

Effectiveness of interventions in teacher education: Synthesis of reviews (2010–2024)

Johannes König  · Sandra Heine  · Jonas Weyers  ·
Carl-Gereon Arntz · Charlotte Kramer  · Sarah Strauß  ·
Kristina Gerhard  · Janine Mühle  · Tim Engartner  ·
Jörg Großschedl  · Charlotte Hanisch  · Petra Hanke  ·
Thomas Hennemann  · Jörg Jost  · Kai Kaspar  · Benjamin Rott 

Received: 1 September 2024 / Revised: 28 May 2025 / Accepted: 2 June 2025 / Published online: 16 July 2025

© The Author(s) 2025

Abstract This paper offers an overview of current research on the effectiveness of interventions in teacher education. We synthesize 27 reviews published between 2010 and 2024, comprising 510 research publications of empirical intervention studies targeting pre-service and in-service teachers’ professional learning. We applied a specific analysis category system to address five research questions concerning (1) justifications and theoretical frameworks, (2) input and output of the intervention studies, (3) study design, (4) effect sizes, and (5) research gaps. Few reviews had developed their own theoretical framework, and justifications for the interventions’ necessity were not always elaborated. Most assumed the existence of an effect chain—namely, that the reviewed interventions help promote teacher variables (mainly competence facets), which are then positively reflected in teaching quality and students’ learning gains. However, study variables typically indicate only segments of the effect chain while comprehensive operationalization of the total effect chain is rare. Interventions are typically organized as coursework, not necessarily connected to teaching practice or student learning. The interventions’ targeted learning outcomes are heterogeneously operationalized. Basic design features appear typical for intervention studies in teacher education, whereas more challenging designs are less common. Self-report instruments appear more common than rigorous measurement approaches. Based on a subsample of 14 reviews reporting summarized effect sizes, we conducted a meta-meta-analysis that shows an overall significant moderate effect of interventions is observable. Differentiation shows stronger effects for proximal outcomes (e.g., change in teacher knowledge or instructional practice) than for student learning. The synthesis of moderator effects yielded mixed results, and the quality of interventions appears more crucial than the length. Research gaps concern further theoretical development, the individual intervention studies’ research methodological standards, and improvement of literature reviews’ methodological quality. The results offer a theoretical background for

Extended author information available on the last page of the article.

future research on teacher education interventions and research progress in teacher education effectiveness. Recommendations for future research are given.

Keywords Effectiveness · Intervention · Meta-meta-analysis · Review · Teacher education · Teacher professional development

Stichwort Wirksamkeit von Interventionen in der Lehrkräftebildung: Synthese von Reviews (2010–2024)

Zusammenfassung Der Beitrag bietet einen Überblick über die aktuelle Forschung zur Wirksamkeit von Interventionen in der Lehrkräftebildung. Auf der Grundlage einer systematischen Literaturrecherche fassen wir 27 Übersichtsartikel (systematische Reviews und Meta-Analysen) zusammen, die zwischen 2010 und 2024 veröffentlicht wurden. Sie umfassen insgesamt 510 Publikationen empirischer Interventionsstudien, die sich mit dem professionellen Lernen von angehenden und berufstätigen Lehrkräften in der Aus- und Fortbildung befassen. Wir verwendeten ein spezifisches Analysekategoriensystem zur Beantwortung von fünf Forschungsfragen, die sich auf (1) Begründungen und theoretische Rahmungen, (2) Input und Output der Interventionsstudien, (3) Studiendesign, (4) Effektgrößen und (5) Forschungslücken beziehen. Kaum eine der Übersichten entwickelte einen eigenen theoretischen Rahmen, und Begründungen für die Notwendigkeit von Interventionen wurden nicht immer ausgearbeitet. Die meisten Übersichtsarbeiten folgten der Annahme einer Wirkungskette, nämlich dass die untersuchten Interventionen zur Förderung von Lehrpersonvariablen (vor allem Kompetenzfacetten) beitragen, die sich wiederum positiv in der Unterrichtsqualität und im Lernzuwachs von Schülerinnen und Schülern niederschlagen sollten. Allerdings geben die Studienvariablen in allen Übersichtsartikeln meist nur Segmente der Wirkungskette an, selten wird die gesamte Wirkungskette umfassend operationalisiert. Die Interventionen sind in der Regel als Lehrveranstaltungen organisiert, die nicht notwendigerweise mit der Unterrichtspraxis oder dem Lernen der Schülerinnen und Schüler verbunden sind. Die Operationalisierungen der angestrebten Lernergebnisse der Interventionen sind heterogen. Grundlegende Design-Merkmale scheinen typisch für Interventionsstudien in der Lehrkräfteausbildung zu sein, während anspruchsvollere Designs seltener zu finden sind. Selbsteinschätzungsinstrumente scheinen häufiger zu sein als testdiagnostische Messansätze. Auf der Grundlage einer Teilstichprobe von 14 Übersichtsarbeiten, die zusammengefasste Effektgrößen aufweisen, kann in einer durchgeführten Meta-Meta-Analyse ein insgesamt signifikanter moderater Effekt von Interventionen festgestellt werden. Eine Differenzierung der Effekte zeigt stärkere Effekte für proximale Ergebnisse (z.B. Veränderung des Lehrkräftewissens oder der Unterrichtspraxis) als für das Lernen der Schülerinnen und Schüler. Die Synthese von Moderatoreffekten ergab gemischte Ergebnisse, lässt aber vermuten, dass eher die Qualität als die Dauer der Interventionen entscheidend ist. Forschungslücken betreffen die Weiterentwicklung der Theorie, die Berücksichtigung forschungsmethodischer Standards der einzelnen Interventionsstudien sowie die methodische Qualitätsverbesserung von Literaturübersichten. Die Ergebnisse können den theoretischen Hintergrund für künftige

Forschungen zu Interventionen in der Lehrkräftebildung und für Fortschritte in der Forschung zur Wirksamkeit der Lehrkräftebildung liefern. Empfehlungen für zukünftige Forschung werden gegeben.

