Studies in Educational Evaluation 81 (2024) 101357

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Measuring the quality of teacher professional development: A large-scale validation study of an 18-item instrument for daily use

Eric Richter [*,1], Dirk Richter [2]

Department of Education, University of Potsdam, Potsdam, Germany

ARTICLE INFO

Keywords:
Teacher professional development
Courses
Quality assessment
Instrument validation
Educational quality

ABSTRACT

This study introduces the Teacher Professional Development (TPD) Monitor, an 18-item instrument designed to assess the quality of formal learning opportunities (e.g. courses, workshops), for in-service teachers across all grade levels of general education. By evaluating these opportunities across four critical dimensions—clarity and structure, cognitive activation, collaboration, practical relevance—the TPD Monitor provides a comprehensive framework for assessing the effectiveness of professional development programs. Developed and validated with data from 2314 in-service teachers actively teaching in primary, secondary, special education, and vocational schools, each participating in one of 173 distinct TPD courses, this instrument is supported by multi-level confirmatory factor analysis that confirmed its structure, offering robust evidence of its validity and reliability. Furthermore, measurement invariance was achieved, allowing for meaningful comparisons between subgroups. The TPD Monitor is a valuable and practical tool enabling TPD providers to design effective courses and allowing to monitor and improve the quality of TPD programs.

1. Introduction

Effective teacher professional development (TPD) can play a pivotal role in transforming schools by fostering innovative teaching approaches and facilitating student growth. Its significance lies in its capacity to enhance teachers' professional knowledge, skills, and classroom practices, generating tangible improvements in student learning outcomes. TPD effectiveness is defined as organized professional learning that leads to changes in teacher practice and improvements in student learning outcomes (Darling-Hammond, Hyler, & Gardner, 2017).

The quality of TPD is crucial for its effectiveness (Sims, Fletcher-Wood, O'Mara-Eves, & Cottingham, 2021). Extensive research has been conducted to identify the features of effective TPD, resulting in lists of characteristics that include active learning and content focus (e.g., Darling-Hammond, Hyler, & Gardner, 2017). TPD quality has emerged from this literature as a multifaceted concept that can be described through various practices of facilitators and instructors. These practices are deemed significant for fostering high-quality instructional processes in TPD courses and are positively associated with desired outcomes, including increased teacher knowledge, skills, and long-term professional success.

Although ample research has investigated the characteristics of high-quality TPD, there is still a lack of standardized measurement instruments capable of precisely assessing TPD quality. To enhance the impact of TPD programs, it is crucial to develop an instrument that enables accurate measurement and evaluation of TPD quality.

The present article addresses this desideratum and seeks to fill the current research gap. Our primary aim in this study was to validate a recently developed instrument for assessing TPD quality by testing the reliability and validity of the instrument and its constituent constructs with the purpose of obtaining objective data on TPD quality. The validated instrument should allow thorough evaluation of the multifaceted aspects of TPD quality, leading to a more comprehensive understanding of TPD effectiveness.

The process of validating the instrument involved a rigorous empirical investigation encompassing a sample of 173 TPD courses with a total of 2314 participants. We used the data to evaluate the validity and reliability of the instrument so that it could be used to assess and measure TPD quality in different contexts. Validating an assessment

* Correspondence to: University of Potsdam, Department of Education, Karl-Liebknecht-Straße 24–25, 14476 Potsdam, Germany.
E-mail address: eric.richter@uni-potsdam.de (E. Richter).
[1] https://orcid.org/0000-0003-3572-2022
[2] https://orcid.org/0000-0002-2384-1588

https://doi.org/10.1016/j.stueduc.2024.101357
Received 17 October 2023; Received in revised form 20 March 2024; Accepted 2 April 2024
Available online 6 April 2024

instrument to measure TPD quality is a vital undertaking, aimed at identifying areas requiring improvement and implementing measures to address weaknesses, thereby elevating the overall quality of TPD courses, and ensuring their efficacy and effectiveness. This process ensures that TPD offerings are aligned with participants' needs and leave a lasting impact on their professional practice. Ultimately, this validation effort will play a vital role in advancing the field of TPD and facilitating continuous improvement in the design and provision of high-quality TPD experiences.

1.1. Perspectives on TPD quality

Our review of the research on TPD quality is based on a definition of TPD as a "structured, facilitated activity for teachers intended to increase their teaching ability" (Sims et al., 2021, p. 7). This encompasses formal courses and workshops for all active members of the teaching workforce, emphasizing the inclusivity of our approach by specifying that it includes in-service teachers across primary, secondary, special, and vocational schools, catering to both beginning and experienced teachers. The term "quality," however, has a wide array of interpretations and is subject to diverse perspectives in educational research (Adams, 1993; Adams, Acedo, & Popa, 2012). Scheerens, Luyten, & van Ravens (2011a) use the concept of educational quality to describe educational processes and outcomes at different levels, including the system, school, and classroom. In this context, education is often understood as a productive system in which inputs are transformed into desired outcomes (Scheerens et al., 2011a), and educational quality is understood as the success of the system, which depends on the attainment of the desired outputs or outcomes (Organisation for Economic Cooperation and Development, 2005).

An alternative perspective on educational quality is associated with instrumental effectiveness (Scheerens, 2004; Scheerens, Luyten, & van Ravens, 2011a,b). In this view, quality is not defined solely by the success of a system, but also by the educational processes and their anticipated influence on outcomes. This understanding of educational quality suggests that the output of education can be specified and predicted by process indicators. From this perspective, process indicators have the potential to replace outcome indicators in describing educational quality. They are also considered highly relevant to educational policy and practice, as they are regarded as malleable characteristics that can be modified and improved (Scheerens, Luyten, & van Ravens, 2011a,b) .

This understanding of the quality of TPD as relating to the process rather than the outcome is also evident in influential works on teacher learning and teacher education. For instance, Darling-Hammond, Hyler, and Gardner (2017) assert that effective TPD is "structured professional learning that results in changes in teacher practices and improvements in student learning outcomes." Effective TPD is therefore not defined solely by the resulting changes, but above all by the organization and structure of the learning experience—the process of interaction between facilitators and participating teachers and among the participating teachers.

Building upon this foundation, we recognize that a process-oriented approach to understanding quality is not confined to the specific context of TPD, extending to broader educational domains. This is articulated through examples in the field of K-12 general education, where instructional quality is similarly recognized as being processual. The foundation for this notion can be traced back to Caroll (1963), who follows a process-product paradigm (Brophy, 2000; Shuell, 2001). Instructional quality, within this framework, encompasses all the actions of teachers within the classroom, with a specific focus on the instructional practices they employ to accomplish specific instructional tasks (the process). These practices play a vital role in optimizing student learning outcomes (the product) (Goe & Stickler, 2008; Senden, Nilsen, & Blömeke, 2021). Moreover, this process-oriented understanding of quality also serves as a basis for research in higher education on the quality of university teaching, as demonstrated by studies such as Abrami, d'Apollonia, and Rosenfield (2007), illustrating the widespread applicability of a process-oriented perspective.

