TYPE Original Research
PUBLISHED 18 June 2025
DOI 10.3389/feduc.2025.1570510

Check for updates

OPEN ACCESS

EDITED BY
Goran Kuvačić,
University of Split, Croatia

REVIEWED BY
Michael Rochnia,
University of Wuppertal, Germany
Bryan Catama,
Saint Louis University, Philippines

*CORRESPONDENCE
Clemens Berthold
✉ clemens.berthold@phsg.ch

RECEIVED 03 February 2025
ACCEPTED 27 May 2025
PUBLISHED 18 June 2025

CITATION
Berthold C, Jeisy E and Baumgartner M (2025) Pre-service teachers' knowledge of evidence-based classroom management practices in physical education: a test validation using item response theory.
Front. Educ. 10:1570510.
doi: 10.3389/feduc.2025.1570510

Pre-service teachers' knowledge of evidence-based classroom management practices in physical education: a test validation using item response theory

Clemens Berthold*, Eric Jeisy and Matthias Baumgartner

Institute of Physical Education, Sports and Health, St. Gallen University of Teacher Education, Rorschach, Switzerland

This study validated an assessment instrument measuring pre-service teachers' professional knowledge of evidence-based classroom management practices in physical education. Drawing on a model of teacher competence that integrates knowledge, situation-specific skills, and performance, the study focused on the competence area of classroom management to ensure conceptual clarity and relevance. Data from 877 pre-service primary education teachers from four universities of teacher education were analyzed using item response theory to examine the instrument's structure and psychometric properties. The findings indicate a unidimensional structure with satisfactory reliability and no evidence of bias related to demographic variables. Test scores showed a small positive correlation with situation-specific skills, reflecting construct validity, as these require additional distinct cognitive abilities while being conceptually related. However, the test's items proved relatively easy, resulting in a mismatch between item difficulty and participant ability levels, and did not capture the expected differences across pre-service teachers at different stages of their training, potentially due to a ceiling effect. Together, these findings limit the test's capacity to differentiate among higher-ability individuals, thereby constraining criterion validity. Despite these limitations, the results demonstrate the instrument's capacity to measure knowledge about evidence-based practices in classroom management. Further refinement could enhance its discriminatory power at advanced knowledge levels. This assessment provides a foundation for exploring how knowledge shapes teachers' perception, interpretation, decision-making, and performance, and could support efforts in teacher education to develop effective classroom management practices.

KEYWORDS

physical education, evidence-based practice, declarative knowledge, test development, test validation, professional competence, classroom management

1 Introduction

Professional knowledge is a central element of teachers' competencies and teacher education, preparing pre-service teachers to address the specific demands of real-world classrooms (Guerriero, 2017). Unlike beliefs or motivational orientation, it grounds educational decisions and instructional strategies in objective, evidence-based insights (e.g., Fenstermacher, 1994). Higher levels of assessed knowledge correlate with higher levels and increased stability of instructional quality (Blömeke et al., 2022; Voss et al., 2022). However, professional knowledge alone cannot ensure effective teaching (Baumgartner, 2018). It forms part of a dynamic continuum of teacher competence that integrates three facets (1) aspects of competency—such as professional knowledge, (2) situation-specific skills—Perception, Interpretation, and Decision-making (PID), and (3) performance in authentic teaching contexts (Baumgartner, 2022; Blömeke et al., 2015a). Within this continuum, knowledge is assumed to shape teachers' PID, and guide effective performance. Together, these three facets constitute professional competence in relation to a competence area[1]; a teacher may excel in Classroom Management (CM) but struggle with providing constructive feedback (Blömeke et al., 2015a; Blömeke et al., 2015b). Consequently, the extent to which knowledge predicts performance, or relates to PID, may depend on the competence area in focus (Blömeke et al., 2015a).

Previous studies applying this continuum of teacher competence have lacked critical aspects affecting their comparability, theoretical alignment, and proximity to practice (Charalambous, 2020). First, they often do not provide a focused examination of one specific competence area, such as CM (Blömeke et al., 2022; Römer and Rothland, 2015). Second, while most studies rely on previously validated instruments, some of these tools fail to distinctly measure one individual facet of competence, such as professional knowledge (Brühwiler et al., 2017; König and Kramer, 2016; Lenske et al., 2016). Third, the measures frequently prioritize theoretical concepts over the real-world demands of classroom teaching, reducing the practical relevance and interpretability of findings (Brühwiler and Hollenstein, 2021; Lüders, 2012).

The Swiss National Science Foundation (SNSF)-funded project *"From Knowledge to Performance in Physical Education: Pre-service PE Teachers' Transformation of Competences – an intervention study on classroom management (WiPe-Sport)"* investigates how pre-service teachers develop and apply their CM-related competence in Physical Education (PE) (Baumgartner et al., 2023). As part of the project, a multi-stage, quasi-experimental intervention study investigates the relationship and development of CM-related knowledge, PID, and performance in teacher education. To address these questions two instruments were developed within the project: one to measure CM-related knowledge and another to assess PID (cf. ibid.; Jeisy et al., in prep.). Both instruments draw on the nine dimensions of effective CM used in the validated observation instrument by Baumgartner et al. (2020). and have previously undergone content validation through a Delphi study (Baumgartner et al., 2023).

This paper focuses on the recently developed knowledge test that targets evidence-based practices in CM for PE. The test is designed to comprehensively measure professional knowledge as a distinct facet of teacher competence. After initial content validation the next methodological step was to administer the test to a sample of pre-service teachers. Psychometric properties are analyzed through Item Response Theory (IRT). Criterion validity is assessed by examining the test's sensitivity to pre-service teachers' educational progression, and construct validity is explored through its relationship with PID. Together, these analyses aim to establish the instrument as a valid, reliable, and objective measure of CM-related knowledge in PE.

2 Theoretical background

2.1 Theoretical framework of teacher competence

Professional competence is a complex, hypothetical construct that cannot be directly observed (Shavelson, 2013). In educational measurement, it is often either holistically inferred from behavior in specific performance situations or analytically pieced together from aspects of competency such as knowledge and cognitive, affective, and motivational dispositions (Baumgartner, 2022; Blömeke et al., 2015a). However, Blömeke et al. (2015a) caution that both approaches have limitations: a sole focus on observable behavior may neglect the underlying aspects of competency and situation-specific skills essential for real-world performance, while an analytic perspective might overlook the dynamic interaction between these facets (Baumgartner, 2022).

To address these issues, Blömeke et al. (2015a) proposed a model viewing teacher competence as a continuum from aspects of competency (e.g., professional knowledge) through situation-specific PID to actual teaching performance. In this model and its adaptation to PE (Baumgartner, 2022), these three facets are assumed to be positively correlated and cumulative. According to this framework, (pre-service) teachers with higher levels of knowledge and PID are likely to perform better in practice (Baumgartner, 2022; Blömeke et al., 2022; König et al., 2021) and targeted improvements in one facet, such as professional knowledge, should enhance performance (e.g., Blömeke et al., 2022).

To better understand how these facets are connected, it is helpful to consider the underlying cognitive mechanisms. Professional knowledge may activate or restructure prior (experiential) knowledge (Boshuizen et al., 2020), correct misconceptions (Fenstermacher, 1994; Kleickmann, 2023) and support the use and adaptation of evidence-based practice (Renkl, 2022; Wilkes and Stark, 2022). Rather than offering ready-made solutions, such knowledge supports the justification, adaptation, and evaluation of instructional decisions (Bauer and Kollar, 2023; Heins and Zabka, 2019). It enhances (pre-service) teachers' capacity to encode and organize complex classroom information, enabling them to process multiple, simultaneous events more efficiently. As information processing becomes more knowledge-driven, activated schemata and scripts guide attention, filter relevant cues, and help structure classroom events into meaningful patterns—thereby enhancing perception and interpretation amid classroom complexity (Gegenfurtner et al., 2023; Heins and Zabka, 2019). These processes strengthen teachers' flexibility, precision, and ability not only to respond appropriately but also to shape their environment through informed instructional actions.

Finally, teacher competence develops and manifests within distinct competence areas (e.g., classroom management):

[1] A competence area denotes a specific cluster of teaching practices and demands essential for successful teaching, such as classroom management (Baumgartner, 2022). By contrast, domain-specificity refers to the field of expertise (e.g., teaching; e.g., Boshuizen et al., 2020), while subject-specificity pertains to knowledge and skills unique to academic disciplines (e.g., mathematics; Jeschke et al., 2019). Finally, situation-specificity emphasizes how performance can vary depending on contextual factors (e.g., Blömeke et al., 2015b).

functionally and thematically defined clusters of teaching practices that require the coordinated use of knowledge, PID, and performance (Baumgartner, 2022). While this assumes that certain dimensions of knowledge, PID, and performance are more strongly connected than others, it does not imply a simple one-to-one mapping between these facets (Renkl, 2022; Wilkes and Stark, 2022).

