.8	100	90.0	84.6
BM	0.68	0.72	0.68	0.65	95.5	97.9	90.0	94.2
PD	0.73	0.56	0.94	0.68	95.5	95.8	100	94.2
ILF	0.46	0.45	0.85	0.18	94.5	100	100	88.5
LM	NA	0.80	NA	NA	NA	97.9	NA	NA
QF	0.69	0.69	0.73	0.61	90.9	91.7	100	88.5
CD	NA	0.70	NA	NA	NA	97.9	NA	NA

PC, positive climate; NC, negative climate; TS, teacher sensitivity; RSP, regard for student perspectives; BM, behavior management; PD, productivity; ILF, instructional learning formats; CD, concept development; QF, quality of feedback; LM, language modeling.

restricted number of observations, it was possible to distinguish between different instructors effectiveley using the CLASS instrument. Notably, substantial variations were observed both within and between contexts in the dimension of Quality of Feedback. Further research should investigate these discrepancies in a larger sample and, if applicable, examine the factors that contribute to these variations in greater depth.

It is not possible to make statistically validated comparisons of means with other studies that have used the CLASS tool due to the limited size and lack of representativeness of the current sample. It would be prudent to refrain from drawing firm conclusions regarding the efficacy of teaching and coaching based on the presented results. Nevertheless, the observed tendencies may offer preliminary insights and indicate potential avenues for further research and improvement. The overall trends indicate higher mean values, particularly in the domains of Classroom Organization (Productivity and Behavior Management), and Instructional Learning Formats, in comparison to a summary of over 4,000 American Pre-K-6 classrooms (Hamre et al., 2013). The reported values from Finnish studies in Mathematics and Language Arts, although based on slightly older students, also demonstrate lower scores in the domain of Classroom Organization but similar scores in Emotional Support and Instructional Support (Pöysä et al., 2019; Virtanen et al., 2018). In addition to the selective sampling employed in the present study, which may have contributed to the overall higher values observed, potential differences in comparison to American classrooms could also be influenced by subject-specific and country-specific factors. It is important to note that our sample includes not only "regular" class sizes of 20–30 children but also sports groups, where significantly smaller groups are common, such as in gymnastics. Such factors may influence the implementation of behavior management strategies and the efficient use of time. In a Swiss context, which shares cultural and educational similarities with Germany, remarkably high values (M = 6.39) have been reported in the domain of Classroom Organization for a comparable age group (Gasser et al., 2018). The Swiss study did not provide details regarding the specific content of the lesson. It can be posited that German educational contexts may place a stronger emphasis on structured procedures to facilitate efficient time

utilization. If the finding of higher values in the Instructional Learning Formats dimension could be replicated in larger samples, this could indicate that the potential of hands-on activities may be more effectively utilized by teachers in highly applied contexts such as science and sports.

As has been demonstrated in previous studies, the scores obtained in the Instructional Support domain are relatively low. Our findings indicate that there is considerable scope for improvement in the quality of feedback provided by the five teachers in physical education lessons, which exhibited the lowest scores across all contexts. One potential explanation for this finding is that some teachers in our sample may not possess a strong instructional ethos or ambition. Rather than viewing physical education lessons as an instructional opportunity, they may see them as a means of balancing the school day and getting children moving. It is important to examine whether this finding is due to the selectivity of the sample or whether it is a subject-specific phenomenon. To do so, larger samples must be examined.

The present study's examination of inter-rater agreement in PWO was notably robust, which is not a common occurrence, given the tendency for observer judgments to diverge (Bell et al., 2015). Although the sample size was limited, all 110 included sequences were fully double-coded, thereby ensuring the dependability of any statements regarding the reliability of the entire sample. The average percent-within-one-agreement (PWO) values exceeded 90% across all three contexts examined, indicating a high level of interrater reliability. In their comprehensive review of studies conducted in the United States, Hamre et al. (2013) reported PWO values ranging from 71 to 91% per study. Similarly, other studies have reported PWO values ranging from 71 to 97% (e.g., in Finland: Virtanen et al., 2018).