The inclusion of evidence from both K-12 and higher education research highlights the universality of the process-oriented perspective in assessing educational quality. By leveraging these broader strands of research, we further our understanding of TPD quality, underscoring that effective teaching and learning principles are consistent across all levels of education. This comparative perspective not only reaffirms the applicability of a process-oriented understanding across different educational contexts but also strengthens the theoretical foundation for our study of TPD quality.

The use of quantifiable indicators is essential in evaluating the process quality of TPD. These process indicators hold great relevance for educational policy and practice, as they are considered flexible attributes that are strongly associated with higher levels of educational achievement (Scheerens, Luyten, & van Ravens, 2011a,b). Over the years, research on TPD has focused on identifying indicators that contribute to the effectiveness of professional development, and has employed various terms to describe them, including "characteristics" (Sims & Fletcher-Wood, 2021), "critical features" (Desimone, 2009; Kraft, Blazar, & Hogan, 2018), "core features" (Garet, Porter, Desimone, Birman, & Yoon, 2001), "design elements" (Darling-Hammond, Hyler, & Gardner, 2017), "design features" (Kennedy, 2006) or "mechanism" (Sims et al., 2021).

A substantial body of literature reviews and meta-analyses has sought to identify process indicators of effective TPD (e.g., Desimone, 2009; Kennedy, 2016; Timperley, Wilson, Barrar, & Fung, 2007; Walter & Briggs, 2012; Wei, Darling-Hammond, Andree, Richardson, & Orphanos, 2009; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). These studies include large- and small-scale investigations, intensive case studies and evaluations of specific approaches to enhancing teaching, and descriptions of "best practices" in TPD based on expert experiences (Garet et al., 2001). They have resulted in the compilation of various process indicators presumed to constitute high-quality TPD (Kennedy, 2016).

However, the research lacks consensus on which process indicators are most meaningful and how they should be evaluated. While some researchers claim that there is broad agreement on process indicators (e.g., Darling-Hammond, Hyler, & Gardner, 2017, Desimone, 2009, Kraft et al., 2018), others criticize the methods used to develop existing indicator lists (e.g., Kennedy, 2016; Sims & Fletcher-Wood, 2021). Sims and Fletcher-Wood (2021), for instance, raise concerns about the inconclusive nature or significant methodological limitations of existing indicator lists, arguing that they lack a clear distinction between causally redundant components of interventions and the "active ingredients" that contribute to improved teaching and student learning (Sims & Fletcher-Wood, 2021).

Providing a comprehensive list of all process indicators that may potentially contribute to positive outcomes in TPD is a challenging endeavor due to the diverse range of indicators identified in the literature. However, research from the United States has highlighted several widely used process indicators of effective TPD. These include content focus, active learning approaches, collaborative structures, models of effective practice, coaching and expert support, as well as feedback and reflection integrated into TPD courses (e.g., Darling-Hammond, Hyler, & Gardner, 2017; Desimone, 2009). These findings have also been referenced extensively in international studies on high-quality TPD (e.g., Kalinowski, Gronostaj, & Vock, 2019; Lipowsky & Rzejak, 2021). Sims et al. (2021), in a recent literature review and meta-analysis, not only included indicators such as practical social support, feedback, and modeling in their list of effective TPD characteristics but also introduced additional indicators. These include managing cognitive load, revisiting material, goal setting, credibility of sources, positive reinforcement, instructional rehearsal, environmental cues, action planning, behavior monitoring, and context-specific repetition.

Building on the approaches proposed by Sims et al. (2021) and Meyer, Kleinknecht, and Richter (2023), it is possible to integrate and map the identified indicators into dimensions of process quality. By combining the indicators from these studies, a comprehensive framework can be developed to assess and evaluate the various facets of process quality in TPD. This mapping approach enables a holistic understanding of the key components that contribute to effective TPD, facilitating the design of targeted interventions and the measurement of TPD course effectiveness.

When revisiting the lists of process indicators, it becomes evident that effective TPD is well-structured, includes activating tasks, provides opportunities for exchange, and establishes connections to participants' professional practice. Based on these observations, we identify four key dimensions of TPD process quality: clarity and structure, practical relevance, cognitive activation, and collaboration. The dimension of *clarity and structure* emphasizes the importance of precise and explicit learning goals, clear explanations of TPD content, efficient time management, and a transparent and logical course structure (e.g., Guskey, 2000). *Practical relevance* encompasses aspects that establish a strong connection between the PD content and participants' professional practice (e.g., Trivette, Dunst, Hamby, & O'Herin, 2009). This includes aligning the PD content with real-life discipline-specific curricula and utilizing practice-based models that are applicable and relevant to real-world teaching situations. *Cognitive activation* involves engaging participants' prior knowledge, incorporating thought-provoking questions and tasks, and providing adaptive feedback and opportunities for reflection on professional activities. Finally, the dimension of *collaboration* emphasizes the promotion of meaningful exchange among participants to foster a collaborative learning community that encourages sharing ideas, experiences, and insights.

The suggested aggregation of process indicators into overarching dimensions of process quality is still tentative, but it is derived from relevant studies on TPD quality. These dimensions can be reasonably assumed to play an important role in the success of TPD. Furthermore, they offer a pragmatic approach for operationalizing and measuring TPD quality.

1.2. Measuring TPD quality

In the realm of educational quality management, regular and standardized evaluations play a central role (Gosling & D'Andrea, 2001; Pratasavitskaya & Stensaker, 2010). Process quality can be assessed using diverse data sources and standardized instruments to understand how instructors, including school and university teachers as well as TPD facilitators, conduct their classroom or course instruction (Praetorius & Charalambous, 2018; Guskey, 2000; O'Hanlon & Mortensen, 1980). Instruments that measure instructional quality often incorporate the perspectives of learners, such as students or TPD participants (Senden et al., 2021). There is relative consensus in the research on instructional quality in both elementary and secondary schools as well as higher education contexts that learners are capable of evaluating the characteristics of learning opportunities (Golding & Adam, 2016; Richardson, 2005; Wagner, Göllner, Helmke, Trautwein, & Lüdtke, 2013).

When applying test instrument-based evaluation in TPD, adherence to established standards is crucial (AERA, 2011; Guskey, 2000; Yarbrough, Shulha, Hopson, & Caruthers, 2010). Psychometrically, these instruments must exhibit objectivity, validity, and reliability (AERA, 2011). Furthermore, TPD evaluations, as a form of course evaluation, should adhere to utility standards (e.g., evaluator credibility), feasibility standards (e.g., cost-effectiveness), propriety standards (e.g., protection of human subjects' rights), and accuracy standards (e.g., ensuring the validity and reliability of information) (Guskey, 2000).