2.2 Classroom management in physical education: a key competence area

CM broadly refers to teachers' efforts to create and sustain an environment that supports students' cognitive, social–emotional, and motor development (Baumgartner et al., 2020; Brophy, 2006). These efforts involve using behavioral and instructional strategies to guide student learning, increase on-task behavior, and preventatively or reactively address student misbehavior (Emmer and Stough, 2001; Korpershoek et al., 2016; Oliver et al., 2011; Simonsen et al., 2008). While CM is considered a generic aspect of teaching, it poses unique challenges in PE due to the subject's distinctive learning settings and demands (Baumgartner et al., 2020; Cothran and Kulinna, 2015; Herrmann and Gerlach, 2020).

Effective CM is crucial for enhancing students' attention, motivation, engagement, and learning outcomes (Korpershoek et al., 2016; Kunter et al., 2007; Oliver et al., 2011). Conversely, classroom disruptions can undermine student self-efficacy and achievement and diminish the positive impact of teacher need support (Burns et al., 2021). For teachers, mastering CM can reduce stress and mitigate the risk of burnout (Aloe et al., 2014; Dicke et al., 2015; König and Rothland, 2016).

Despite the importance of CM, many teachers, including those in training, continue to describe it as a significant professional challenge, often feeling unprepared to manage classrooms effectively (Dicke et al., 2015; Ulferts, 2019; Stokking et al., 2003). This sense of unpreparedness contrasts sharply with recent research indicating high levels of CM-related knowledge (Dückers et al., 2022; Junker et al., 2021; Schlag and Glock, 2019), and CM-related performance (Gold et al. 2021; Junker et al., 2021) among (pre-service) teachers.

Given these challenges, PE teachers need to implement strategies tailored to the unique demands of their teaching context (cf. Baumgartner et al., 2020; Cothran and Kulinna, 2015). They need to manage the high noise levels and sustain communication with physically active students (Ryan and Swartz, 2018). Teachers have to establish distinct rules and routines for varying environments, including gymnasiums, fields, and swimming pools (Hummel and Krüger, 2015). The dynamic and fast-paced nature of PE demands active supervision, which involves constant movement, strategic positioning, and frequent interactions with students to maintain rapport and ensure safety (Arbogast and Chandler, 2005; van der Mars et al., 1994). PE-specific CM further emphasizes efficient transitions including frequent student grouping and the cooperative handling of bulky equipment or large amounts of materials (Giessing, 2010; Raith, 2017). Teachers must establish and enforce specific safety protocols tailored to different types of sports to minimize risks and ensure a secure learning environment. Additionally, they have to attend to students who do not actively participate (Wolters, 2021).

2.3 The role of professional knowledge in CM-related competence

Professional knowledge is typically organized into three main categories: subject-specific knowledge, pedagogical content knowledge, and General Pedagogical Knowledge (GPK; Shulman, 1986; Guerriero, 2017). Subject-specific and pedagogical content knowledge relates directly to the subject being taught, whereas GPK constitutes the "specialized knowledge of teachers for creating effective teaching and learning environments for all students, independent of subject matter" (Guerriero, 2017, p. 80). CM is mostly seen as an important area for the practical application of GPK (Leijen et al., 2022; Voss et al., 2015).

Meta-analytical findings suggest that GPK, which generally includes knowledge about CM, has moderate effects on teaching quality and a small impact on student academic and social–emotional outcomes (König, 2014; Ulferts, 2019). When focusing specifically on CM-related performance, studies indicate that GPK has small to moderate correlations with CM-related performance as perceived by students (König and Pflanzl, 2016). Observational studies also show that GPK positively influences CM-related performance, often as part of broader instructional quality. For example, König et al. (2021) and Voss et al. (2014) highlighted the role of GPK in shaping instructional quality, particularly in the competence area of CM. Furthermore, Lenske et al. (2016) demonstrated that GPK has both direct effects on student outcomes and indirect effects mediated through observed CM-related performance. In contrast, Blömeke et al. (2022) find these effects mediated via PID rather than through performative aspects, suggesting that the role of mediating factors in linking GPK to student outcomes is not yet fully clarified. The relationship between GPK and situation-specific skills varies considerably. Correlations range from low to moderate ($r = 0.13$ to 0.36) to high ($r = 0.56$; cf. Müller and Gold, 2022), depending on factors such as the type of knowledge assessed, the configuration of skills (e.g., PID), and teachers' professional development level (Bastian et al., 2024; Junker et al., 2021; Weber et al., 2023).

2.4 Measurement of CM-related knowledge

Despite its acknowledged importance, GPK remains underexplored (Ulferts, 2019), and the generalizability of findings is limited by variability in GPK conceptualization and assessment (Brühwiler et al., 2017; Leijen et al., 2022; Voss et al., 2015). Differences in contextualization, assessment design, and data collection approaches further contribute to contradictory results, reducing comparability across studies (Brühwiler et al., 2017; Brühwiler and Hollenstein, 2021). These inconsistencies complicate the interpretation of the relationship between teachers' knowledge and their classroom performance (Charalambous, 2020), highlighting the complexity of choices that must be made when developing and validating assessment instruments (Brühwiler and Hollenstein, 2021).

2.4.1 Challenges in comparing GPK conceptualizations and designs

GPK is inherently broad and generic, making direct comparisons across studies difficult. Existing measurement instruments often differentiate between multiple, yet inconsistent, dimensions (Leijen et al., 2022; Pollmeier et al., 2024; Voss et al., 2015), which can lead to outcomes that fail to correlate meaningfully (König and Seifert, 2012). This limits insights into specific areas, such as CM (Brühwiler and Hollenstein, 2021; Römer and Rothland, 2015). While CM-related knowledge is typically embedded in broader GPK assessments, it is rarely applied as an independent dimension. For example, when reporting the effects of GPK on CM-related performance using tests from the COACTIVE-R study (Voss et al., 2011, 2014), the TEDS-M study (König et al., 2011; König and Kramer, 2016), or the ProwiN study (Lenske et al., 2015, 2016), the specific contribution of CM-related knowledge is not disentangled from other aspects of GPK. Moreover, when these dimensions are not empirically separable within a given instrument (e.g., Lenske et al., 2015; Voss et al., 2014), their individual use can pose challenges in their interpretation and application (e.g., Junker et al., 2021).

2.4.2 Aligning contextualization with cognitive demands

Another challenge lies in the alignment of test formats with the cognitive demands placed on teachers. Contextualized assessments, such as text- or video-based formats, present teachers with realistic classroom scenarios. These formats additionally require situation-specific skills, which can blur the boundaries between declarative knowledge and PID (Brühwiler and Hollenstein, 2021; Gold and Holodynski, 2015; Kaiser et al., 2017; König and Kramer, 2016). While such assessments provide richer insights into teacher competence and have been shown to improve the prediction of CM-related performance (König and Kramer, 2016; Lenske et al., 2016), they can reduce comparability across studies due to their unique contextual features and varying levels of cognitive demands (Brühwiler et al., 2017; Brühwiler and Hollenstein, 2021).

2.4.3 Proximity of knowledge to performance

The proximity between teacher knowledge assessments and actual teaching performance is crucial for understanding their relationship (Charalambous, 2020; Lüders, 2012). Instruments focusing on theoretical scientific knowledge may capture teacher education outcomes but often fail to reflect the demands of real-world classroom situations (Brühwiler and Hollenstein, 2021; Lüders, 2012). The current push towards evidence-based teaching emphasizes the need for professional knowledge that directly informs and improves classroom practices and is grounded in empirical findings (Knogler et al., 2022; Prenzel, 2020; Slavin, 2002; Smith, 2024).

2.5 Evidence-based practices in classroom management

Most research on CM focuses on identifying effective practices and strategies that produce measurable positive effects on student behavior and learning outcomes (Emmer and Stough, 2001; Korpershoek et al., 2016; Simonsen et al., 2008). Such "professional behaviors, decisions, and practices oriented towards improving school or classroom practices and based on relevant empirical findings and scientific facts" (Zlatkin-Troitschanskaia et al., 2016, p. 61) are understood as evidence-based practices. Intervention studies targeting teachers' CM strategies show strong evidence for their effectiveness in controlled conditions (Korpershoek et al., 2016). Together, a solid body of scientific knowledge about evidence-based practices exists, providing a foundation for assessing knowledge, PID (e.g., Weyers et al., 2023), and performance (Albu and Lindmeier, 2023). However, in the field of PE, the specific database is considerably less extensive, particularly regarding subject-specific dimensions of CM.