Schlüsselwörter Intervention · Lehrkräftebildung · Lehrkräftefortbildung · Meta-Meta-Analyse · Übersicht · Wirksamkeit

1 Introduction

Numerous educational reforms in multiple countries worldwide target pre-service and in-service teachers' professional learning, highlighting the significance of the efforts undertaken in initial teacher education (ITE) and teacher professional development (TPD). One reason for this is the widespread and basic assumption of an underlying effect chain—that is, the idea that ITE and TPD learning opportunities enhance the competence of teachers, who, in turn, are better positioned to deliver high-quality instruction to their students, ultimately resulting in more favorable student learning outcomes at school (e.g., Desimone 2009; Lipowsky and Rzejak 2015; Scheerens and Blömeke 2016). In spite of this assumption's potential plausibility, however, the actual effectiveness of these efforts in ITE and TPD continues to be debated in educational science, indicating that our scientific knowledge about teacher education effectiveness remains limited (Blömeke et al. 2008; König et al. 2024).

Researchers in recent decades have conducted empirical research studies that aim to evaluate the impact of interventions in both ITE and TPD, yielding scientific insights into teacher education effectiveness (Hascher 2014; König and Blömeke 2020; König et al. 2024). Intervention research studies in this sense typically relate to a clearly defined unit of purposeful learning opportunities for pre-service or in-service teachers—usually training or coursework, sometimes combined with practice (for an overview, see König et al. 2024). Intervention researchers design such units with the aim of effecting change in specific outcome variables reflecting teacher learning and further consequences for their teaching and their students' learning (e.g., Leutner 2013; McBride 2016). Pre-service teachers usually benefit from an intervention that is outside, in addition to, or partly embedded in existing conventional teacher education programs, courses, or field experiences in ITE. Given that TPD, unlike ITE, does not necessarily follow a formally structured program path (in the German context, at least), a TPD measure can serve as an intervention for in-service teachers or an intervention may be embedded into an overall opportunity in TPD (e.g., Ansyari et al. 2022; Kalinowski et al. 2019). Intervention research studies are expected to follow a particular experimental study design methodology (Campbell and Stanley 1963; Hill et al. 2021; König 2013).

Scholars have begun synthesizing research intervention studies in teacher education in a bid to obtain systematic insights into teacher education effectiveness (e.g., Cochran-Smith and Villegas 2015). Such literature reviews often confront particularly prominent challenges in educational reform, such as inclusion (e.g., Dignath et al. 2022) or digitalization in education (e.g., Wilson et al. 2020). Therefore, they make incontrovertibly significant contributions to the ongoing question of

how teacher education measures can most effectively prepare and educate teachers for upcoming challenges (Darling-Hammond and Bransford 2007; Schleicher 2012, 2016). However, when considering a more general and overarching perspective of interventions in teacher education, our scientific understanding of teacher education effectiveness as a research field remains relatively limited (König et al. 2024). One reason for this is the heterogeneity of research approaches adopted in these intervention studies, which can impede comparison (e.g., Kennedy 2016). Nevertheless, the number of literature reviews synthesizing single research intervention studies has shown a remarkable increase in recent years (König et al. 2024). Such an increase in reviews generally opens up the opportunity to carry out a synthesis or meta-meta-analysis of the effectiveness of interventions in teacher education—partly following those approaches that have long been carried out for effective teaching in schools (Hattie 2009, 2023). This would provide overarching orientation for the constantly growing field of research on teacher education effectiveness (König et al. 2024, p. 373).

Against this background, the present paper uses a total of 27 such literature reviews, selected by means of a systematic literature search, with the aim of providing an overarching perspective on the state of interventions in teacher education research. As for example Rowan et al. (2015, p. 275) maintain, this can facilitate our understanding of how the “large number of small scale but often disconnected studies of teacher education practice” can be evaluated. Specifically, regarding the question of teacher education effectiveness, the present paper aims to provide a more robust foundation from which to address five key research questions (see Sect. 4) relating to (1) justifications and theoretical frameworks, (2) input and output of the intervention studies, (3) study design, (4) effect sizes, and (5) research gaps. Findings are discussed and recommendations are provided for future research to be carried out by scholars who are committed to improve the quality of interventions in teacher education and to enhance our understanding of teacher education effectiveness in general.

2 Teacher education as an object for effectiveness examination

2.1 Analysis of the status quo: Survey studies

Teacher education effectiveness research has largely been undertaken in the form of survey studies (Blömeke et al. 2008), most prominently demonstrated by the Teacher Education and Development Study: Learning to Teach Mathematics (TEDS-M), a comparative large-scale study of 17 countries worldwide led by the International Association for the Evaluation of Educational Achievement (IEA) in 2008 (Tatto et al. 2008; Tatto and Senk 2011). Teacher professional knowledge was directly assessed using representative country samples with future teachers selected shortly prior to attaining full certification as teaching professionals. Their scores from standardized tests were linked to survey data on the learning opportunities to which the future teachers had been exposed during their ITE. Teacher education effectiveness was thus examined through cross-sectional correlational analyses between

processes and characteristics of ITE programs that future teachers had followed and the outcomes of ITE (Blömeke et al. 2012).

Survey studies—in particular those in the tradition of the large-scale assessments in TEDS-M (for a recent overview, see Kaiser 2024) and studies extending cross-sectional to longitudinal design (e.g., Kauper et al. 2023; Mahler et al. 2024; Weyers et al. 2024)—undoubtedly furnish the research-based evidence that is urgently required for teacher education reform and to inform educational policy (Kaiser and König 2019). These studies function to analyze and reflect on the status quo. However, regarding the emerging teacher education effectiveness research paradigm (Cochran-Smith and Villegas 2015), intervention studies are also needed to generate detailed insights into teacher education effectiveness that exceed our status quo knowledge of existing or conventional teacher education (König et al. 2024). As this is clearly a fundamental question of initiating research (e.g., Coe et al. 2021), following the components of a logical sequence of empirical investigations (e.g., Friedrichs 1990, p. 51), it seems important to first account for possible starting points for researchers to develop an intervention study in teacher education—before going into further components such as research methods and utilization of results (see Sect. 3).