In our literature review on TPD quality, we found numerous studies that aimed to measure TPD quality. These studies can be categorized into two groups: those that lack a clear claim to present a standardized, valid, and reliable instrument capturing TPD quality, and those that explicitly propose standardized, valid, and reliable instruments capturing TPD quality. An example from the first group is the study by Fischer et al. (2018), which explored the link between participation in overall TPD and teachers' classroom practices. In it, variables assessing the qualitative aspects of TPD participation draw inspiration from the work of Desimone (2009) and Darling-Hammond, Hyler, and Gardner (2017) on frameworks of design features for high-quality TPD activities. Fischer et al. (2018) combined these variables using single items to create an overall quality score, providing a comprehensive measure of the TPD quality experienced by teachers. The measurement approach they utilized is simplistic and does not fully capture the complexities of the processes involved in high-quality TPD, but it serves as a basic approximation to address their research question effectively.

Other studies have made efforts to develop evidence-based measurement tools capable of capturing the multidimensional aspects of TPD quality more comprehensively. Although these studies have strengths, they also have weaknesses concerning evaluation standards. In the following, we provide a brief overview of the different approaches to capturing TPD and highlight their limitations.

Gaumer Erickson, Noonan, Brussow, and Supon Carter (2017) introduced the Observation Checklist for High-Quality Professional Development Training (HQPD Checklist). Through a rigorous four-year statewide field test and refinement process, the authors established a six-domain (preparation, introduction, demonstration, engagement, evaluation, mastery), 22-item instrument that utilizes a dichotomous choice format (yes vs. no). They involved diverse stakeholders in the development process, including TPD evaluators, TPD providers, and staff from state departments of education that fund TPD, and found their HQPD Checklist to show high validity. They also conducted an intraclass correlation analysis to assess reliability. While the authors invested considerable effort in piloting the instrument, their investigation lacks crucial information needed to estimate the psychometric quality of the tool. Notably, they did not report any tests of factor structure or provide further information about scale reliability. As a result, additional research may be necessary to thoroughly evaluate the psychometric properties of the HQPD Checklist.

The same is true of a reflection tool presented by Main and Pendergast (2015). Drawing on the so-called five core features of effective TPD (content focus, active learning, coherence, duration, and collective participation; Desimone, 2009; Garet et al., 2001), they developed 38 items to measure TPD quality. Each item was rated on a five-point Likert scale (ranging from strongly disagree to strongly agree). However, the study lacks construct validation, descriptive or correlational analyses of the scales, and statistical evidence on the relationships of the scales with other criteria. As such, further research is needed to establish the psychometric properties and relationships of the scales in the tool.

Drawing from the same theoretical framework, Soine and Lumpe (2014) developed the Characteristics of Teacher Professional Development (CTPD) instrument, comprising 61 items to measure five features of TPD. The investigation employed rigorous data analysis, including exploratory factor analysis, and provided sufficient information about the scales. Additionally, the authors conducted correlational analysis with other criteria, such as data from Washington State's Measures of Student Progress (MSP) or the Skills/knowledge, Thinking, Application, and Relationships (STAR) protocol. Overall, the authors concluded that the CTPD instrument appears to be a viable tool for capturing teacher perceptions of professional development characteristics. However, it should be noted that the instrument's focus is mainly limited to content-specific TPD courses. Moreover, there may be some concerns regarding construct validity. For instance, when assessing the feature "content focus", the instrument includes items such as "gained a deeper understanding of the subject […]" or "learned more about the content […]" (Soine & Lumpe, 2014, p. 322). Such wording indicates an attempt to describe outcome quality, that is, TPD effects, rather than focusing on the process characteristics of a course.

Meyer et al. (2023) offer an instrument developed through a rigorous

validation process. Drawing on the Community of Inquiry framework (Garrison, Anderson, & Archer, 1999; Garrison, Anderson, & Archer, 2001; Garrison, Anderson, & Archer, 2003), they designed eleven items that map the three quality dimensions of clarity and structure, cognitive activation, and collaboration in the context of online TPD. The instrument's structure was validated using confirmatory factor analysis. Additionally, a structural equation model was employed to assess the external validity of the instrument by examining the relationship between online TPD quality attributes, teacher satisfaction, and changes in teachers' professional practice. Although this instrument provides a valid and reliable measure of TPD quality, it is important to note that the entire tool and its validation process are specifically focused on online TPD, which limits its applicability to other types of TPD settings.

To the best of our knowledge, there is currently no comprehensive tool available for measuring TPD quality across various settings, including face-to-face and online formats as well as content-specific and domain-general contexts, that is firmly grounded in a robust theoretical framework and has undergone a rigorous, large-scale validation process.

2. Present study

Assessing the quality of TPD is crucial to ensure teachers have access to high-quality learning opportunities. However, there is a lack of validated evaluation instruments that meet the required standards. This study aimed to address this gap by validating a recently developed TPD quality assessment instrument, focusing on four dimensions: clarity and structure, practical relevance, cognitive activation, and collaboration. In our study, we meticulously acknowledge the inherent multilevel structure of our dataset, with teachers (level 1) nested within PD courses (level 2). This necessitates a nuanced validation process to accurately account for the hierarchical relationships present. Such a multilevel approach is reflective of common research designs in educational studies, where, for instance, students (level 1) are clustered within classes (level 2). This methodology is crucial for exploring dynamics such as instructional quality and learning environments across various educational levels. Similar methodical settings are found in higher education studies, exemplified by research examining instructional quality through data from university students taught in various courses (e.g., Daumiller et al., 2022). However, the applicability of multilevel research extends beyond specific instances and concerns populations with a hierarchical structure.

Traditionally, Level 2 constructs in educational research have been derived from aggregating Level 1 variables, such as student ratings of teaching quality, to form broader classroom-level variables (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011). However, critical reviews (e.g., Hox, Moerbeek, & van de Schoot, 2017; Lüdtke et al., 2011) have highlighted two prevalent issues: the oversight of appropriate levels of analysis and the lack of control for measurement and sampling error. Our study seeks to address these challenges by ensuring a rigorous multilevel analysis that accurately reflects and analyzes the hierarchical data structure. The following hypotheses were tested to validate the instrument:

H1: The theoretical reasoning suggests a four-factor structure, which is expected to demonstrate validity against an aggregate single-factor solution.

H2: Both the TPD courses examined and the teacher respondents within the sample can be considered heterogeneous with respect to a variety of variables such as the format of the course (face-to-face vs. online), and the gender and school type of the respondents. For this reason, the variables were selected to illustrate measurement invariance across as diverse a set of groups as possible. We expect measurement invariance with respect to the variables mentioned to be able to use the instrument for difference measures in further evaluation studies.