While high-quality evidence derived from meta-analyses and randomized controlled trials is critical for deriving evidence-based practices, relying solely on such broad synthesis can overlook the contextual nuances of teaching (Renkl, 2022). Additionally, the effectiveness of these practices depends on teachers' ability to implement them with fidelity and adapt them to real-world contexts (Cook et al., 2012; Renkl, 2022). Therefore, there is a need for syntheses that balance robust empirical support with practical, context-sensitive relevance (Knogler et al., 2022; Smith, 2024).

The challenges posed by contextualization, cognitive demands, and varying proximities to actual performance in conceptualizing and measuring CM-related knowledge underscore the need for more nuanced assessment approaches. Additionally, the growing emphasis on evidence-based teaching highlights the urgency of refining these assessments to better align with the realities of classroom practice. Integrating empirically validated CM strategies into assessment designs can help future research and test development thereby creating measures that are scientifically grounded, ecologically valid, and better predictors of CM-related performance.

3 Development and validation of the CM-related knowledge test

Building on the above considerations, this section introduces the CM-related knowledge test. It first outlines the test's theoretical framework and summarizes its development and content validation process (see Baumgartner et al., 2023). Second, it details the objectives and hypotheses of the current validation approach.

3.1 Prior steps: test development and content validation

The CM-related knowledge test was developed as part of the SNSF-funded "WiPe-Sport" project, which investigates the development and application of CM-related competence in pre-service PE teachers. It is grounded in the nine observable dimensions of good CM in PE identified by Baumgartner et al. (2020). These dimensions define the scope of all instruments within the project and include general pedagogical skills, such as monitoring, and two PE-specific dimensions: ensuring safety and managing equipment. Each dimension is represented by a set of evidence-based, actionable strategies tailored to the unique demands of PE. Together, these CM strategies emphasize an evidence-based, "technical" perspective on teaching, focusing on basic techniques that are proven effective in practice. For example, the observation instrument includes the rating of the monitoring strategy *"The PE teacher chooses positions in the room from which she/he has a good overview of what is going on in the class."*

The development of the CM-related knowledge test began with identifying evidence on effective strategies across the nine dimensions of CM. Evidence was selected and analyzed from a range of high-quality sources, including meta-analyses (Hattie, 2010; Marzano et al., 2003), systematic reviews (Landrum and Kauffman, 2006; Simonsen et al., 2008) and original research (e.g., van der Mars et al., 1994). Due to the scarcity of empirical research specific to PE—particularly concerning safety and equipment management—practice-oriented sources such as normative criteria for good CM (Ophardt and Thiel, 2013) and practical recommendations (Söll and Kern, 1999) were also incorporated to gather the best available information on these critical dimensions (Knogler et al., 2022; Smith, 2024).

Test items were constructed to reflect a single CM strategy, requiring participants to judge whether or not it represents an effective, evidence-based instructional practice (true/false format)., For example: "To monitor the classroom, a teacher should choose a fixed position that allows him/her to keep all students in sight" (false, dimension of monitoring). In contrast to established GPK tests, which often rely on broad, theoretically derived constructs (Lüders, 2012) this test focuses exclusively on declarative knowledge about CM strategies assessed in a non-contextualized format. This design aims to isolate professional knowledge as a distinct facet of professional competence by reducing the additional cognitive demands associated with contextualized assessment (Brühwiler and Hollenstein, 2021). In contrast, the PID instrument used in the project elicits reflective, situation-specific responses. For example, participants are prompted with: *"If you were the teacher in this situation, what would you do differently to improve classroom management?"* A typical answer aligned with the monitoring dimension might be: *"When instructing and demonstrating, I position myself in such a way that I also keep an eye on the small group playing."*

To ensure content validity, a Delphi study involving experts in teacher education, PE pedagogy, and CM research was conducted. Multiple rounds of feedback resulted in a consensus on the appropriateness and quality of the test items. This iterative process resulted in an instrument consisting of 104 items that provide a comprehensive representation of CM-related knowledge aligned with empirical evidence and firmly rooted in the realities of PE classrooms (Baumgartner et al., 2023). All items of the final test are available in the Supplementary material.

3.2 Study objectives, hypotheses and design

The objectives of this study are to evaluate the test's (internal) psychometric properties and provide evidence for (external) criterion and construct validity. First (H1), the test is expected to capture the construct of CM-related knowledge, demonstrating adequate reliability and model parameters within a unidimensional model. Second (H2), the test is expected to reflect criterion validity by effectively differentiating between knowledge levels of pre-service teachers at various stages of their education, with higher scores indicating the accumulation of knowledge over time (König et al., 2024; Weyers et al., 2024). Third (H3), construct validity focuses on the relationship between CM-related knowledge and PID, hypothesizing a positive, yet small, correlation (cf. Müller and Gold, 2022).

4 Method

4.1 Participants

877 pre-service teachers, specializing in primary education (740 = *female,* 130 = *male*, 7 = *divers*) participated in this study. At the time, they were enrolled in one of four participating Swiss Universities of Teacher Education (UTEs) (UTE St. Gallen [473], UTE Lucerne [357], UTE Fribourg [41], UTE Grisons [6]). Participants were evenly distributed across the first ($n = 277$), second ($n = 283$) and third ($n = 313$) year of study (four unreported). Among them, 275 were training to teach at the kindergarten level and 602 at the primary level.[2] Their average age was 23.4 years ($SD = 4.0$, $range = 18–54$).

4.2 Measurement

4.2.1 CM-related knowledge

The CM-related knowledge test, evaluated in this study, measures teachers' declarative, non-situated knowledge about effective CM practices, focusing on the nine dimensions outlined in Baumgartner et al. (2020). Initially, the test consisted of 104 dichotomous items, scored as correct or incorrect.

Following IRT analysis of local and global model fit (see section 4.4), a refined set of items was used to assess criterion and construct validity (see section 5).

4.2.2 CM-related PID

The CM-related PID test evaluates teachers' situation-specific skills in CM using seven video vignettes (duration: 1:19–3:27 min). Each vignette covers at least two of the nine CM dimensions, ensuring that all dimensions are addressed multiple times. After viewing, participants answered dichotomous items targeting the three cognitive demands of PID. For example, participants interpreted a situation by selecting appropriate responses to the question: "Which of the following CM-related teachers' actions happened?" One example, focusing on monitoring, was: "The teacher concentrates on the group doing gymnastics on the rings without losing sight of the class." Psychometric analysis by Jeisy et al. (sub.) supported a one-dimensional solution for the situation-specific skills, indicating that the PID can be treated as a unified construct. Reliability was acceptable ($EAP = 0.674$; $WLE = 0.639$).

4.3 Data collection

Data were collected online using the LimeSurvey (LimeSurvey GmbH, n.d.) software between March and April 2022. The CM-related

[2] In Switzerland, primary education spans 8 years, divided into kindergarten (2 years) and primary school (6 years). The system is decentralized, with cantons overseeing curricula and teacher qualifications. Teacher education at UTEs combines coursework, pedagogical training, and internships, leading to a Bachelor's degree in Primary Education. While preservice teachers usually study all subjects in the curriculum, qualifications may focus on specific levels, such as lower or upper primary grades, rather than all primary levels (IDES, n.d.).

knowledge test was surveyed alongside the PID test and a section on personal information such as teaching experience in PE and self-assessed quality of CM.

The knowledge test employed a booklet design. Items were grouped into sets according to their CM dimension and assigned to seven booklets. Following a balanced incomplete block design (Frey et al., 2009), each booklet included three sets, with all dimensions occurring equally across booklets. Each participant completed items from three to five CM dimensions, resulting in 360–386 responses per item. To mitigate order effects, the sequence of the tests was randomized. On average, participants took 33 min (*SD* = 13.0) to complete the survey, with 17.1 min (*SD* = 8.0) on the PID test and 5.4 min (*SD* = 4.5) on the knowledge test.

Some participants completed the test during a course (UTE St. Gallen, UTE Lucerne) and others in their free time (UTE Fribourg, UTE Grisons). Of the 1,473 registered participants, 1,076 completed the survey (73% response rate), and 877 (59%) were included in the final analysis after data cleaning. Only participants who completed all items and met realistic completion times—determined as 15 min total, with 2 min for the knowledge test and 8 min for the PID test, based on video runtime and task completion estimates—were included to ensure data quality.

4.4 Data analysis

The internal psychometric properties and structure of the test instrument were assessed using a combination of exploratory and confirmatory approaches using IRT and interferential statistics. IRT models describe the probabilistic relationship between individuals' latent traits (e.g., ability or proficiency) and their performance on test items, characterized by parameters representing item difficulty, discrimination, and guessing. A good model fit implies that the model parameters adequately explain test outcomes (Moosbrugger and Kelava, 2020).