2.2 Starting points to develop an intervention study

Intervention studies typically depart from a specific starting point, two of which are particularly prominent on the basis of their potential to enrich our knowledge derived from survey study design in educational research (König 2013). One of these starting points emerges from a mismatch between the conventional curriculum and new demands imposed on teacher education by educational reform (e.g., Little 1993). The missing fit has been identified based on findings from survey design studies—recently, for example, studies have highlighted that the teacher education curriculum appears to lag behind the overall digital transformation in education (e.g., Fernández-Batanero et al. 2022; Gerhard et al. 2023). Moreover, a specific curricular innovation ensuing from the reform imperative must be positively evaluated with respect to its effectiveness before it can be further implemented into the existing curriculum (where it may also be evaluated further using survey study designs).

Moreover, the reform imperative leading to the necessity of interventions may be accompanied by other factors. For example, a certain topic may reflect recent societal or cultural developments not yet covered adequately by the teacher education curriculum. Recent developments, however, have urged the prioritization of such topics in teacher education. Inclusive education and digitalization have been prominent examples of such topics over the last decade, with more presumably to come, such as navigating the consequences of flight and migration, climate change and sustainability, democracy education, or mental and physical health.

The application of survey studies to investigate effect mechanisms can primarily provide indications of causality, but influences of unconsidered third variables, for example, cannot be ruled out. Only experimental studies—recognized or at least discussed as the gold standard for testing causal assumptions (e.g., Colnet et al. 2024)—make it possible to actually investigate the chain of effects in the causal

sense. Therefore, another key starting point concerns the need to provide causal evidence for effects previously found in other study designs (e.g., survey studies or non-experimental studies), giving rise to intervention research implemented using experimental study designs in teacher education. However, given that the second starting point is more methodological in nature, it may be less prominent in the current discourse surrounding interventions in teacher education. The investigation of theory-based assumptions using experimental study design may offer a further perspective that prompts the implementation of intervention research in teacher education. However, the question of the effectiveness of interventions in teacher education relates primarily to questions of curricular or methodological innovation.

Many educational reforms targeting pre-service and in-service teachers' learning are aimed at improving instructional quality and supporting students' learning progress (e.g., Lieberman and Pointer Mace 2008). Curricular innovations in ITE (e.g., extending curricular content for preparing future teachers for inclusive teaching) and novel TPD measures (e.g., in the field of digitalization at school) may serve as mediators in the effectiveness of educational policy for pre-service and in-service teacher professionalization, frequently linked to assumed impact on teaching and learning at school. Therefore, it is important to understand what factors contribute to the effectiveness of innovations in the ITE curriculum and novel TPD measures, since only then can the success or failure of educational reforms be fully understood (Desimone 2009).

Educational research relating to teaching and teachers is required to contribute to such an understanding, both from a theoretical perspective, which helps, for example, to apply a well-defined understanding of "innovation" in teacher education (Ellis et al. 2023), and through empirical research that fulfills the specific expectations of methodological rigor and rigorous intervention standards (e.g., Hill et al. 2021). One reason is that behind general agreements in education, such as the suggested roles that ITE and TPD play in assuring, preparing and supporting teachers to meet the new expectations created by educational reforms, may also lie competing expectations, foci, and approaches to educational reforms (Cochran-Smith 2001; Wang et al. 2010). For example, central definitions of effectiveness may vary, implying different topics and suggesting different priorities for educational reform. As a concrete example, we may cite the question of the extent to which artificial intelligence should form part of the overall digital transformation in education (Sperling et al. 2024). Another question relates to the actual capacity of teachers to implement educational reforms. Teachers may play a key role in translating ambitious reform efforts into everyday teaching practices in the school setting (see, for example, in relation to inclusive education: Rouse 2017). Meanwhile, they may also experience various barriers that hinder the effective integration of innovation delivered to them via an intervention (Ellis et al. 2023).

3 Intervention studies in teacher education

3.1 Research methods: Experimental study design

Teacher education effectiveness research is committed to the application of not only survey methods but also experimental study designs (König and Blömeke 2020). Other approaches to research on teachers and their profession leave experimental research virtually unconsidered (see, for example, the comprehensive overview for the situation in Austria, Germany, and Switzerland by Cramer et al. 2020). One reason for this is that the majority of teacher education and teacher professionalism research has its origins in disciplines that espouse methodologies other than experimental research (Cramer et al. 2023) and, as a possible consequence, seldom refer to the state of empirical findings proliferated by studies using experimental designs. However, owing to the increase in relevant empirical studies and literature reviews aggregating empirical findings, research on the effectiveness of teacher education, including experimental study designs, may today be acknowledged as an emerging research paradigm (König et al. 2024).

At its core, educational intervention research concerns the classic experiment in the sense of a randomized controlled study (Campbell and Stanley 1963; Coe et al. 2021; Hanushek and Jackson 2013; Leutner 2013; Shadish et al. 2002; Slavin 2002). Applied to teacher education, this usually involves specifying an intervention compared with a non-intervention (specified as the control group) or different variants of an intervention (e.g., competing interventions—multiple treatments—to test a specific hypothesis) as the independent variable. Then, for example, using a pre-post measurement design, the study will test whether the variation in the independent variable causes a variation in the relevant dependent variable, whereby the latter reflects the intervention's objective of the intervention—for example, an expected increase in teacher competence or a desired improvement of teaching practice on the part of the intervention's participants. Pre-post design may undergo various expansions, such as the addition of follow-up measurements that serve to examine the scope of potential intervention effects.

Intervention studies in teacher education can thus be further defined as content-based training (which might or might not relate to evidence-based concepts or practice) as part of an intervention with the outcome measure (pre-post) of teacher variables (e.g., their knowledge), evaluating the link between such measures and variables of their subsequent teaching (and the learning of their students). Meanwhile, the term “intervention” is also used in the literature with another meaning: to refer to approaches that apply evidence-based teaching concepts or practices to be (learned and) applied by teachers, possibly in combination with evaluating implementation for teaching concept or practice (and their students' learning). Given that interventions of this nature are located primarily on the level of teaching rather than teacher education, their specific understanding are not a focal point—or at least not the primary focal point—of the present synthesis (e.g., Brock and Carter 2017).