As anticipated, the ICC values exhibit some degree of variation. While the overall ICC across all dimensions is rated as very good at 0.9, the values for individual dimensions within specific contexts are notably lower. In this study, the values ranged from 0.45 to 0.73. When considered independently of one another, the three settings exhibit lower values in the Instructional Learning Format and the Negative Climate dimension. Virtanen et al. (2018) similarly documented a considerable range of ICC values, spanning from 0.25 to 0.75 across observed dimensions. This variability is to be expected, given that ICCs are influenced by both sample size and variance. The reliability of ICC values increases with larger sample sizes, estimating ICCs more precisely and less susceptible to random variations. Additionally, low ICC values can occur when the variability within groups is high compared to the variability between groups—a phenomenon likely applicable for example to the Negative Climate dimension. Nevertheless, future research should consider ways to enhance interrater reliability especially in the Instructional Learning Format dimension in sports contexts.

The favorable outcomes pertaining to rater reliability observed in the present sample can be attributed to extensive observation training that was provided. In addition to the two-day CLASS training program, which was followed by reliability testing, additional calibration sessions were conducted using video material from the three contexts under consideration. Based on this foundation and strict adherence to the CLASS Manual, it seems reasonable to conclude that reliable assessments are achievable. Therefore, although the sample size is limited, the present data substantiate the conclusion that the reliable utilization of CLASS across disparate contexts is feasible when raters have undergone comprehensive training. The exploratory data presented here provide an important foundation for future research projects seeking to utilize CLASS across various contexts.

The depth concepts delineated in CLASS are highly transferable to disparate contexts, as background constructs at the depth level are not subject-specific. This enables prospective cross-country comparisons, as previously indicated by Hofkens et al. (2023). Additionally, in Germany, where studying and teaching at least two subjects is common, generic teaching competencies that are applicable across subjects are of paramount importance in the context of teacher education. Furthermore, research indicates that training in general teaching methods is more effective in promoting ambitious teaching practices than training in content-specific instructional strategies (Youngs et al., 2022). Accordingly, a general instrument such as the CLASS can offer valuable insights into feedback and learning processes across subjects, particularly for students engaged in teacher education programs.

5.1 Limitations

The presented data is limited by the absence of representative samples, which constraints the extent to which the findings can be generalized. The results provide preliminary indications of the utility of the methodological approach, but further investigation in larger samples is necessary.

Furthermore, we elected to exclude the Concept Development and Language Modeling dimensions from our analysis, which constraints the scope for comparisons with earlier studies with respect to the Instructional Support domain. Consequently, our analysis was primarily focused on dimension scores, rather than domain scores.

6 Conclusion

In conclusion, the findings of this study illustrate the adaptability of the generic CLASS observation tool, indicating preliminary evidence of its utility in diverse educational contexts with respect to differentiability and observer agreement. Moreover, the potential strengths and weaknesses identified offer valuable insights for future studies that can inform teacher training and further education initiatives.

Nevertheless, the integration of generic observations with subject-specific dimensions appears to be a particularly fruitful approach in the domain of "Instructional Support." This adaptation is crucial for the effective assessment of physical education and training, given the distinctive characteristics of movement learning and game tactical learning. The introduction of an additional dimension, which focuses on the quality of instructions in sports, has the potential to enhance the assessment of movement-based instruction. This dimension could encompass elements such as the perception of the effects of a movement, grading of difficulty, and the provision of sufficient and variable practice opportunities (Richartz and Kohake, 2023). Notwithstanding the aforementioned potential adaptations, comparisons between subjects in the Emotional Support and Classroom Organization domains would remain feasible. This approach allows for the advantages of generic and subject-specific approaches to be combined.

Looking ahead, in order to gain a deeper understanding of the CLASS scores, it would be beneficial to investigate their stability over different lessons and the school day. Within this, it would also be of interest to observe the same teacher teaching different subjects, as is common practice in Germany, where teachers typically teach at least two subjects. This approach could elucidate cross-disciplinary strengths and weaknesses, thereby furnishing invaluable insights for teacher education and professional development programs.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Local Ethics Committee of the Faculty of Psychology and Human Movement Science, University of Hamburg. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

KK: Writing – review & editing, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

I would like to thank Alfred Richartz and Jessica Maier for their substantial contribution to the video recordings and CLASS ratings. I would also like to thank the ProSach team of the IQB for providing the video recordings from science classes. Further thanks to Bob Pianta for his advice and support during my stay at UVA.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ainsworth, M. D. S. (2003). "Feinfühligkeit versus Unfeinfühligkeit gegenüber den Mitteilungen des Babys (1974)" in *Bindung und menschliche Entwicklung. John Bowlby, Mary Ainsworth und die Grundlagen der Bindungungstheorie*. eds. K. E. Grossmann and K. Grossmann (Stuttgart: Klett-Cotta), 414–421.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., Mccaffrey, D. F., Gitomer, D. H., et al. (2015). "Improving observational score quality" in *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. eds. T. J. Kane, K. A. Kerr and R. C. Pianta (San Francisco: John Wiley & Sons, Inc.), 50–97.