The most appropriate model was identified by comparing: (a) IRT models with different parameter specifications, (b) global model fit, and (c) local model fit. (a) The three-parameter logistic model (3*PL*) was theoretically expected to describe the data best due to the test's dichotomous items (allowing for guessing) and its broad content range (indicating variations in item discrimination and difficulty). This model was compared with simpler, nested solutions: the two-parameter logistic model (*2PL*), which accounts for varying item discrimination and difficulty, and the one-parameter logistic model (*1PL*), which assumes uniform discrimination across items (Bond et al., 2021). Model fit was compared using likelihood ratio (*LR*) tests, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), expected a posteriori (EAP) and weighted likelihood estimates (WLE) reliability indices, and theta variance. (b) Global model fit was assessed using chi-square statistics with multiple testing corrections, as implemented in the TAM package. Based on these results, the item whose removal led to the greatest improvement in global model fit was excluded, and the analysis iteratively repeated until an acceptable global model fit was reached (cf. Nielsen and Dammeyer, 2019). (c) Local item fit was evaluated using infit and outfit statistics, which reflect the alignment of the items with the model expectations, ensuring values fell within acceptable ranges (0.8–1.2, or *t*-standardized values between −1.96 and 1.96; Bond et al., 2021). Differential Item Functioning (DIF) analysis using the Mantel–Haenszel test was conducted to examine whether item responses were conditionally independent of demographic variables such as gender, mother tongue, university placement, and participation type.

The test structure was examined by comparing a one-dimensional model with two multidimensional models to determine if accounting for either specific CM dimensions or booklet improved fit. Specifically, comparisons were made between the one-dimensional solution and: (a) a nine-dimensional model based on the nine CM-related dimensions, and (b) a seven-dimensional model aligning items with their respective test booklets. Models were compared using LR tests, AIC, and BIC to evaluate whether the more complex models provided a significant improvement in fit.

A Wright Map was used to evaluate the alignment between item difficulty and participant ability, providing insight into the overall fit between the test and the sample. In this map, the mean of the person ability is set as the zero point on the logit scale, and item locations are plotted relative to this origin. A person located at the same point as an item has a 50% probability of correctly answering it. This allows for a visual interpretation of the relationship between item difficulty and participant ability (Bond et al., 2021).

The (external) construct validity of the instrument was assessed by calculating correlations between person ability estimates of knowledge and PID. Additionally, the capacity of the test to distinguish between pre-service teachers at different stages of their studies was examined using ANOVA, with post-hoc tests conducted as necessary, as an indicator for criterion validity.

All analyses were conducted using R Project for Statistical Computing (RRID: SCR_001905) in RStudio (2023.06.01). The TAM package (Version 4.1.4), with marginal maximal likelihood estimation (e.g., tam.mml.2pl()) was used for IRT modeling.

5 Results

Model selection began with the more complex 3PL model but favored simpler models based on LR tests (see Table 1). While both the 2PL and 1PL models showed no significant loss in fit, the 1PL model lacked sufficient reliability. Consequently, the 2PL model was selected for further analysis.

Based on global fit measures, 25 items were iteratively eliminated until a non-significant result of the global model fit test X^2 (3084) = 17.34; *p* = 0.068 was achieved, indicating no significant deviation from model assumptions. This refined version showed acceptable reliabilities (EAP = 0.603, WLE = 0.569). Local fit analysis using mean squared residual statistics (tam:msq.itemfit()) further supported model assumptions, with outfit values ranging from 0.85 to 1.05 (*t*-standardized: −0.31 to 0.66) and infit values from 0.99 to 1.01 (*t*-standardized: −0.19 to 0.19). The remaining 79 items assess CM across the nine dimensions of: monitoring (12), dealing with disruptions (10), clarity of announcements (11), group mobilization (8), momentum (10), overlapping (3), smooth transitions (4), safety (12), and use of material (9). Full item information can be found in the Supplementary material.

Comparison of the nested, multi-dimensional models with the unidimensional model revealed no significant improvement in model fit, supporting the assumption of unidimensionality (see Table 2).

TABLE 1 Model comparison and fit indices for 3PL, 2PL, and 1PL item response models.

Model	loglike	Deviance	Parameters	*N*	*AIC*	*BIC*	Model fit test	*EAP* reliability	*WLE* reliability	Theta variance	Likelihood ratio test
3PL	−17,511	35,021	221	877	35,463	36,519	$\chi^2(5995) = 122.6$ $p = 1$	0.677	–	0.275	
2PL	−17,512	35,023	220	877	35,463	36,514	$\chi^2(5995) = 123.1$ $p = 1$	0.673	0.343	1	$\chi^2 = 2.11$, df = 1, $p = 0.14$
1PL	−16,185	32,370	111	877	32,592	33,122	$\chi^2(5995) = 120.6$ $p = 1$	0	0.325	0.001	$\chi^2 = -2{,}653$, df = 109, $p = 1$

loglike, natural logarithm of the likelihood function; *N*, number of participants; AIC, Akaike information criterion; BIC, Bayesian information criterion; EAP, expected a posteriori; WLE, weighted likelihood estimate; PL, parameter logistic.
Model fit Test: Test statistic of global fit based on multiple testing correction of χ^2 statistics.
EAP and WLE measures can be interpreted like Cronbach's alpha values.

Further item-level analysis did not reveal any indication of differential item functioning regarding gender, mother tongue, university placement, or participation type.

The Wright Map (see Figure 1) showed that item difficulty (right) and person ability (left) appear to be normally distributed. However, item difficulty is lacking range, with most items being easier than the participants' ability levels. This suggests the test may struggle to differentiate between participants with higher ability levels.

Consistent with this, ANOVA across study years (1, 2, and 3) revealed no significant differences in performance (F (2, 860) = 0.33; $p = 0.72$). Finally, construct validity was supported by a significant but small correlation between person estimates from the PID and knowledge tests ($r = 0.16$; $p < 0.001$), suggesting that, while related, professional knowledge and PID are distinct facets of competence.

6 Discussion

This study provides evidence about the validity of a test instrument designed to measure professional knowledge about evidence-based CM practices in PE. The test, previously content validated through expert consensus in a Delphi study, was administered to a sample of 877 pre-service teachers. Results show adequate psychometric properties and reliability using IRT and indicate construct validity, confirming its effectiveness in measuring CM-related declarative knowledge. However, the findings also highlight areas for refinements that could improve its applicability and ability to distinguish between varying levels of participant knowledge.

6.1 Validation: insights and challenges

The psychometric properties of the test confirm Hypothesis H1, demonstrating that it effectively captures the construct of CM-related knowledge within a unidimensional model. The CM-related knowledge test contains 79 items, demonstrating adequate psychometric properties under a 2PL IRT model. Both global and local fit assessments confirm that the model assumptions were met with acceptable infit and outfit statistics. Additionally, no evidence of DIF across demographic subgroups underscores the robustness of the test across the diverse pre-service teacher population.

The results do not support H2, as the expected differences between participants across study years were not observed. The Wright Map reveals a mismatch between item difficulty and participant ability, with most items being easier than the abilities demonstrated by the participants. The limited range of item difficulty likely restricted the test's ability to distinguish among higher-ability individuals, likely due to ceiling effects. This impacts the test's criterion validity and questions its sensitivity. However, the challenge of capturing advanced CM-related knowledge is not unique to this study, as other CM-related studies have reported similar issues. For instance, Dückers et al. (2022) observed ceiling effects in declarative knowledge assessments. Schlag and Glock (2019) found that pre-service teachers' strategic knowledge often matched or exceeded that of in-service teachers. Junker et al. (2021) noted that both pre-service and beginning teachers demonstrated high levels of pedagogical knowledge, with minimal differences between these groups. Another possible explanation for this lack of differentiation is the study's context—a teacher education

TABLE 2 Comparison of unidimensional and multidimensional IRT models.

Model	loglike	Deviance	Parameters	*N*	*AIC*	*BIC*	Likelihood ratio test
2*PL* – global (unidimensional)	−12,765	25,529	158	877	25,845	26,600	
2*PL* – booklet (7 – dimensional)	−13,629	27,258	335	877	27,928	29,528	$\chi^2 = -1728$, df = 177, $p = 1$
2*PL* – content (9 – dimensional)	−13,088	26,175	194	877	26,563	27,490	$\chi^2 = -646$, df = 36, $p = 1$

loglike, natural logarithm of the likelihood function; *N*, number of participants; AIC, Akaike information criterion; BIC, Bayesian information criterion; EAP, expected a posteriori; WLE, weighted likelihood estimate; PL, parameter logistic.