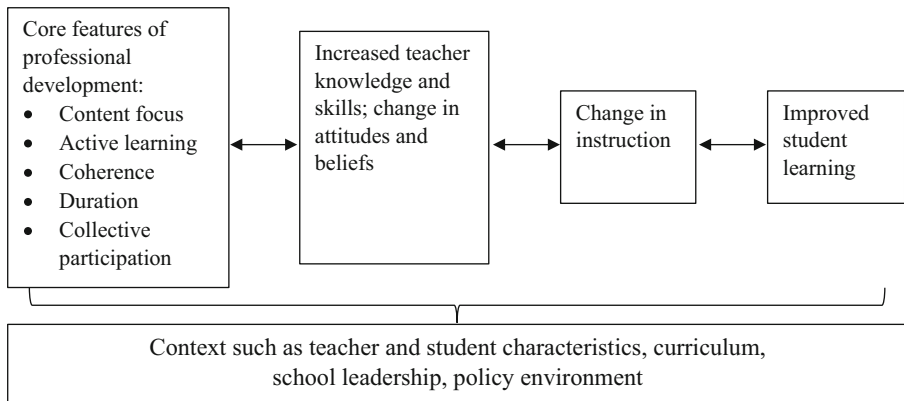


Fig. 1 Core conceptual framework for studying the effects of professional development on teachers and students by Desimone (2009, p. 185)

3.2 Scope and utilization of interventions: Evaluation and validation

Researchers have long assumed the existence of an underlying effect chain in teacher education comprising components such as teacher education opportunities to learn, increase of teacher competence, which in turn affects instructional quality and the learning progress of students (e.g., Desimone 2009; Lipowsky and Rzejak 2015; Scheerens and Blömeke 2016). One of the most prominent examples is presumably the conceptual framework proposed by Desimone (2009, p. 185) that is presented in Fig. 1. It suggests a sequential relationship, that is, an effective intervention enhances teacher variables, classroom teaching, and student learning. According to Desimone (2009, p. 185), it “allows testing both a theory of teacher change (e.g., that professional development alters teacher knowledge, beliefs, or practice) and a theory of instruction (e.g., that changed practice influences student achievement), both of which are necessary to complete our understanding of how professional development works”. However, such relationships among the four levels illustrated in Fig. 1 are neither unidirectional nor linear. For example, Desimone (2009) maintains that changes in beliefs may, for instance, be both a cause and an effect of improved student performance.

These assumptions may highlight the need to elaborate why the outcome of any intervention in ITE or TPD—in particular, those interventions that claim to be innovative—is desirable and for whom it is so (Ellis et al. 2023). Regarding intervention studies that adhere to the current expectations of teacher education effectiveness research (König et al. 2024), it is necessary not only to evaluate the proximal outcomes of an intervention but also to implement study design features that allow the efforts undertaken by the intervention (e.g., increased academic achievement of students) to be validated.

Meanwhile, the proximal outcome measurement of an intervention—for example, the direct acquisition of specific knowledge by means of TPD training among the participating teachers—appears significant as well. In particular, interventions that aim to affect educational practices and student learning may not be sufficiently

evaluated only through an empirical investigation of these practices in the teaching delivered by the intervention's participants. One reason is that teachers might profit from an intervention (e.g., in terms of competence increase), but at the same time, they may then struggle to implement practices for various reasons (e.g., Brock et al. 2014; Brock and Carter 2017; Cook and Cook 2013). It might therefore be important to provide a more comprehensive operationalization of the sequential relationship's components (Fig. 1) when creating intervention study research designs.

3.3 Characteristics of interventions

Scholars have reviewed and summarized central characteristics of interventions. For example, as presented in Fig. 1, Desimone (2009) proposes core characteristics that are more on a generic level focusing content relevant for school learning, active learning of teachers during professional development, coherence of interventions, adequate length and collective participation of teachers (e.g., from the same school). Active learning implies learning opportunities to connect professional learning and practical application, emphasizing the link between theory and practice. Parsons et al. (2016) outline general characteristics for TPD interventions “to be (a) sustained, (b) linked to students’ learning goals, (c) based on best practices, (d) delivered by coaches/experts, (e) collaborative participation among teachers, (f) based on student needs, (g) implemented with school support including solid leadership, and (h) reflective of teachers’ practice” (see Basma and Savage 2023, p. 214). Examination of TPD intervention characteristics (i.e. moderation effects) may be found in the reviews by Basma and Savage (2024) or Didion et al. (2020)—for example, with respect to variables such as study, program, or participant characteristics. As in the following synthesis of reviews a detailed examination of such characteristics does not seem to be feasible for various reasons (e.g., heterogeneity of approaches, large number of individual interventions, various topical foci), we follow the more sparse conceptualization developed by Kalinowski et al. (2019), who maintain that characteristics “can be organized into features that relate to the framework and structure of the intervention (structural features), the content of the intervention (content-related features), and the way that the content is delivered (didactic features).”

4 Research questions

An overview of the state of research using intervention studies in teacher education reveals a remarkably large corpus of publications. Numerous intervention studies in teacher education have been conducted and published and have even been synthesized by reviews to a remarkable extent. However, overarching perspectives (i.e., beyond a specific intervention, such as coaching, or beyond a specific topic, such as inclusive education) are rarely elaborated. Existing reviews and meta-analyses can therefore be used to investigate the state of intervention research without the need to simultaneously synthesize hundreds to thousands of studies. While this is a well-known approach in empirical educational research focusing effective teaching (Hattie 2009, 2023), to our knowledge, reviews of teacher education interventions

focusing on teacher education effectiveness have not been synthesized in such a way yet (König et al. 2024, p. 373). Using a substantial number of systematic and scoping literature reviews as well as meta-analyses of interventions in teacher education, our synthesis of these reviews focuses on research questions (RQ) that are framed by the following heuristic for reviewing interventions in teacher education (Fig. 2).