H3: The subscales of the newly developed instrument are related to external variables that measure participants' engagement in the TPD courses. A positive relationship with behavioral engagement is expected.

3. Methods

3.1. Study design: instrument development and piloting procedure

The development of the instrument involved stakeholders from educational administration, educational research, and educational practice and aimed at creating a practical, evidence-based tool for educational quality management. The main focus was on standardizing the assessment of process quality from the participants' perspectives, while ensuring the suitability of the instrument for different course formats, including online and face-to-face TPD, as well as general (e.g., lesson planning, classroom management) and content-specific TPD (e.g., use of GeoGebra in elementary-level mathematics education). To ensure practicality, the instrument was designed to be short so that course participants could complete it within ten minutes at the end of a course.

A multi-stage study design was used to create the TPD Monitor, an instrument for assessing the quality of TPD processes. In the first phase, the authors organized a workshop and invited researchers, administrators from the ministry of education, representatives of TPD institutes, TPD facilitators, and different groups of teachers (e.g., elementary teachers, secondary teachers, teachers with disabilities) to discuss the quality of TPD. A key outcome of the workshop was the focus on four overarching dimensions of process quality. These dimensions included the facilitator's ability to present content in a clear and structured way and practical relevance, the importance of opportunities for active learning and collaboration among participants. In a second phase, the authors formed a smaller multidisciplinary group to operationalize the four quality dimensions identified. In an iterative process, items were developed and refined through multiple feedback loops. The items developed through this process were presented to a wider audience for feedback in a second workshop. Suggestions for improvement were actively incorporated, resulting in an 18-item instrument that accurately captures the four dimensions of quality: clarity and structure, practical relevance, cognitive activation, and collaboration. The full instrument can be found in Appendix (Table A). The third phase was to pilot the TPD Monitor from October 2021 to July 2022. Prior to the pilot, facilitators and teacher representatives were informed about the process. A random sample of 286 TPD courses was drawn from the course catalogues of state TPD institutions, covering different course topics and target groups (e.g., subject-specific teachers or all teachers). Different course formats were considered, including face-to-face and online. The TPD Monitor was administered via an online questionnaire completed by TPD participants at the end of the course. Facilitators provided access to the survey and participation was voluntary, in accordance with all data protection and ethical regulations. Approval to conduct the study was obtained from the Ministry of Education.

3.2. Sample

During the pilot phase, participants in 286 TPD courses were invited by their course facilitators to participate in the survey. Due to the voluntary nature of the survey and factors such as course cancellations (e.g., due to facilitator illness or low registration), the survey was ultimately conducted in a subset of 200 of the originally targeted 286 TPD courses. Of the 86 courses that were not included in the study, 36 were cancelled. For the other 50 courses, the course facilitators did not offer participants the chance to evaluate the course for this study. The reasons for this decision remain unknown. Of the 200 TPD courses that participated in the study, a total of 2381 participants responded to the TPD Monitor survey. All participants were in-service teachers actively teaching across a variety of educational settings at the time of the study. The courses varied significantly in size, with an average of 11.91 participants per course ($SD = 8.06$), ranging from a minimum of 1 to a maximum of 71 participants. To address methodological considerations and ensure robustness in our analysis, we adjusted the dataset to establish a minimum cluster size of 5. This adjustment resulted in a

revised sample size of 2314 participants across 173 TPD courses, with an updated average of 13.37 participants per course (SD = 7.67).

As part of the survey, participants were asked to provide personal information about their gender, job experience, and the type of school in which they were currently employed. The results showed that 70.5% identified as female, 28.7% as male, and 0.8% as non-binary. In terms of job experience, most participants (40.4%) had between 5 and 15 years of job experience as a teacher; 36.0% had more than 15 years of experience; and 23.6% had fewer than 5 years. Teachers from all types of schools in the German school system participated in the survey, with secondary school teachers representing the majority (39.5%). Other school types strongly represented were elementary school (30.1%) and vocational school (22.3%).

Of the 173 courses evaluated, 68.6% were face-to-face, 27.3% were online, and 4.1% were in a hybrid format. The format of two courses is unknown (Table 1). In addition, 25.7% of the TPD courses were open to teachers from all types of schools, while 32.2% targeted secondary school teachers. The remaining courses were divided between teachers at elementary schools and teachers at schools for children with special needs.

3.3. Measures

The study assessed the socio-demographic characteristics (age, gender, teaching experience) of TPD participants and evaluated the four quality dimensions of the TPD Monitor—clarity and structure, practical relevance, cognitive activation, and collaboration—using 18 items (Richter & Richter, 2023). Clarity and structure measured the presentation of course objectives and the comprehensibility of the course. Practical relevance assessed the alignment of course content with participants' professional practice and the applicability of course content in the school setting. Cognitive activation assessed whether the course was intellectually stimulating for participants and offered them the chance to reflect on their own practice. Collaboration assessed whether the course gave participants the opportunity to interact with others and included periods of small group work. Respondents rated all items on a four-point Likert scale ranging from 1 (strongly disagree) to 4 (strongly agree). The research question included an analysis of the internal structure, and the results section presents information on the factor structure, means, standard deviations, and scale reliability.

To validate the TPD Monitor with external criteria, participants' behavioral engagement in the course was assessed using four items (adapted from Chan, Maneewan, & Koul, 2023). These items measured whether participants actively participated in the course, for instance: "I actively participate in the course activities." Respondents rated all items on a four-point Likert scale ranging from 1 (strongly disagree) to 4 (strongly agree). Reliability, measured by internal consistency, was found to be satisfactory (ω = 0.78).

Table 1
Description of the sample of event formats.

	Total group (n = 173)
Teacher target group	
All school types	25.7
Elementary schools	21.0
Secondary schools	32.2
Vocational schools	15.8
Schools for students with SEN[1]	4.7
Other school types	0.6
Event formats	
Face-to-face	68.6
Online	27.3
Hybrid	4.1

Note. [1]SEN=Special educational needs.

3.4. Data analysis

To evaluate the measurement quality of our assessment tool, considering the unique hierarchical structure of our data, we examined its reliability and validity. Participants' responses are nested within courses, requiring us to aggregate individual ratings at the course level for analysis. Consequently, our analyses are conducted using multilevel approaches, as recommended by Marsh et al. (2009).

We utilized Multilevel Structural Equation Modeling (ML-SEM) to test our first hypothesis, following the framework proposed by Marsh et al. (2009). These models control for measurement error at both individual and course levels by employing multiple indicators for each construct and aggregating individual responses to represent course-level characteristics (Morin, Marsh, Nagengast, & Scalas, 2014).