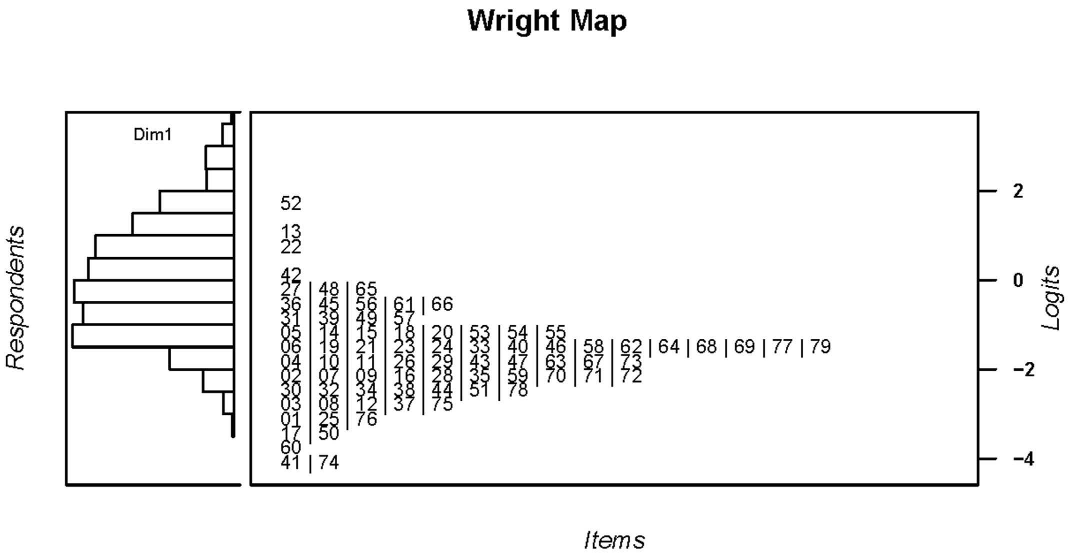

FIGURE 1
Wright map illustrating the alignment between item difficulty and participant ability. The mean of participant ability is set as the zero point on the logit scale, with item locations plotted relative to this origin. A participant positioned at the same point as an item has a 50% probability of correctly answering it.

program that integrates theoretical coursework with practical experience. This structure may contribute to higher levels of practice-related knowledge across all stages of training, making differences between groups less pronounced. Compared to validation studies of similar instruments that used broader and possibly more heterogeneous samples—ranging from first-year students to advanced in-service teachers—our more homogeneous sample likely reduced the variance in participant ability (Gold and Holodynski, 2015; Lenske et al., 2015), thereby increasing demands on the test's sensitivity.

Furthermore, the correlation between CM-related knowledge and PID is significant but small, which aligns with Hypothesis H3 and supports construct validity. While professional knowledge and PID are conceptualized as distinct yet related facets of teacher competence (Blömeke et al., 2015a), the small size is unexpected, given that the instruments were designed to align closely. This finding, however, is consistent with prior research indicating that while declarative knowledge may be sufficient for responding to predetermined situational interpretations, it is less effective for independently generating context-sensitive interpretations (Müller and Gold, 2022). Similarly, weak, or non-significant links between declarative knowledge and the ability to interpret or react to CM-specific events have been reported (Junker et al., 2021; Weber et al., 2023). These insights highlight the need for clearer distinctions between knowledge types and more precise measurements of their influence on the dimensions of PID (Weber et al., 2023). Additionally, the modest effect size observed suggests that other factors, such as self-efficacy beliefs, may influence this relationship (Depaepe and König, 2018; Junker et al., 2021; Leijen et al., 2024).

Further research is needed to better understand how different types of knowledge and PID skills interact and shape effective teaching practice. Given the correlational methodology of most studies, it remains unclear whether these facets co-evolve (Boshuizen et al., 2020) or follow a sequential development process (Blömeke et al., 2022). For now, it is expected that pre-service teachers are likely to benefit most when drawing on multiple sources of knowledge, including scientific research, experiential insights, and contextual understanding (Renkl, 2022).

6.2 Limitations

Although the instrument is based on a broad definition of CM, it primarily reflects a teacher-centered, method-focused approach by emphasizing evidence-based strategies, which may be particularly relevant in the early stages of teacher education (König, 2023). More collaborative frameworks, such as social and emotional learning were not included, risking missing the conditions that lead to off-task behaviors (Freiberg et al., 2020; Freiberg and Lamb, 2009).

The use of chi-square statistics to assess global fit is debated in the context of IRT. Disagreements persist regarding the appropriate specification of degrees of freedom for the null chi-squared distribution, and there are concerns about its sensitivity to sample size (Ranger and Much, 2020; Stone and Zhang, 2003). Yet, due to the study's balanced incomplete booklet design, other tests, like the M_2 – test (Maydeu-Olivares and Joe, 2006) or the Hausmann test (Ranger and Much, 2020), cannot be used as they require the data to be full rank (Zhao, 2006).

Furthermore, the reliance on dichotomous items may constrain the instrument's ability to capture the nuanced understanding of the assessed teaching practices, since passively identifying correct strategies is inherently easier than actively generating them (Klemenz and König, 2019). Finally, to maintain high ecological validity, no items were excluded based on their discrimination parameters (see Supplementary material). While this approach preserved the content across all dimensions of CM, the differential weighting of items based on their discrimination in the 2PL model may impact reliability.

6.3 Future directions and practical implications

Future work could explore different methodological approaches to broaden the test's applicability. For example, if only two or three strategies for managing disruptions are recalled, it might indicate an insufficient preparation for dealing with the multifaceted challenges encountered in CM (Baier-Mosch and Kunter, 2024). Integrating multiple knowledge elements or adopting different perspectives would increase cognitive complexity (Klemenz and König, 2019). This could be implemented through items that require evaluating or ranking different instructional strategies or by incorporating open-ended questions could encourage participants to actively demonstrate their knowledge. While coding open-ended responses could rely on our previously content-validated criteria, such a procedure reduces scalability and practicality. Nevertheless, since we seek to maintain a clear distinction between the measurement of the facets of knowledge and PID, any adaptations to the test should be made with careful consideration to ensure that blurring of these facets is intentional. At the same time, the current test is positioned as a complementary tool alongside contextualized approaches, particularly when aiming to predict performance outcomes.

Practical implications build on a growing consensus that systematically accumulated evidence on "what works" should inform both the creation of measurement instruments and the design, implementation, and evaluation of teacher education programs (Hill et al., 2024). This test contributes to this broader effort by aligning assessment and real-world teaching demands. Aligned with a shared framework of key teaching practices (referred to as "core practices"; Grossman et al., 2009), such tools and measures can help disentangle the specific effects of more complex, real-world interventions that combine effective elements such as video-based feedback, peer coaching, or direct instruction (cf. Wilkinson et al., 2020). Future research should continue to expand these efforts to additional competence areas. In doing so, these instruments may not only enhance evaluation and accountability in teacher education but also foster stronger synergies between research and practice (Hill et al., 2024; Baumgartner et al., in revision).

7 Conclusion

In conclusion, this study contributes to the field of teacher education research by further validating a recently developed instrument that assesses professional knowledge in CM within PE. While the test demonstrates solid psychometric properties and construct validity, further refinements could enhance its capacity to differentiate among varying levels of participant ability and to capture more complex aspects of teacher knowledge. By emphasizing declarative knowledge on evidence-based practices and maintaining a specific focus on CM in PE, the instrument establishes a foundation for aligning additional assessment of distinct facets of competence in this area. For example, within the SNSF project "WiPe-Sport," this test will be used alongside a PID test and an observational rating instrument to provide a more comprehensive understanding of the types of knowledge and skills teachers need to improve their CM-related performance effectively.

The development and validation of this instrument serve as an example of centering teacher education research around practical demands. Creating assessment instruments that bridge the gap between theoretical knowledge and teaching practices can foster stronger synergies between research and practice, and strengthen evaluation and accountability in teacher education.

Data availability statement

The data presented in this study can be found at: https://doi.org/10.48573/j4bn-xr96.

Ethics statement

Ethical approval was not required for the studies involving humans because this study was planned and conducted in accordance with the ethical requirements of the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

CB: Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. EJ: Resources, Methodology, Writing – original draft, Writing – review & editing.

MB: Conceptualization, Funding acquisition, Methodology, Project administration, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research is funded by the Swiss National Science Foundation (SNSF). The grant 192397 has been awarded to MB (Pädagogische Hochschule St. Gallen). For further information see: https://p3.snf.ch/project-192397.

Acknowledgments

The authors want to thank Prof. Dr. Jan Hochweber and Substitue Professor Dr. Alexander Naumann for their support during data analysis.