As Fig. 2 illustrates, interventions in teacher education need theoretical background and starting points first. RQ1 therefore seeks to identify justifications and theoretical frameworks of each review with the aim to synthesize a selected number of intervention studies; in particular, it seeks to determine what frameworks and concepts are applied and to what extent the review's topical focus concerns innovations in teacher education curriculum and methodological advances, given that these constitute important starting points for conducting interventions (as outlined before in Sect. 2).

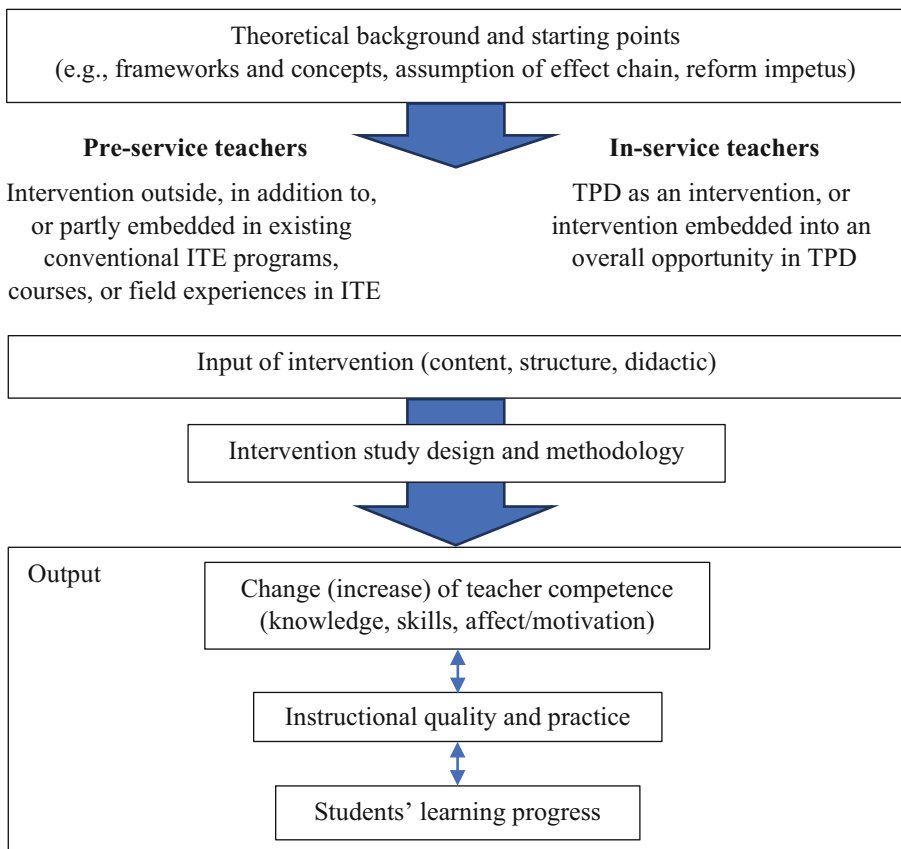


Fig. 2 Heuristic for reviewing interventions in teacher education. (*ITE* initial teacher education (target group: student teachers, pre-service teachers, teacher candidates etc.), *TPD* teacher professional development (target group: in-service teachers))

RQ1 Which frameworks exist and are used to reason regarding the characteristics of the specific interventions (e.g., curriculum innovation)?

RQ2 relates to the conceptualization, implementation, and measurement quality of dependent and independent variables. First, regarding our review of the input (independent variables), basic characteristics are focused rather than comprehensive records of intervention features. This specific focus is chosen as our synthesis of literature reviews takes on a more general and overarching perspective, gathering heterogeneous approaches and thus making it necessary to refrain from perspectives too detailed to compare (e.g., Kennedy 2016). As the following synthesis is not capable to record characteristics of interventions in great detail, we follow the more bundled feature organization proposed by Kalinowski et al. (2019)—as outlined before in Sect. 3: *content*, *structure*, *didactic*—when coding a selection of input variables describing basic characteristics of interventions. This includes features that relate to the framework of the intervention, namely the topical focus indicating the *content* of the intervention (topical focus and its area, see Table 3), the duration as a *structural* feature, and basic characteristics indicating *didactic* features (provision of course work, transfer into teaching practice, using digital resources, see Table 4). Second, regarding our review of the output (dependent variables), we are interested to value, among other aspects, the scope of study designs regarding evaluation and validation (Galluzzo and Craig 1990; König et al. 2024). That is, following model assumptions as outlined by Desimone (2009) and illustrated in Fig. 1, we are not only interested to review proximal output variables. Instead, we wish to determine to what extent the intervention studies encompass the following assumed effect chain through the operationalization of study variables: intervention (as a learning opportunity) *affects* teacher outcome measures (teacher competence, comprising knowledge, skills, and affect/motivation) *affects* observable performance of teachers and teaching in terms of teacher-student interaction (instructional quality and practice) *affects* student outcome (learning progress of students; e.g., Desimone 2009; Kaiser and König 2019).

RQ2 What are basic characteristics of the interventions (input)—in terms of content, structure, didactic—and which outcome measures (output)—in terms teacher competence, instructional quality, and learning progress of students—are covered?

The quality of research methods applied in intervention studies in teacher education has been criticized (e.g., Hill et al. 2021); therefore, with RQ3, we examine the reviews regarding the quality of study design and how intervention studies fulfill the expectations of experimental study designs (e.g., the inclusion of a control group).

RQ3 Do the interventions fulfill the rigorous standards in place for interventions aimed at providing evidence to judge teacher education effectiveness in the context of experimental study designs?

Given that several literature reviews of empirical intervention research studies in teacher education have applied meta-analyses, RQ4 aims at a synthesis of aggregated

effects towards outcome measures (output) as reported by the reviews, extended by a synthesis of moderator analyses.

RQ4 What are the interventions' effects on the specified outcomes, and what moderator effects have been identified in existing meta-analyses?