To assess the validity of our measures, we conducted Multilevel Confirmatory Factor Analysis (MCFA), using scale items as indicators for each latent variable. We evaluated model fit using standard criteria for single-level analyses (CFI close to.95, RMSEA close to.06, SRMR close to.08; Hu & Bentler, 1999). However, it's worth noting that interpreting global fit indices in multilevel models is still an area of ongoing research (Marsh, Hau, & Wen, 2004).

Descriptive analyses (means, standard deviations) were conducted at both individual and course levels to better understand our scales and items. Additionally, we assessed scale reliability using McDonald's Omega (ω) (Dunn, Baguley, & Brunsden, 2014).

To explore participants' shared perceptions of TPD quality within the same course, we calculated Intraclass Correlation Coefficients (ICC1 and ICC2). ICC1 represents the proportion of total variance occurring at the course level, while ICC2 indicates the reliability of the group average (Lüdtke et al., 2011). A recommended threshold for ICC1 is close to or greater than 0.10, whereas ICC2 values should exceed 0.70 (Lüdtke et al., 2011; Marsh et al., 2012).

Ensuring the comparability of constructs across levels and addressing our second hypothesis, we tested the invariance of factor loadings. This simplifies the model and enhances parameter estimate accuracy at the course level (Morin et al., 2014).

For our third hypothesis, we employed four multilevel regression models, each including a dimension of TPD process quality as an independent variable and participants' behavioral engagement as the dependent variable. We also included participants' gender, teaching experience, and school type as covariates at the individual level to control for potential influences.

Statistical analyses were conducted using Mplus 8.9 (Muthén & Muthén, 1998–2023) employing the Weighted Least Squares Mean and Variance adjusted (WLSMV) estimation method. To address missing data, we utilized full information maximum likelihood estimation (FIML; Enders, 2010).

4. Results

4.1. Factorial structure of TPD process quality

In the first phase, we conducted an analysis of the factorial structure of TPD process quality by testing a measurement model that included all four dimensions of TPD process quality, with each dimension represented by its respective items (clarity and structure, practical relevance, cognitive activation, and collaboration). MCFA were conducted to assess the validity of the instrument and its latent structure. The items were used as manifest indicators of the latent variables at both levels.

The measurement model including all four dimensions of TPD process quality showed an acceptable model fit (Fig. 1): $\chi 2(274) = 466.97$, CFI = .98, RMSEA = .01, $SRMR_{within}$ = .04 and $SRMR_{between}$ = .09. In this model, all factor loadings fell within acceptable ranges for both L1 variables (.72–.83 for clarity and structure, .79–.85 for practical relevance, .70–.80 for cognitive activation .85–.91 for collaboration) and L2 variables (.79–.99 for clarity and structure, .84–.99 for practical

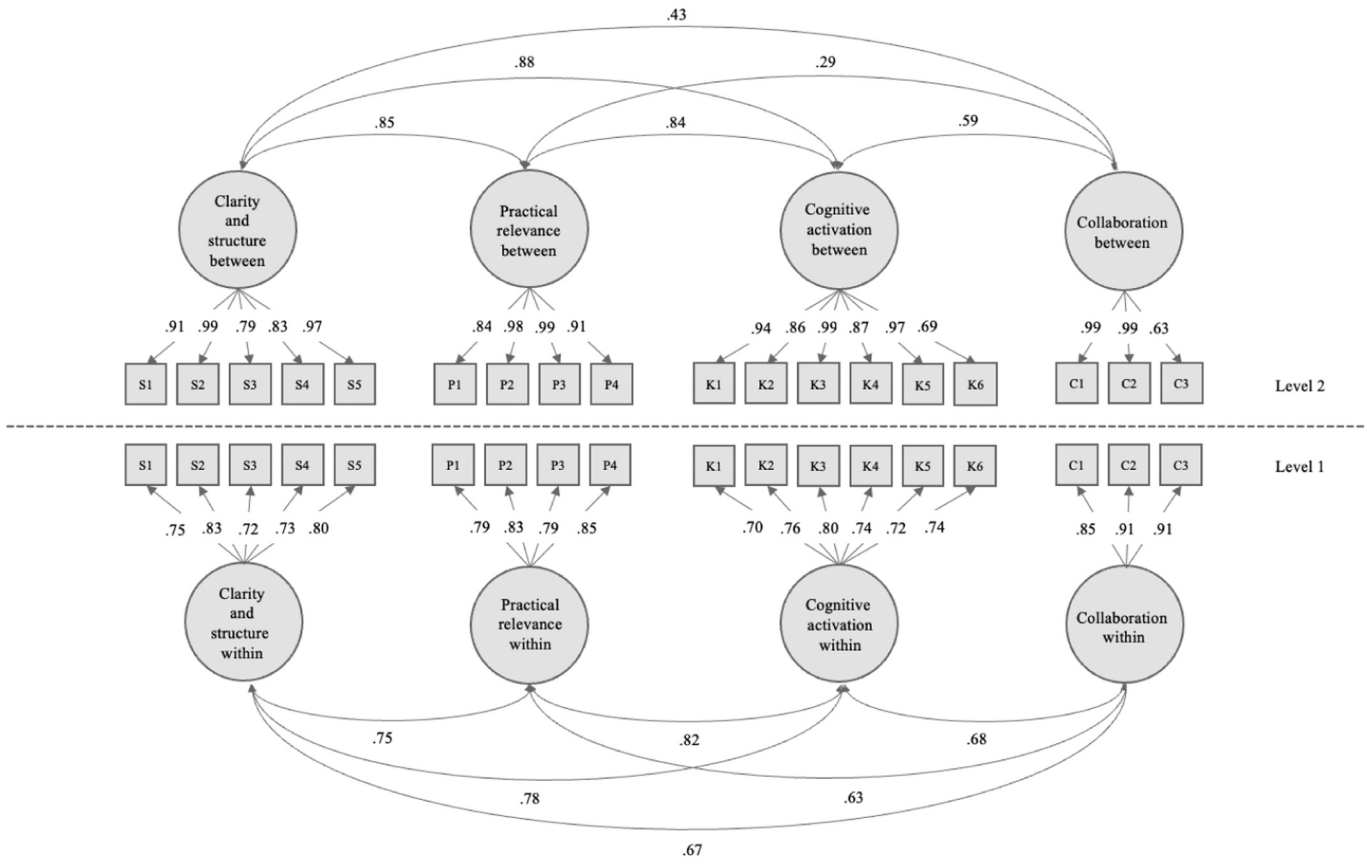

Fig. 1. Four-factor model of participants' ratings of TPD process quality. *Notes.* The figure represents the latent correlations between the four factors of process quality at the teacher and course levels. The factor structure was estimated at both levels due to the hierarchical structure of the data. The numbers between the double-headed arrows represent the latent correlations between the latent constructs and the numbers on the single headed arrows represent the factor loadings.

relevance, .69–.99 for cognitive activation, .63–.99 for collaboration).