Bowlby, J. (1991). "Ethologisches Licht auf psychoanalytische Probleme" in *Bindung und menschliche Entwicklung: John Bowlby, Mary Ainsworth und die Grundlagen der Bindungstheorie*. eds. K. E. Grossmann and K. Grossmann (Stuttgart: Klett-Cotta), 55–69.

Charalambous, C. Y., and Praetorius, A. K. (2022). Synthesizing collaborative reflections on classroom observation frameworks and reflecting on the necessity of synthesized frameworks. *Stud. Educ. Eval.* 75:101202. doi: 10.1016/j.stueduc.2022.101202

Downer, J. T., Doyle, N. B., Pianta, R. C., Burchinal, M., Field, S., Hamre, B. K., et al. (2024). Coaching and coursework focused on teacher–child interactions during language/literacy instruction: effects on teacher outcomes and Children's classroom engagement. *Early Educ. Dev.* 35, 1032–1062. doi: 10.1080/10409289.2024.2303604

Downer, J. T., Sabol, T. J., and Hamre, B. (2010). Teacher-child interactions in the classroom: toward a theory of within- and cross-domain links to Children's developmental outcomes. *Early Educ. Dev.* 21, 699–6723. doi: 10.1080/10409289.2010.497453

Evertson, C. M., and Emmer, E. T. (2000). *Classroom Management for Elementary Teachers*. London: Pearson.

Gasser, L., Grütter, J., Buholzer, A., and Wettstein, A. (2018). Emotionally supportive classroom interactions and students' perceptions of their teachers as caring and just. *Learning Instruction* 54, 82–92. doi: 10.1016/j.learninstruc.2017.08.003

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., and Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: revisiting the factor structure and practical application of the classroom assessment scoring system-secondary. *J. Early Adolesc.* 35, 651–680. doi: 10.1177/0272431614537117

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A., Jones, S. M., et al. (2013). Teaching through interactions. Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *Elem. Sch. J.* 113, 461–487. doi: 10.1086/669616

Hamre, B. K., Pianta, R. C., Mashburn, A. J., and Downer, J. T. (2007). Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. early childhood and elementary classrooms. *Foundation for Childhood Development*. Available at: https://www.fcd-us.org/wp-content/uploads/2016/04/BuildingAScienceOfClassroomsPiantaHamre.pdf (Accessed May 05, 2024).

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Herrmann, C., and Gerlach, E. (2020). Unterrichtsqualität im Fach Sport – Ein Überblicksbeitrag zum Forschungsstand in Theorie und Empirie. *Unterrichtswissenschaft* 48, 361–384. doi: 10.1007/s42010-020-00080-w

Hofkens, T., Pianta, R. C., and Hamre, B. (2023). "Teacher-student interactions: theory, measurement, and evidence for universal properties that support students' learning across countries and cultures" in *Effective teaching around the world: Theoretical, empirical, methodological and practical insights* (Cham: Springer International Publishing), 399–422. doi: 10.1007/978-3-031-31678-4_18

Hossner, E.-J., and Künzell, S. (2022). Einführung in die Bewegungswissenschaft. *Limpert*.

Kane, T. J., and Staiger, D. O. (2012). Gathering feedback for teaching. Research Paper. MET Project. Bill and Melinda Gates Foundation. Available at: http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf (Accessed May 05, 2024).

Klieme, E., Pauli, C., and Reusser, K. (2009). "The Pythagoras study: investigating effects of teaching and learning in Swiss and German mathematics classrooms" in *The power of video studies in investigating teaching and learning in the classroom*. eds. T. Janik and T. Seidel (Münster: Waxmann).