Finally, with RQ5, we aim to provide an overview of the key research gaps that future research should address, based on the conclusions of the selected literature reviews. As we code such conclusions, we directly address this as another RQ in our findings section before such gaps are generally discussed at the end of the present paper.

RQ5 What gaps should future research address?

5 Methodology

5.1 Systematic literature search

In March 2024, one member of the authorial team conducted a systematic literature search to identify the literature relevant to our research questions. The keywords essential to the search had been determined previously by the team. We used the databases Web of Science (WoS) and Education Resources Information Center (ERIC). Given that the two databases differ with respect to their structure, search capabilities, and the export of search results, we shall first describe our individual approach within each database before outlining the selection procedure. To ensure a certain standard of quality, we decided to include only articles that were subjected to a peer-review process.

The aim of our search was to identify articles that compiled results from intervention studies evaluating the effectiveness of teacher education measures, either in the form of a literature review or meta-analysis. This approach incorporates several aspects that refine the literature selection. Therefore, we decided to run the various queries in WoS using the "All Fields" option to avoid further narrowing the search radius. We used the filters "Document Types" and "Languages" to limit our search to "Article" and "Review Article," as well as to "English" and "German." Additionally, we repeated our search queries for combinations of various synonyms to precisely reflect our research interest without unnecessary restrictions. The various inputs, detailed in Table 1, resulted in different queries. We worked with truncation (i.e., the "*" symbol) to capture the different endings of words in our results. Using the AND operator, we ensured the necessary precision of the query by combining the words. For example, we used the input "chang*" to account for the implicit mention of an effect measurement. We used "teacher training," "[...] education," and "[...] coaching" as synonyms for teacher training measures in the different stages of teacher (initial) training. With "effect size*," we also considered the outcome of a study's evaluation without explicitly referring to an intervention. Similarly, "impact" was used to capture a possible synonym for "effectiveness". These various

Table 1 Inputs used for running queries in Web of Science and ERIC

Web of Science	ERIC ^a
“teacher education” AND effect size* AND effectiveness	“teacher education” AND “effect size” AND “effectiveness”
“teacher education” AND effect size* AND impact	“teacher training” AND “effect size” AND “effectiveness”
“teacher education” AND intervention* AND chang* AND review	“teacher coaching” AND “effect size” AND “effectiveness”
“teacher education” AND intervention* AND chang* AND meta-analys*	“teacher coaching” AND “effect size” AND “impact”
“teacher training” AND effect size* AND effectiveness	“teacher education” AND “effect size” AND “impact”
“teacher training” AND effect size* AND impact	“teacher training” AND “effect size” AND “impact”
“teacher training” AND intervention* AND chang* AND review	“teacher training” AND “intervention” AND “review”
“teacher training” AND intervention* AND chang* AND meta-analys*	“teacher coaching” AND “intervention” AND “review”
“teacher coaching” AND effect size* AND effectiveness	“teacher education” AND “intervention” AND “review”
“teacher coaching” AND effect size* AND impact	“teacher education” AND “intervention” AND “meta”
“teacher coaching” AND intervention* AND chang* AND review	“teacher training” AND “intervention” AND “meta”
“teacher coaching” AND intervention* AND chang* AND meta-analys*	“teacher coaching” AND “intervention” AND “meta”

^aFor a clearer presentation, we have omitted the “subject” or “description” in front of each item

queries yielded a total of $n = 261$ results in WoS (Fig. 3), which we exported to Excel.

To simplify the process of exporting the search results to Excel in ERIC, we utilized the ERIC API. Since the use of truncations is not possible here, we used “meta” instead of “meta-analys*” to keep the search as open-ended as possible. Furthermore, we searched within abstracts by prefixing each search term with the API Attribute Name “description,” as well as within ERIC’s own category “Descriptors” using the API Attribute Name “subject.” The latter refers to “Terms from the ERIC Thesaurus added to records to help identify materials on topics in the field of education” (ERIC 2024). Table 1 also lists the individual queries conducted for both “subject” and “description.” These searches resulted in $n = 233$ and $n = 72$ hits, respectively (Fig. 3), which we also exported to Excel.

These 566 titles were reduced to $n = 350$ after non-peer-reviewed articles and duplicate records were removed. Subsequently, independent screening of the abstracts by two of the authors further reduced the number to 47 titles. The reason for this initial narrowing of the corpus was that it quickly became apparent that the articles were not review articles or meta-analyses or that the overarching focus was on, for example, the medical context. The same members of the authorial team then carried out a full-text screening of the 47 articles. A further ten articles were excluded by this process. The criteria for the articles’ inclusion or exclusion were based on the