Conflict of interest

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

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. This manuscript utilized ChatGPT, versions 4o and o1, for language editing to enhance clarity, coherence, and overall readability. The AI-assisted editing focused on refining sentence structure, improving phrasing, and ensuring academic rigor, while maintaining the integrity of the original content.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1570510/full#supplementary-material

References

Albu, C., and Lindmeier, A. (2023). Performance assessment in teacher education research—a scoping review of characteristics of assessment instruments in the DACH region. *Zeitschrift für Erziehungswissenschaft* 26, 751–778. doi: 10.1007/s11618-023-01167-7

Aloe, A. M., Amo, L. C., and Shanahan, M. E. (2014). Classroom management self-efficacy and burnout: a multivariate meta-analysis. *Educ. Psychol. Rev.* 26, 101–126. doi: 10.1007/s10648-013-9244-0

Arbogast, G., and Chandler, J. P. (2005). Class management behaviors of effective physical educators. *Strategies* 19, 7–11. doi: 10.1080/08924562.2005.11000381

Baier-Mosch, F., and Kunter, M. (2024). Pre-service teachers' knowledge about classroom management from university studies and own schooling experiences—Content and effects of their activation. *Front. Educ.* 9:1365005. doi: 10.3389/feduc.2024.1365005

Bastian, A., König, J., Weyers, J., Siller, H.-S., and Kaiser, G. (2024). Effects of teaching internships on preservice teachers' noticing in secondary mathematics education. *Front. Educ.* 9:1360315. doi: 10.3389/feduc.2024.1360315

Bauer, J., and Kollar, I. (2023). (Wie) kann die Nutzung bildungswissenschaftlicher Evidenz Lehren und Lernen verbessern? Thesen und Fragen zur Diskussion um evidenzorientiertes Denken und Handeln von Lehrkräften. *Unterrichtswissenschaft* 51, 123–147. doi: 10.1007/s42010-023-00166-1

Baumgartner, M. (2018). "… Kompetenz ohne Performanz ist leer! Performanz ohne Kompetenz ist blind …!" Zu einem integrativen Kompetenz-Strukturmodell von posrtlehrkräften. *Zeitschrift für sportpädagogische Forschung* 6, 49–68. doi: 10.5771/2196-5218-2018-1-49

Baumgartner, M. (2022). Professional competence(s) of physical education teachers: terms, traditions, modelling and perspectives. *Germ. J. Exerc. Sport Res.* 52, 550–557. doi: 10.1007/s12662-022-00840-z

Baumgartner, M., Berthold, C., and Jeisy, E. (in revision). The effectiveness of different interventions on the classroom management-related performance of pre-service teachers in physical education: improvement of knowledge and situation-specific skills alone is not sufficient.

Baumgartner, M., Jeisy, E., and Berthold, C. (2023). From knowledge to performance in physical teacher education: a Delphi study and a pretest for the content validation of the test instruments. *Swiss J. Educ. Res.* 45, 151–163. doi: 10.24452/sjer.45.2.6

Baumgartner, M., Oesterhelt, V., and Reuker, S. (2020). Konstruktion und Validierung eines multidimensionalen Beobachtungsinstruments zur Erfassung der klassenführungsbezogenen Performanzen von sportunterrichtenden Lehrkräften (KlaPe-Sport). *Ger. J. Exerc. Sport Res.* 50, 511–522. doi: 10.1007/s12662-020-00675-6

Blömeke, S., Gustafsson, J.-E., and Shavelson, R. J. (2015a). Beyond dichotomies: competence viewed as a continuum. *Z. Psychol.* 223, 3–13. doi: 10.1027/2151-2604/a000194

Blömeke, S., Jentsch, A., Ross, N., Kaiser, G., and König, J. (2022). Opening up the black box: teacher competence, instructional quality, and students' learning progress. *Learn. Instr.* 79:101600. doi: 10.1016/j.learninstruc.2022.101600

Blömeke, S., König, J., Suhl, U., Hoth, J., and Döhrmann, M. (2015b). Wie situationsbezogen ist die Kompetenz von Lehrkräften? Zur Generalisierbarkeit der Ergebnisse von videobasierten Performanztests. *Zeitschrift für Pädagogik* 61, 310–327. doi: 10.25656/01:15350

Bond, T. G., Yan, Z., and Heene, M. (2021). Applying the Rasch model: Fundamental measurement in the human sciences. *Fourth* Edn. New York: Routledge Taylor & Francis Group.

Boshuizen, H. P. A., Gruber, H., and Strasser, J. (2020). Knowledge restructuring through case processing: the key to generalise expertise development theory across domains? *Educ. Res. Rev.* 29:100310. doi: 10.1016/j.edurev.2020.100310

Brophy, J. E. (2006). "History of research on classroom management" in Handbook of classroom management: Research, practice, and contemporary issues. eds. C. M. Evertson and C. E. Weinstein (New York: Lawrence Erlbaum Associates Publishers), 17–43.

Brühwiler, C., and Hollenstein, L. (2021). The contextualised measuring of general pedagogical knowledge and skills: exploring the use of knowledge in practice. In H. Ulferts Teaching as a knowledge profession: studying pedagogical knowledge across education systems. OECD. Available at: https://doi.org/10.1787/e823ef6e-en (Accessed June 3, 2025).

Brühwiler, C., Hollenstein, L., Affolter, B., Biedermann, H., and Oser, F. (2017). Welches Wissen ist unterrichtsrelevant?: Prädiktive Validität dreier Messinstrumente zur Erfassung des pädagogisch-psychologischen Wissens von Lehrpersonen. *Zeitschrift für Bildungsforschung* 7, 209–228. doi: 10.1007/s35834-017-0196-1

Burns, E. C., Martin, A. J., Collie, R. J., and Mainhard, T. (2021). Perceived classroom disruption undermines the positive educational effects of perceived need-supportive teaching in science. *Learn. Instr.* 75:101498. doi: 10.1016/j.learninstruc.2021.101498

Charalambous, C. Y. (2020). Reflecting on the troubling relationship between teacher knowledge and instructional quality and making a case for using an animated teaching simulation to disentangle this relationship. *ZDM* 52, 219–240. doi: 10.1007/s11858-019-01089-x

Cook, B. G., Smith, G. J., and Tankersley, M. (2012). Evidence-based practices in education. In K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra and J. Sweller (Eds.), APA educational psychology handbook, Vol 1: Theories, constructs, and critical issues. 495–527. American Psychological Association. doi: 10.1037/13273-017

Cothran, D., and Kulinna, P. (2015). Classroom management in physical education. In E. T. Emmer and E. J. Sabornie (Eds.), Handbook of classroom management (2nd ed., pp. 239–260). New York: Routledge.

Depaepe, F., and König, J. (2018). General pedagogical knowledge, self-efficacy and instructional practice: disentangling their relationship in pre-service teacher education. *Teach. Teach. Educ.* 69, 177–190. doi: 10.1016/j.tate.2017.10.003

Dicke, T., Elling, J., Schmeck, A., and Leutner, D. (2015). Reducing reality shock: the effects of classroom management skills training on beginning teachers. *Teach. Teach. Educ.* 48, 1–12. doi: 10.1016/j.tate.2015.01.013

Dückers, C., Hörter, P., Junker, R., and Holodynski, M. (2022). Professional vision of teaching as a focus-specific or focus-integrated skill – conceptual considerations and video-based assessment. *Teach. Teach. Educ.* 117:103797. doi: 10.1016/j.tate.2022.103797

Emmer, E. T., and Stough, L. M. (2001). Classroom management: a critical part of educational psychology, with implications for teacher education. *Educ. Psychol.* 36, 103–112. doi: 10.1207/S15326985EP3602_5

Fenstermacher, G. D. (1994). The knower and the known: the nature of knowledge in research on teaching. *Rev. Res. Educ.* 20, 3–56. doi: 10.3102/0091732X020001003

Freiberg, H. J., and Lamb, S. M. (2009). Dimensions of person-centered classroom management. *Theory Pract.* 48, 99–105. doi: 10.1080/00405840902776228

Freiberg, H. J., Oviatt, D., and Naveira, E. (2020). Classroom management meta-review continuation of research-based programs for preventing and solving discipline problems. *J. Educ. Stud. Placed Risk* 25, 319–337. doi: 10.1080/10824669.2020.1757454

Frey, A., Hartig, J., and Rupp, A. A. (2009). An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice* 28, 39–53. doi: 10.1111/j.1745-3992.2009.00154.x

Gegenfurtner, A., Gruber, H., Holzberger, D., Keskin, Ö., Lehtinen, E., Seidel, T., et al. (2023). "Towards a cognitive theory of visual expertise: methods of inquiry" Damşa, C., Rajala, A., Ritella, G., & Brouwer, J. (Eds.). (2023). *Re-theorising Learning and Research Methods in Learning Research*. (1st ed.). Routledge. 142–158. doi: 10.4324/9781003205838

Giessing, J. (2010). Mannschaften und Gruppen im Sport—Unterricht einteilen. *Sport Praxis* 51, 19–22.