Kohake, K., Richartz, A., and Maier, J. (2022). Measuring pedagogical quality in children's sports: validity and reliability of the classroom assessment scoring system K-3 in extracurricular sports training. *Ger. J. Exerc. Sport Res.* 53, 47–58. doi: 10.1007/s12662-022-00836-9

Kounin, J. (1976). *Techniken der Klassenführung* (Nachdruck) Bern: Waxmann.

Kunter, M., and Ewald, S. (2016). "Bedingungen und Effekte von Unterricht: Aktuelle Forschungsperspektiven aus der pädagogischen Psychologie" in *Bedingungen und Effekte guten Unterrichts*. eds. N. McElvany, W. Bos, H. G. Holtappel, M. M. Gebauer and F. Schwabe (Münster: Waxmann), 9–31.

Li, H., Liu, J., and Hunter, C. V. (2020). A Meta-analysis of the factor structure of the classroom assessment scoring system (CLASS). *J. Exp. Educ.* 88, 265–287. doi: 10.1080/00220973.2018.1551184

Maier, J. (2023). *Individual video-supported learning guidance to improve the teaching quality in physical education. Further development, implementation and evaluation of video-based teaching-learning environments for physical education* Teachers [Dissertation]. Hamburg: Universität Hamburg.

Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., et al. (2010). A validation of the classroom assessment scoring system in Finnish kindergartens. *Early Educ. Dev.* 21, 95–124. doi: 10.1080/10409280902858764

Pianta, R. C., DeCoster, J., Cabell, S., Burchinal, M., Hamre, B. K., Downer, J. T., et al. (2014). Dose-response relations between preschool teachers' exposure to components of professional development and increases in quality of their interactions with children. *Early Child. Res. Q.* 29, 499–508. doi: 10.1016/j.ecresq.2014.06.001

Pianta, R. C., and Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: standardized observation can leverage capacity. *Educ. Res.* 38, 109–119. doi: 10.3102/0013189X09332374

Pianta, R. C., La Paro, K. M., and Hamre, B. K. (2008). *Classroom assessment scoring system. Manual K-3*. Baltimore: Paul H. Brookes.

Pöysä, S., Vasalampi, K., Muotka, J., Lerkkanen, M. K., Poikkeus, A. M., and Nurmi, J. E. (2019). Teacher–student interaction and lower secondary school students' situational engagement. *Br. J. Educ. Psychol.* 89, 374–392. doi: 10.1111/bjep.12244

Praetorius, A. K., and Gräsel, C. (2021). Noch immer auf der Suche nach dem heiligen Gral: Wie generisch oder fachspezifisch sind Dimensionen der Unterrichtsqualität? *Unterrichtswissenschaft* 49, 167–188. doi: 10.1007/s42010-021-00119-6

Richartz, A., and Kohake, K. (2021). Zur (Fach-)Spezifität von Unterrichtsqualität im Fach Sport. *Unterrichtswissenschaft* 49, 243–251. doi: 10.1007/s42010-021-00112-z

Richartz, A., and Kohake, K. (2023). *Beyond facilitating higher order thinking: How does effective instructional support look like in motor learning, game-play-learning, and perceptual learning? In abstract to the 7th workshop on systematic observation in educational research*. Norway: University of Stavanger.

Richartz, A., Maier, J., and Kohake, K. (2021). "Pädagogische Qualität des Trainings im Kinder- und Jugendsport – normative und wirksamkeitsorientierte Kriterien" in *Kinder- und Jugendsportforschung in Deutschland – Bilanz und Perspektive*. ed. N. Neuber (Wiesbaden: Springer), 171–201. doi: 10.1007/978-3-658-30776-9_9

Ryan, R. M., and Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. New York: Guilford Press.

Virtanen, T. E., Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Siekkinen, M., and Nurmi, J. E. (2018). A Validation Study of Classroom Assessment Scoring System–Secondary in the Finnish School Context. *JEA*. 38, 849–880. doi: 10.1177/0272431617699944

Youngs, P., Elreda, L. M., Anagnostopoulos, D., Cohen, J., Drake, C., and Konstantopoulos, S. (2022). The development of ambitious instruction: how beginning elementary teachers' preparation experiences are associated with their mathematics and English language arts instructional practices. *Teach. Teach. Educ.* 110:103576. doi: 10.1016/j.tate.2021.103576