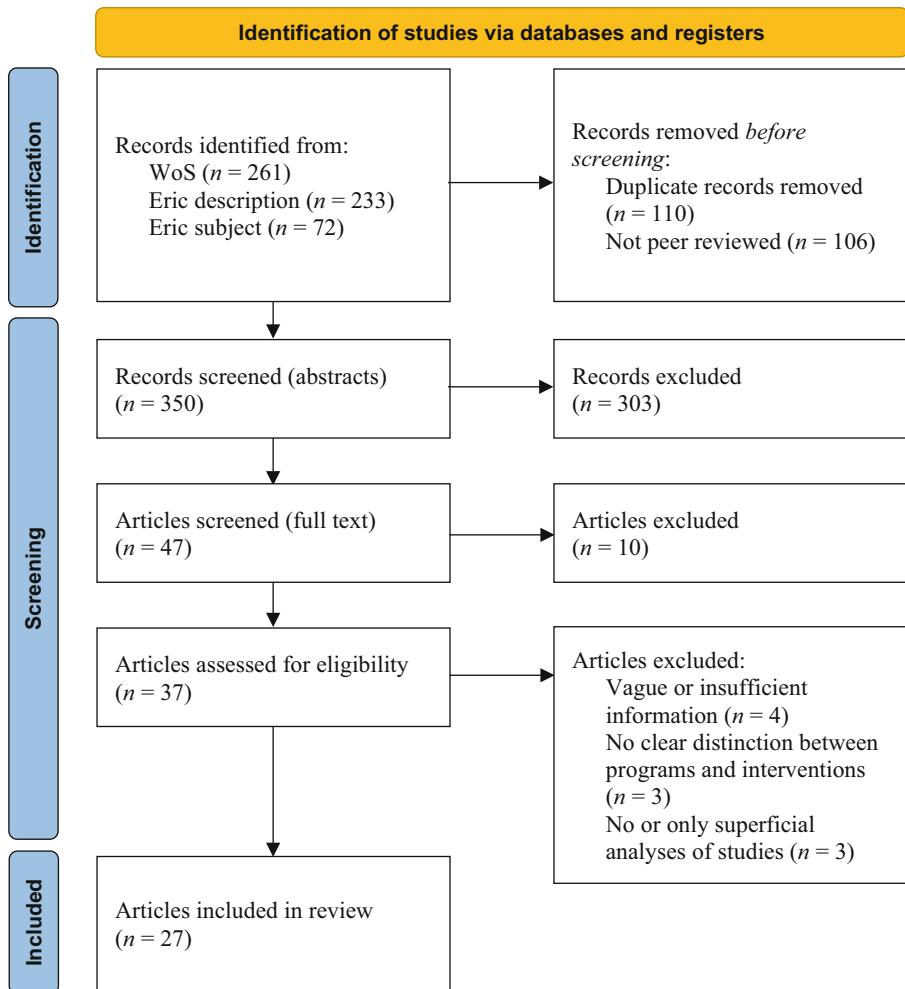


Fig. 3 Flow diagram (following the guidelines of Page et al. 2021)

criteria listed in Table 2. Whereas a single exclusion criterion was already sufficient to deselect articles, those articles to be selected had to fulfil virtually all the inclusion criteria listed in Table 2.

5.2 Description of text corpus

The remaining 37 articles were carefully examined by both authors again to create the final corpus. Of these, four were discarded on the grounds that they provided only vague or insufficient information about the intervention studies (Cummings et al. 2007; Hikida et al. 2019; Thomas and Rieth 2011; Tristani and Bassett-Gunter 2020). Three other articles (Barrio et al. 2015; Dunst et al. 2020; Österling and Christiansen 2022) focused more on programs than on interventions, or the distinction was not

Table 2 Exclusion and inclusion criteria of published reviews and meta-analyses

<i>Exclusion criteria</i>	Focus on early childhood
	Vague or insufficient information about the studies
	No focus on intervention studies
	Different understanding of the term “intervention”—i.e., not related to experimental study design, pre-post etc.
	Review of (school development) programs, single case studies, or student teachers’ projects
<i>Inclusion criteria</i>	No or only superficial analyses of studies
	Initial teacher education (ITE), early career teaching and/or professional development
	Pre-post design and/or group-design (randomized controlled trials and/or quasi-experimental studies)
	If possible (aggregated) effect sizes
	Transparent and clear analyses of studies
	Articles that were subjected to a peer-review process

clearly identifiable, resulting in their exclusion. Given our objective of providing an overarching analysis of the articles in this paper, we excluded three articles that contained no or only very superficial analyses of the studies they addressed (Accurso and Gebhard 2021; Bottiani et al. 2017; Pisaneschi 2020). Thus, the final corpus comprises 27 systematic reviews, scoping reviews, or meta-analyses, which we shall examine in greater detail below.

As a first step, we classified the 27 reviews with respect to their topical focus (Table 3). It transpired that approximately half of all reviews (14 out of 27) could easily be classified to the area of digitalization or the area of inclusive education, with 7 reviews representing each demand imposed on teachers in recent decades. For the remaining reviews, it was then possible to group two or three reviews each into the following areas: language education (3 reviews), research-based teacher approaches (2 reviews), classroom management (2 reviews), coaching (2 reviews), and STEM (2 reviews). The topical foci of teacher burnout and family engagement were represented by only one review each.

Our synthesis builds on a broad understanding of teacher education—that is, comprising ITE and TPD—and therefore includes reviews of interventions targeting pre-service and/or in-service teachers. As Table 3 illustrates, this broad understanding is consistent with the selected reviews, as more than one third (10 out of 27) shared this understanding with including interventions in ITE and TPD. Most reviews (22 out of 27) included in-service teachers, indicating that interventions might be more prominent in TPD than in ITE. Interestingly, the topical foci of digitalization and inclusive education are clearly directed toward ITE rather than other topical foci, such as teacher burnout or family engagement, that address the need for in-service teachers’ professional development.

In developing and applying our coding procedure (Sect. 5.3), it was necessary to identify and determine the precise number of individual studies underlying conclusions drawn by each review. For reasons of transparency and methodological rigor, we traced back the number of studies, leading, in some cases, to a corrected number of studies (Table 3). For example, the number decreased when we de-selected qualitative studies or survey studies that were also part of the review, and the number

Table 3 Classification of papers by authors, year of publication, area, topical focus, target group, time frame of selected studies, number of studies, review with meta-analysis

Authors (year)	Area	Topical focus	ITE	TPD	Time frame	<i>n</i>	MA
Billingsley et al. (2019)	D	Immersive VR technology	x	x	1999–2013	10 ^a	–
König et al. (2022)	D	ICT integration in lesson plans	x	x	2015–2021	4 ^d	–
Ma et al. (2023)	D	Blended teaching	x	–	2004–2021	15	x
Ning et al. (2022)	D	Teacher TPACK	x	x	2010–2022	59	x
Wilson (2023)	D	Technology integration: attitudes/beliefs	x	–	2007–2016	30	x
Wilson et al. (2020)	D	Technology integration: knowledge	x	–	2007–2016	38	x
Young et al. (2013)	D	Teacher TPACK	x	–	2009–2011	8	–
Alexander et al. (2015)	I	Training teachers in evidence-based practices for individuals with Autism Spectrum Disorder (ASD)	–	x	1977–2013	23	–
Allsopp and Haley (2015)	I	Teach mathematics for students with learning disabilities	x	x	2005–2013	7 ^d	–
Brock and Carter (2017)	I	Implementation of interventions for students with disabilities	x	x	1980–2015	12	x
Dignath et al. (2022)	I	Teachers' beliefs about inclusive education	x	x	2004–2019	22 ^d	x
Khamzina et al. (2024)	I	Attitudes toward inclusion	x	–	1996–2020	36 ^d	–
Kurniawati et al. (2014)	I	Inclusive education	x	x	2000–2011	10 ^d	–
Lautenbach and Heyder (2019)	I	Attitudes to inclusion	x	–	1997–2016	23	–
Basma and Savage (2023)	L	Student reading	–	x	2009–2016	14 ^d	x
Didion et al. (2020)	L	Student reading	–	x	1981–2018	28	x
Kalinowski et al. (2019)	L	Students' academic language proficiency	–	x	2003–2015	38	–
Kretlow et al. (2010)	C	Teacher coaching/evidence-based practices	x	x	1989–2010	13	–
Kraft et al. (2018)	C	Teacher coaching on instruction and achievement	–	x	2006–2018	60	x
Whear et al. (2013)	M	Student SE outcomes/Classroom management	–	x	1970–2013	14 ^b	–
McGuire et al. (2024)	M	Classroom management	x	x	2008–2020	6 ^d	–