Gold, B., and Holodynski, M. (2015). Development and construct validation of a situational judgment test of strategic knowledge of classroom management in elementary schools. *Educ. Assess.* 20, 226–248. doi: 10.1080/10627197.2015.1062087

Gold, B., Junker, R., Wissemann, M., Klassen, C., and Holodynski, M. (2021). Are good observers good classroom managers? The relationship between teachers' professional vision and their students' ratings on classroom management. *Int. J. Educ. Res.* 109:101811. doi: 10.1016/j.ijer.2021.101811

Grossman, P., Hammerness, K., and McDonald, M. (2009). Redefining teaching, re-imagining teacher education. *Teach. Teach.* 15, 273–289. doi: 10.1080/13540600902875340

Guerriero, S. (ed.) (2017). Pedagogical Knowledge and the Changing Nature of the Teaching Profession. Paris: Publishing, Paris. doi: 10.1787/9789264270695-en (Accessed June 3, 2025).

Hattie, J. (2010). Visible learning: A synthesis of over 800 meta-analyses relating to achievement (Reprinted). Routledge: London.

Heins, J., and Zabka, T. (2019). Mentale Prozesse bei der Unterrichtsbeobachtung. Theoretische Klärungen und ein Fallbeispiel zum Literaturunterricht. *Zeitschrift für Pädagogik* 65, 904–925. doi: 10.25656/01:24155

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

Hill, H. C., Mancenido, Z., and Loeb, S. (2024). Effectiveness research for teacher education. *Educ. Res.* 53, 370–377. doi: 10.3102/0013189X241260393

Hummel, A., and Krüger, M. (2015). Rituale im Schulsport. Einführung in das Themenheft. *Sportunterricht* 64, 34–35.

IDES. (n.d.) Swiss education system. Available online at: https://www.edk.ch/en/education-system-ch?set_language=en (accessed January 20, 2025).

Jeisy, E., Berthold, C., and Baumgartner, M. (in prep.). Validation of a video-based test instrument to assess preservice teacher's situation-specific skills in classroom management in physical education.

Jeschke, C., Kuhn, C., Lindmeier, A., Zlatkin-Troitschanskaia, O., Saas, H., and Heinze, A. (2019). Performance assessment to investigate the domain specificity of instructional skills among pre-service and in-service teachers of mathematics and economics. *Br. J. Educ. Psychol.* 89, 538–550. doi: 10.1111/bjep.12277

Junker, R., Gold, B., and Holodynski, M. (2021). Classroom management of pre-service and beginning teachers: from dispositions to performance. *Int. J. Mod. Educ. Stud.* 5, 339–363. doi: 10.51383/ijonmes.2021.137

Kaiser, G., Blömeke, S., König, J., Busse, A., Döhrmann, M., and Hoth, J. (2017). Professional competencies of (prospective) mathematics teachers—cognitive versus situated approaches. *Educ. Stud. Math.* 94, 161–182. doi: 10.1007/s10649-016-9713-8

Kleickmann, T. (2023). Acht Irrtümer über Klassenführung. *Pädagogik* 1, 22–25. doi: 10.3262/PAED2301022

Klemenz, S., and König, J. (2019). Modellierung von Kompetenzniveaus im pädagogischen Wissen bei angehenden Lehrkräften. Zur kriterialen Beschreibung von Lernergebnissen der fächerübergreifenden Lehramtsausbildung. *Zeitschrift für Pädagogik* 65, 355–378. doi: 10.25656/01:23947

Knogler, M., Hetmanek, A., and Seidel, T. (2022). "Bestimmung und Bereitstellung der "best available" Evidenz für bestimmte Praxisfelder im Bildungsbereich" in Optimierung schulischer Bildungsprozesse—What works? eds. N. McElvany, M. Becker, F. Lauermann, H. Gaspard and A. Ohle-Peters. (Münster: Waxmann), 135–144.

König, J. (2014). Designing an international instrument to assess teachers' general pedagogical knowledge (GPK): review of studies, considerations, and recommendations. Technical paper prepared for the ITEL project. (OECD, Ed.). Available online at: https://web-archive.oecd.org/2019-03-28/513002-Assessing%20Teachers'%20General%20Pedagogical%20Knowledge.pdf (Accessed June 3, 2025).

König, J. (2023). Lehrer:innenexpertise und Lehrer:innenkompetenz. In M. Rothland (M. Rothland Ed.), *Beruf Lehrer:in* (2. aktual. u. erw. Aufl.), pp. 148–171). Waxmann. Available online at: https://elibrary.utb.de/doi/10.36198/9783838588216-148-171.

König, J., Blömeke, S., Jentsch, A., Schlesinger, L., née, C., Musekamp, F., et al. (2021). The links between pedagogical competence, instructional quality, and mathematics achievement in the lower secondary classroom. *Educ. Stud. Math.* 107, 189–212. doi: 10.1007/s10649-020-10021-0

König, J., Blömeke, S., Paine, L., Schmidt, W. H., and Hsieh, F.-J. (2011). 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.* 62, 188–201. doi: 10.1177/0022487110388664

König, J., and Kramer, C. (2016). Teacher professional knowledge and classroom management: on the relation of general pedagogical knowledge (GPK) and classroom management expertise (CME). *ZDM* 48, 139–151. doi: 10.1007/s11858-015-0705-4

König, J., Ligtvoet, R., Klemenz, S., and Rothland, M. (2024). Discontinued knowledge growth: on the development of teachers' general pedagogical knowledge at the transition from higher education into teaching practice. *Teach. Teach.* 1–19, 1–19. doi: 10.1080/13540602.2024.2308895

König, J., and Pflanzl, B. (2016). Is teacher knowledge associated with performance? On the relationship between teachers' general pedagogical knowledge and instructional quality. *Eur. J. Teach. Educ.* 39, 419–436. doi: 10.1080/02619768.2016.1214128

König, J., and Rothland, M. (2016). Klassenführungswissen als Ressource der Burnout-Prävention? Zum Nutzen von pädagogisch-psychologischem Wissen im Lehrberuf. *Unterrichtswissenschaft* 44, 425–441.

König, J., and Seifert, A. (2012). Lehramtsstudierende erwerben pädagogisches Professionswissen: Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung. New York: Waxmann. doi: 10.25656/01:21029

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., and Doolaard, S. (2016). 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.* 86, 643–680. doi: 10.3102/0034654315626799

Kunter, M., Baumert, J., and Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learn. Instr.* 17, 494–509. doi: 10.1016/j.learninstruc.2007.09.002

Landrum, T. J., and Kauffman, J. M. (2006). Behavioral Approaches to Classroom Management. I Handbook of classroom management: Research, practice, and contemporary issues. eds. C. M. Evertson and C. E. Weinstein (New York: Lawrence Erlbaum Associates Publishers), 47–71.

Leijen, Ä., Malva, L., Pedaste, M., and Mikser, R. (2022). What constitutes teachers' general pedagogical knowledge and how it can be assessed: a literature review. *Teach. Teach.* 28, 206–225. doi: 10.1080/13540602.2022.2062710

Leijen, Ä., Pedaste, M., Baucal, A., Poom-Valickis, K., and Lepp, L. (2024). What predicts instructional quality and commitments to teaching: self-efficacy, pedagogical knowledge or integration of the two? *Front. Psychol.* 15:1287313. doi: 10.3389/fpsyg.2024.1287313

Lenske, G., Thillmann, H., Wirth, J., Dicke, T., and Leutner, D. (2015). Pädagogisch-psychologisches Professionswissen von Lehrkräften: Evaluation des ProwiN-Tests. *Zeitschrift für Erziehungswissenschaft* 18, 225–245. doi: 10.1007/s11618-015-0627-5

Lenske, G., Wagner, W., Wirth, J., Thillmann, H., Cauet, E., Liepertz, S., et al. (2016). Die Bedeutung des pädagogisch-psychologischen Wissens für die Qualität der Klassenführung und den Lernzuwachs der Schüler/innen im Physikunterricht. *Zeitschrift für Erziehungswissenschaft* 19, 211–233. doi: 10.1007/s11618-015-0659-x

LimeSurvey GmbH (n.d.) LimeSurvey: an open source survey tool LimeSurvey GmbH. Available online at: https://www.limesurvey.org (Accessed June 3, 2025).