To examine the robustness of the four-factor model, we conducted a comparison with an alternative single-factor model, where all items were assumed to load on a single factor (Table 2). The results showed that the single-factor model had a poorer fit to the data (ΔCFI = .05 and ΔRMSEA = .02), indicating the need to differentiate between items assessing different dimensions of TPD process quality.

In addition to examining the latent factor structure, we also report the descriptive findings for individual items and scales, with the scale values representing the mean of all associated items. The results show that the mean values of the scales are consistently above the theoretical mean of 2.5, indicating a positive assessment of the quality of the TPD process by the participants (Table 3). Furthermore, all reliability values of the scales are good to very good at both participant and course levels (Table 3). This indicates that the scales consistently measure the intended constructs with high precision and consistency, enhancing the credibility of our findings. Good reliability ensures that the measurement tool reliably captures the nuances of the TPD process quality, thereby bolstering the validity of our assessments and the confidence in our results.

In terms of intraclass correlation coefficients, the results indicate that the ICC2 values for all scales are within a satisfactory range, reflecting a reasonable level of agreement among participants within a course (Lüdtke et al., 2011). In addition, the analysis of ICC1 values showed significant differences between courses for all scales. Approximately 26 to 38% of the variance in participants' ratings can be attributed to the courses and their characteristics. Thus, the instrument appears to be well suited to identifying differences in process quality between TPD courses.

Table 2
Fit indices of multilevel confirmatory factor analyses: participants (n = 2314) in courses (n = 173).

	χ^2	*df*	*p*	CFI	RMSEA	SRMR	
						L1	L2
4-factor model	466.97	274	< .001	.98	.01	.04	.09
1-factor model	1035.54	288	< .001	.93	.03	.09	.17

4.2. Measurement invariance of TPD process quality

To assess measurement invariance, we examined whether scalar measurement invariance was present. This type of invariance requires that the factor structure, factor loadings, and intercepts of the items remain consistent across the subgroups being compared. Scalar invariance ensures that not only the relationships between items and latent constructs are equivalent across groups, but also that the item means are comparable. This is particularly relevant when comparing means, as it allows for valid comparisons of group differences without bias introduced by differential item functioning. Table 4 shows the goodness-of-fit indices of the models with scalar measurement invariance. The results indicate that all models with scalar invariance showed a sufficiently good fit to allow comparison of both the manifest and latent means of process quality between the different groups (i.e., based on gender, type of school).

4.3. Regression analysis: TPD process quality dimension and external criteria

To test the hypothesized relationship between the TPD process quality dimensions and the external criteria behavioral engagement,

Table 3
Descriptive statistics of different dimension of TPD process quality (participant and course level).

	Items	Example item	Participant level (L1) ($n = 2314$)					Course level (L2) ($n = 173$)		
			M	*SD*	ω	ICC1	ICC2	*M*	*SD*	ω
Clarity and structure	5	The objectives of the PD course were clearly articulated.	3.51	.69	.85	.33	.86	3.52	.33	.93
Practical relevance	4	The aspects addressed in the PD course were relevant to my current professional practice.	3.42	.69	.85	.28	.84	3.44	.32	.95
Cognitive activation	6	The PD course allowed me to familiarize myself with unfamiliar concepts.	3.27	.74	.86	.26	.82	3.27	.33	.94
Collaboration	3	The PD course provided an opportunity to collaborate with other participants.	3.65	.66	.80	.38	.89	3.61	.42	.88

Table 4
Invariance tests of the four-factor model.

							Model comparison	
	χ^2	*df*	*p*	CFI	RMSEA	SRMR	ΔCFI	ΔRMSEA
Gender								
Configural	583.76	258	< .001	.967	.034	.055	-	-
Metric	599.96	272	< .001	.966	.033	.060	-.001	-.001
Scalar	620.02	286	< .001	.966	.033	.060	.000	.000
School type								
Configural	657.45	258	< .001	.963	.038	.078	-	-
Metric	701.43	272	< .001	.961	.038	.097	-.002	.000
Scalar	725.89	286	< .001	.960	.038	.100	-.001	.000

four multilevel regression analyses were conducted (Table 5). The results show positive and statistically significant relationships between each of the four quality dimensions and behavioral engagement. Specifically, higher levels of each quality dimension are associated with higher levels of participant behavioral engagement within a course. In our multilevel regression models, covariates play a crucial role in accounting for potential influences on teachers' behavioral engagement within the TPD courses. Specifically, we found that teaching experience exhibited a positive relationship with teachers' behavioral engagement, indicating that more experienced teachers tended to demonstrate higher levels of engagement. Additionally, our analysis revealed that teachers from elementary schools exhibited slightly lower levels of engagement compared to those from other school types.

5. Discussion

This study's objective was to explore the measurement properties and the quality of an evaluation instrument designed to gauge TPD process quality. The development of this instrument took place in a participatory process involving school administrators and representatives of institutions of TPD to ensure the constructs' validity and future users' acceptance of the instrument. Based on the defining features of effective TPD identified in the research literature (e.g., Darling-Hammond, Hyler, & Gardner, 2017; Meyer et al., 2023), the resultant questionnaire comprises four dimensions of process quality (clarity and structure, practical relevance, cognitive activation, and collaboration) which are measured by 18 items. This instrument enables a time-efficient and practical evaluation of TPD process quality and serves as a foundation for continuous quality assurance of in-service training within school districts or states.

Although our focus was primarily on process quality, the relevant literature has identified various characteristics of effective TPD (Darling-Hammond, Hyler, & Gardner, 2017). The empirical substantiation for these characteristics is somewhat tenuous, however. Retrospective reviews of evaluation studies form the basis of many investigations into the characteristics of effective TPD courses. Yet these studies rarely use

Table 5
Results of the multilevel regression analyses: Dimensions of TPD process quality as predictors of external criteria.

Predictor	Behavioral Engagement			
	Model 1 β (*SE*)	*Model 2* β (*SE*)	*Model 3* β (*SE*)	*Model 4* β (*SE*)
Participant level (L1)				
Clarity and structure	.30** (.03)			
Practical relevance		.41** (.03)		
Cognitive activation			.45** (.03)	
Collaboration				.35** (.02)
Gender (female)	-.02 (.02)	-.03 (.02)	-.01 (.02)	-.01 (.02)
Teaching experience	.12** (.02)	.13** (.02)	.13** (.02)	.13** (.02)
School type (elementary school)	-.04 (.03)	-.09** (.03)	-.05 (.03)	-.05 (.03)
R^2 (L1)	.10 (.01)	.18 (.02)	.22 (.02)	.14 (.02)
χ^2	401.53	313.96	395.08	288.00
df	86	68	106	52
p	< .001	< .001	< .001	< .001
CFI	.90	.92	.92	.91
RMSEA	.04	.04	.03	.04
SRMR (L1)	.11	.05	.05	.06
SRMR (L2)	.11	.10	.11	.07

Note. * $p < .05$, ** $p < 0.01$; standardized regression weights; standard errors are in parentheses

participants' ratings of TPD quality; instead, they rely on external raters' evaluations of quality characteristics. This methodology raises questions about whether ascribed quality characteristics align with participant ratings of process quality. Therefore, valid questionnaires are required to measure TPD course quality and study designs that gauge and juxtapose process quality from varying perspectives.

The resultant reliable and valid instrument from our study can serve in evaluating the process quality of TPD courses. Our validation study utilized a randomly selected, extensive sample of TPD courses, ensuring the applicability of the instrument across diverse content, formats, and target groups. Throughout the validation phase, rigorous analyses verified the factor structure of the four dimensions, considering the multilevel structure of the data. Furthermore, analyses were conducted to determine whether the tool's measurement properties differed among teacher groups (i.e., elementary vs. secondary teachers; female vs. male). This research thus conforms to stringent methodological standards often overlooked in other studies featuring TPD measurement tools. The only previous works to provide comprehensive statistical tests for the validity of a TPD measurement instrument, Soine and Lumpe (2014) and Meyer et al. (2023), focused exclusively on subject-specific and online TPD courses, respectively. Our study presents an evaluation tool of high measurement quality that is applicable to online and face-to-face settings as well as to subject-specific and domain-general TPD courses.

However, some limitations must be considered. The present study explored learning process quality in TPD courses with a wide array of objectives and course content. It is therefore not possible to draw any conclusions about the skills and knowledge that the teachers gained in the courses. The study was also conducted in a single German federal state, limiting the generalizability of the findings. As a result, further research is necessary to examine the tool's measurement properties in different populations. In this study, we relied solely on the responses of participating teachers to glean information about process quality. Given the unique perspectives of facilitators and external observers and their opportunities for comparison, it would be beneficial to include their assessments of process quality in future studies.

The results of this study serve as the foundation for future research and for the application of the tool in TPD practice. Further research should include the additional attributes considered in studies of the effectiveness of TPD (e.g., coherence, content focus) in addition to process quality, to better understand the relationships between these characteristics and process quality and to develop a more comprehensive tool for evaluating TPD quality. Additionally, longitudinal studies could be conducted to examine how process quality changes within and across multiple events in a series of TPD courses. Furthermore, investigating the relationships between process quality and the learning outcomes of participating teachers could yield valuable insights. Tools such as subjective assessments of competence gains or knowledge tests could be employed for this purpose.

The evaluation tool developed in this study also holds potential for enhancing the quality of TPD courses. It could be used to benchmark and compare different TPD courses on a school district, state, or even national level to inform decisions about course adoption, funding, and policy. Moreover, facilitators could employ the tool to obtain feedback on the process quality of their TPD courses, offering valuable insights for their professional development and for the refinement of their TPD strategies. Additionally, the results from the tool could be used to guide and inform the design of new TPD courses that are likely to be of high quality from the outset. Finally, should a standard for TPD process quality be established, this tool could play a part in a certification or accreditation process for TPD courses, ensuring a consistent quality standard across these courses.

CRediT authorship contribution statement

Eric Richter: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dirk Richter:** Writing – review & editing, Project administration, Funding acquisition, Resources, Supervision, Writing – original draft.

Declaration of Competing Interest

We have no conflicts of interest to disclose. This research received funding from the Ministry for School and Education of the State of North Rhine-Westphalia (Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.stueduc.2024.101357.

References

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry, & J. C. Smart (Eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (Vol. 6, pp. 385–456). Springer Netherlands. https://doi.org/10.1007/1-4020-5742-3_10.

Adams, D. (1993). Defining educational quality. *Improving Educational Quality Project Publication, 1*, 3–24.

Adams, D., Acedo, C., & Popa, S. (2012). In search of quality education. In C. Acedo, D. Adams, & S. Popa (Eds.), *Quality and Qualities* (Vol. 54, pp. 1–22). SensePublishers. https://doi.org/10.1007/978-94-6091-951-0_1.

American Educational Research Association (AERA). (2011). Report and recommendations for the reauthorization of the institute of education sciences. American Educational Research Association.

Brophy, J. (2000). Teaching. Educational practices series-1. International Bureau of Education.

Caroll, J. B. (1963). A primer of programmed instruction in foreign language teaching. *International Review of Applied Linguistics in Language Teaching, 1*(2), 115–141.

Chan, S., Maneewan, S., & Koul, R. (2023). Teacher educators' teaching styles: Relation with learning motivation and academic engagement in pre-service teachers. *Teaching in Higher Education, 28*(8), 2044–2065. https://doi.org/10.1080/13562517.2021.1947226

Darling-Hammond, L., Hyler, M., & Gardner, M. (2017). Effective teacher professional development. *Learning Policy Institute*. https://doi.org/10.54300/122.311

Daumiller, M., Janke, S., Hein, J., Rinas, R., Dickhäuser, O., & Dresel, M. (2022). Teaching quality in higher education: Agreement between teacher self-reports and student evaluations. *European Journal of Psychological Assessment, 39*(3), 176–181. https://doi.org/10.1027/1015-5759/a000700

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199. https://doi.org/10.3102/0013189X08331140

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Enders, C.K. (2010). Applied missing data analysis. Guilford Press.

Fischer, C., Fishman, B., Dede, C., Eisenkraft, A., Frumin, K., Foster, B., … McCoy, A. (2018). Investigating relationships between school context, teacher professional development, teaching practices, and student achievement in response to a nationwide science reform. *Teaching and Teacher Education, 72*(1), 107–121. https://doi.org/10.1016/j.tate.2018.02.011

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945. https://doi.org/10.3102/00028312038004915

Garrison, D. R., Anderson, T., & Archer, W. (2003). A theory of critical inquiry in online distance education. In M. G. Moore, & W. G. Anderson (Eds.), *Handbook of distance education* (pp. 113–127). Lawrence Erlbaum Associates, Inc.

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education, 15*(1), 7–23. https://doi.org/10.1080/08923640109527071

Garrison, D., Anderson, T., & Archer, W. (1999). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education, 2*(2-3), 87–105. https://doi.org/10.1016/S1096-7516(00)00016-6

Gaumer Erickson, A. S., Noonan, P. M., Brussow, J., & Supon Carter, K. (2017). Measuring the quality of professional development training. *Professional Development in Education, 43*(4), 685–688. https://doi.org/10.1080/19415257.2016.1179665

Goe, L., & Stickler, L.M. (2008). Teacher quality and student achievment: Making the most of recent research. National Comprehensive Center for Teacher Quality.