Lüders, M. (2012). "Pädagogisches Unterrichtswissen" – eine Testkritik. *Zeitschrift für Erziehungswissenschaft* 15, 775–791. doi: 10.1007/s11618-012-0329-1

Marzano, R. J., Marzano, J. S., and Pickering, D. J. (2003). Classroom management that works: Research-based strategies for every teacher. Association for Supervision and Curriculum Development: Alexandria, Virginia USA.

Maydeu-Olivares, A., and Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika* 71, 713–732. doi: 10.1007/s11336-005-1295-9

Moosbrugger, H., and Kelava, A. (Eds.) (2020). (Testtheorie und Fragebogenkonstruktion. (3rd ed.). Springer. doi: 10.1007/978-3-662-61532-4

Müller, M. M., and Gold, B. (2022). Videobasierte Erfassung wissensbasierten Verarbeitens als Teilprozess der professionellen Unterrichtswahrnehmung – Analyse eines geschlossenen und offenen Verfahrens. *Zeitschrift für Erziehungswissenschaft* 26, 7–29. doi: 10.1007/s11618-022-01128-6

Nielsen, T., and Dammeyer, J. (2019). Measuring higher education students' perceived stress: an IRT-based construct validity study of the PSS-10. *Stud. Educ. Eval.* 63, 17–25. doi: 10.1016/j.stueduc.2019.06.007

Oliver, R. M., Wehby, J. H., and Reschly, D. J. (2011). Teacher classroom management practices: effects on disruptive or aggressive student behavior. *Campbell Syst. Rev.* 7, 1–55. doi: 10.4073/csr.2011.4

Ophardt, D., and Thiel, F. (2013). Klassenmanagement: Ein Handbuch für Studium und Praxis. Kohlhammer Verlag: Stuttgart.

Pollmeier, J., Kleickmann, T., Zimmermann, F., Möller, J., and Köller, O. (2024). Preservice teachers' pedagogical and psychological knowledge: structure and learning opportunities. *Stud. Educ. Eval.* 83:101424. doi: 10.1016/j.stueduc.2024.101424

Prenzel, M. (2020). «Nützlich, praktisch, gut»: Erwartungen an die Forschung in der Lehrerinnen- und Lehrerbildung. *Beiträge zur Lehrerinnen-und Lehrerbildung* 38, 8–20. doi: 10.25656/01:21771

Raith, E. (2017). Geräteaufbau und Geräteabbau in der Grundschule. Ein Lernfeld als didaktische Herausforderung. *Bewegung Sport* 71, 37–41.

Ranger, J., and Much, S. (2020). Analyzing the fit of IRT models with the Hausman test. *Front. Psychol.* 11:149. doi: 10.3389/fpsyg.2020.00149

Renkl, A. (2022). Meta-analyses as a privileged information source for informing teachers' practice?: a plea for theories as primus inter pares. *Zeitschrift für Pädagogische Psychologie* 36, 217–231. doi: 10.1024/1010-0652/a000345

Römer, J., and Rothland, M. (2015). Klassenführung in der deutschsprachigen Unterrichtsforschung. Ein kritischer Überblick zur Operationalisierung und empirischen Erfassung. *Empirische Pädagogik* 29, 266–287.

Ryan, S., and Swartz, D. (2018). Solving the acoustic issue in physical education settings. *Int. J. Phys. Educ. Fit. Sports* 7, 11–16. doi: 10.26524/ijpefs1813

Schlag, S., and Glock, S. (2019). Entwicklung von Wissen und selbsteingeschätztem Wissen zur Klassenführung während des Praxissemesters im Lehramtsstudium. *Unterrichtswissenschaft* 47, 221–241. doi: 10.1007/s42010-019-00037-8

Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educ. Psychol.* 48, 73–86. doi: 10.1080/00461520.2013.779483

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15, 4–14. doi: 10.3102/0013189X015002004

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., and Sugai, G. (2008). Evidence-based practices in classroom management: considerations for research to practice. *Educ. Treat. Child.* 31, 351–380. doi: 10.1353/etc.0.0007

Slavin, R. E. (2002). Evidence-based education policies: transforming educational practice and research. *Educ. Res.* 31, 15–21. doi: 10.3102/0013189X031007015

Smith, K. (2024). Evidence-based, evidence-informed or evidence-ignored teacher education? The role of research in teacher education. In V. Symeonidis, Enhancing the value of teacher education research: Implications for policy and practice (23–41). BRILL. Available at: https://doi.org/10.1163/9789004689992.

Söll, W., and Kern, U. (1999). Alltagsprobleme des Sportunterrichts. Hofmann: Schoerndorf.

Stone, C. A., and Zhang, B. (2003). Assessing goodness of fit of item response theory models: a comparison of traditional and alternative procedures. *J. Educ. Meas.* 40, 331–352. doi: 10.1111/j.1745-3984.2003.tb01150.x

Stokking, K., Leenders, F., De Jong, J., and Van Tartwijk, J. (2003). From student to teacher: Reducing practice shock and early dropout in the teaching profession. *Eur. J. Teach. Educ.* 26, 329–350. doi: 10.1080/0261976032000128175

Ulferts, H. (2019). The relevance of general pedagogical knowledge for successful teaching: Systematic review and meta-analysis of the international evidence from primary to tertiary education", *OECD Education Working Papers*. Paris: OECD Publishing. doi: 10.1787/ede8feb6-en (Accessed June 3, 2025).

van der Mars, H., Darst, P., Vogler, B., and Cusimano, B. (1994). Active supervision patterns of physical education teachers and their relationship with student behaviors. *J. Teach. Phys. Educ.* 14, 99–112. doi: 10.1123/jtpe.14.1.99

Voss, T., Kunina-Habenicht, O., Hoehne, V., and Kunter, M. (2015). Stichwort Pädagogisches Wissen von Lehrkräften: Empirische Zugänge und Befunde. *Zeitschrift für Erziehungswissenschaft* 18, 187–223. doi: 10.1007/s11618-015-0626-6

Voss, T., Kunter, M., and Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: test construction and validation. *J. Educ. Psychol.* 103, 952–969. doi: 10.1037/a0025125

Voss, T., Kunter, M., Seiz, J., Hoehne, V., and Baumert, J. (2014). Die Bedeutung des pädagogisch-psychologischen Wissens von angehenden Lehrkräften für die Unterrichtsqualität. *Zeitschrift für Pädagogik* 60, 184–201. doi: 10.25656/01:14653

Voss, T., Zachrich, L., Fauth, B., and Wittwer, J. (2022). The same yet different? Teaching quality differs across a teacher's classes, but teachers with higher knowledge make teaching quality more similar. *Learn. Instr.* 80:101614. doi: 10.1016/j.learninstruc.2022.101614

Weber, K. E., Neuber, K., and Prilop, C. N. (2023). Videobasierte Reflexion von klassenführungsspezifischen Ereignissen – Welche Rolle spielen Wissen und Reflexionsbereitschaft von Lehramtsstudierenden? *Zeitschrift für Erziehungswissenschaft* 26, 1235–1257. doi: 10.1007/s11618-023-01195-3

Weyers, J., König, J., Santagata, R., Scheiner, T., and Kaiser, G. (2023). Measuring teacher noticing: a scoping review of standardized instruments. *Teach. Teach. Educ.* 122:103970. doi: 10.1016/j.tate.2022.103970

Weyers, J., Ligtvoet, R., and König, J. (2024). How does pre-service teachers' general pedagogical knowledge develop during university teacher education? Examining the impact of learning opportunities and entry characteristics over five time points using longitudinal models. *J. Curric. Stud.* 56, 448–467. doi: 10.1080/00220272.2024.2355923

Wilkes, T., and Stark, R. (2022). Probleme evidenzorientierter Unterrichtspraxis: Anregungen und Lösungsvorschläge. *Unterrichtswissenschaft* 51, 289–313. doi: 10.1007/s42010-022-00150-1

Wilkinson, S., Freeman, J., Simonsen, B., Sears, S., Byun, S. G., Xu, X., et al. (2020). Professional development for classroom management: A review of the literature. *Educ Res Eval.* 26, 182–212. doi: 10.1080/13803611.2021.1934034

Wolters, P. (2021). "Bankdrücker" als Herausforderung für Sportlehrkräfte. *Sportunterricht* 70, 538–543. doi: 10.30426/SU-2021-12-2

Zhao, Y. (2006). Approaches for addressing the fit of item response theory models to educational test data. [University of Massachusetts Amherst]. Available online at: https://doi.org/10.7275/18739815 (Accessed June 3, 2025).

Zlatkin-Troitschanskaia, O., Förster, M., Preuße, D., and Mater, O. (2016). The relationship between teachers' evidence-based actions and communication, cooperation, and participation structures at schools. *J. Educ. Res. Online* 8, 59–79. doi: 10.25656/01:12806