Golding, C., & Adam, L. (2016). Evaluate to improve: Useful approaches to student evaluation. *Assessment & Evaluation in Higher Education, 41*(1), 1–14. https://doi.org/10.1080/02602938.2014.976810

Gosling, D., & D'Andrea, V. M. (2001). Quality development: A new concept for higher education. *Quality in Higher Education, 7*(1), 7–17. https://doi.org/10.1080/13538320120045049

Guskey, T.R., (2000). Evaluating professional development. Corwin Press.

Hox, J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel Analysis Techniques and Applications*. Routledge. https://doi.org/10.4324/9781315650982

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kalinowski, E., Gronostaj, A., & Vock, M. (2019). Effective professional development for teachers to foster students' academic language proficiency across the curriculum: A systematic review, 2332858419828691 *AERA Open, 5*(1). https://doi.org/10.1177/2332858419828691.

Kennedy, M. M. (2006). Knowledge and vision in teaching. *Journal of Teacher Education, 57*(3), 205–211. https://doi.org/10.1177/0022487105285639

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research, 86*(4), 945–980. https://doi.org/10.3102/0034654315626800

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Lipowsky, F., & Rzejak, D. (2021). Fortbildungen für Lehrpersonen wirksam gestalten - ein praxisorientierter und forschungsgestützter Leitfaden [Designing effective teacher professional development programs - a practice-oriented and research-based guide]. Bertelsmann Stiftung. ⟨https://doi.org/10.11586/2020080⟩.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16*(4), 444–467. https://doi.org/10.1037/a0024376

Main, K., & Pendergast, D. (2015). Core features of effective continuing professional development for the middle years: A tool for reflection. *RMLE Online, 38*(10), 1–18. https://doi.org/10.1080/19404476.2015.11658177

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

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

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*(6), 764–802. https://doi.org/10.1080/00273170903333665

Meyer, A., Kleinknecht, M., & Richter, D. (2023). What makes online professional development effective? The effect of quality characteristics on teachers' satisfaction and changes in their professional practices. *Computers & Education, 200*(3), Article 104805. https://doi.org/10.1016/j.compedu.2023.104805

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education, 82*(2), 143–167. https://doi.org/10.1080/00220973.2013.769412

Muthén, L.K., & Muthén, B. (1998-2023). Mplus user's guide. Los Angeles, CA: Muthén & Muthén.

O'Hanlon, J., & Mortensen, L. (1980). Making teacher evaluation work. *The Journal of Higher Education, 51*(6), 664–672. https://doi.org/10.1080/00221546.1980.11780102

Organisation for Economic Cooperation and Development (O.E.C.D.) (2005). *School factors related to quality and equity: Results from PIS*A 2000. Organisation for Economic Cooperation and Development.

Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM, 50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Pratasavitskaya, H., & Stensaker, B. (2010). Quality management in higher education: Towards a better understanding of an emerging field. *Quality in Higher Education, 16*(1), 37–50. https://doi.org/10.1080/13538321003679465

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education, 30*(4), 387–415. https://doi.org/10.1080/02602930500099193

Richter, E., & Richter, D. (2023). *Fortbildungsmonitor. Ein Instrument zur Erfassung der Prozessqualität von Lehrkräftefortbildungen* [The teacher professional development monitor. An instrument for measuring the quality of teacher professional development]. Potsdam. https://doi.org/10.25656/01:27640.

Scheerens, J. (2004). Perspectives on education quality, education indicators and benchmarking. *European Educational Research Journal, 3*(1), 115–138. https://doi.org/10.2304/eerj.2004.3.1.3

Scheerens, J., Luyten, H., & van Ravens, J. (2011a). Perspectives on educational quality. In J. Scheerens, H. Luyten, & J. van Ravens (Eds.), *SpringerBriefs in Education. Perspectives on Educational Quality* (Vol. 1, pp. 3–33). Springer Netherlands. https://doi.org/10.1007/978-94-007-0926-3_1.

Scheerens, J., Luyten, H., & van Ravens, J. (2011b). Measuring educational quality by means of indicators. In J. Scheerens, H. Luyten, & J. van Ravens (Eds.), *SpringerBriefs in Education. Perspectives on Educational Quality* (Vol. 1, pp. 35–50). Springer Netherlands. https://doi.org/10.1007/978-94-007-0926-3_2.

Senden, B., Nilsen, T., & Blömeke, S. (2021). 5. Instructional quality: A review of conceptualizations, measurement approaches, and research findings. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Ways of Analyzing Teaching Quality* (pp. 140–172). Scandinavian University Press. https://doi.org/10.18261/9788215045054-2021-05.

Shuell, T. J. (2001). Learning theories and educational paradigms. In Paul B. Baltes (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (Vol. 1, pp. 8613–8620). Elsevier. https://doi.org/10.1016/B0-08-043076-7/02385-8.

Sims, S., & Fletcher-Wood, H. (2021). Identifying the characteristics of effective teacher professional development: A critical review. *School Effectiveness and School Improvement, 32*(1), 47–63. https://doi.org/10.1080/09243453.2020.1772841

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S. (2021). What are the characteristics of teacher professional development that increase pupil achievement? A systematic review and meta-analysis. London.

Soine, K. M., & Lumpe, A. (2014). Measuring characteristics of teacher professional development. *Teacher Development, 18*(3), 303–333. https://doi.org/10.1080/13664530.2014.911775

Timperley, H.,Wilson, A., Barrar, H. & Fung, I. (2007). Teacher professional learning and development: Best evidence synthesis iteration. Wellington, New Zealand. University of Auckland.

Trivette, C. M., Dunst, C. J., Hamby, D. W., & O'Herin, C. E. (2009). Characteristics and consequences of adult learning methods and strategies. *Research Briefing, 3*(1).

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28*(3)), 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003

Walter, C., & Briggs, J. (2012). What professional development makes the most difference to teachers? Oxford. University of Oxford Department of Education.

Wei, R.C., Darling-Hammond, L., Andree, A., Richardson, N., & Orphanos, S. (2009). Professional learning in the learning profession: A status report on teacher development in the United States and abroad. Dallas.

Yarbrough, D.B., Shulha, L.M., Hopson, R.K., & Caruthers, F.A. (2010). The program evaluation standards: A guide for evaluators and evaluation users (3rd ed.). SAGE.

Yoon, K.S., Duncan, T., Lee, S.W.-Y., Scarloss, B., & Shapley, K. (2007). Reviewing the evidence on how teacher professional development affects student achievement (Issues & Answers No. 33). San Antonio.