Table 3 (Continued)

Authors (year)	Area	Topical focus	ITE	TPD	Time frame	<i>n</i>	MA
Ansyari et al. (2020)	R	Data use on teacher and student outcomes	–	x	2009–2018	17	–
Ansyari et al. (2022)	R	Data use on student learning outcomes	–	x	2009–2018	10 ^d	x
Zhang et al. (2023)	S	Teacher perceptions about STEM	x	x	2011–2021	23	x
Zhou et al. (2023)	S	STEM teacher self-efficacy	–	x	1997–2022	21 ^c	x
Oliveira et al. (2021)	B	Social/emotional learning on burnout	–	x	2013–2020	13	x
Smith and Sheridan (2019)	F	Family Engagement	–	x	1990–2016	39	x

Area: *D* digitalization, *I* inclusive education, *L* language education, *R* research-based teacher approaches, *S* STEM education, *C* coaching, *M* classroom management, *B* teacher burnout, *F* family engagement. *ITE* initial teacher education (target group: student teachers, pre-service teachers, teacher candidates etc.), *TPD* teacher professional development (target group: in-service teachers), *MA* meta-analysis, *n* number of studies

^aOne of the 7 articles includes 4 empirical studies based on separate samples, therefore 10 studies

^b14 publications including 8 different interventions

^c21 studies including 48 effect sizes (effect sizes are not to be seen as individual studies as the same studies was counted based on multiple outcomes)

^dNumber of intervention studies selected

increased when, for example, one publication consisted of several independent intervention studies that had to be counted individually. Further details may be found in the notes below Table 3. Overall, all individual studies identified in all 27 reviews amount to 593 intervention studies (the sum of the penultimate column in Table 3). These 593 studies are derived from 581 papers (without accounting for the overlap between the reviews).

Considering the overlap (i.e., individual publications that appeared in more than one review), 510 different publications were identified to form the basis of all reviews. 453 of these papers (89%) were only selected by one of the 27 reviews, whereas 45 (9%) were selected by two reviews, 10 (2%) by three reviews, and 2 (<1%) by four reviews ($581 = 453 * 1 + 45 * 2 + 10 * 3 + 2 * 4$ and $510 = 453 + 45 + 10 + 2$). This overlap across all 27 reviews is relatively limited but will be accounted for in the subsequent meta-meta-analysis.

5.3 Data analysis

5.3.1 Development of analysis categories and coding procedure

Based on a comprehensive examination of the reviews, we developed a coding scheme for the purpose of extracting information relevant to our research questions. The coding categories (Table 4) were newly developed specifically for the synthesis of the 27 reviews. Besides coding participants (pre-service vs. in-service teachers),

Table 4 Analysis dimensions and categories for coding publications with reference to research questions (including item format and with reference to findings presented in Tables and Figures)

Research question	Dimension	Category	Item format	Table (Fig.)
–	Participants	Pre-service teachers	Yes/No	Tab. 3
		In-service teachers	Yes/No	
1) Justifications and theoretical frameworks	Reason for intervention studies	Curriculum innovation	Yes/No	Tab. 5
		Need for causal evidence/(quasi-) experimental studies	Yes/No	
	Theoretical framework	Effect chain assumption	Yes/No	
		Theoretical framework referenced	Yes/No	
		Theoretical framework developed	Yes/No	
2) Input and output	Input— <i>Content</i>	Area of content, topical focus	Open	Tab. 3
	Input— <i>Structure</i>	Duration—Length of shortest intervention in the review	Open	–
		Duration—Length of longest intervention in the review	Open	
	Input— <i>Didactic</i>	Course work	Multi-level	Tab. 6
		Teaching practice	Multi-level	
		Digital resources	Multi-level	
	Output— <i>Teacher competence</i>	Change of knowledge	Multi-level	Tab. 7
		Change of skills	Multi-level	
		Change of attitudes/beliefs	Multi-level	
	Output— <i>Process</i>	Teaching	Multi-level	
	Output— <i>Students</i>	Student learning	Multi-level	
		Study design	Multi-level	Tab. 8
	Assessment quality	Pre-post design	Multi-level	
		Control group included	Multi-level	
		True experimental design with randomized grouping	Multi-level	
		Multiple treatments	Multi-level	
		Self-report instruments	Multi-level	
4) Effect analysis	Effects	Proficiency testing	Multi-level	
		(Standardized) observation	Multi-level	
	Moderators	Aggregated effect size(s)	Open	(Fig. 4–8)
		Significant moderators/covariates	Open	Tab. 9
5) Research gaps		Non-significant moderators/covariates	Open	
	Theory development		Open	–
	Methodology		Open	
	Topical focus		Open	

In case findings are not presented in Tables or Figures, see corresponding sections for description of findings.

the coding categories addressed *justifications and theoretical frameworks* (RQ1), such as reasons or purposes for conducting the respective literature review, the *input and output of the intervention studies* (RQ2), including characteristics of the interventions and the targeted learning outcomes, the *designs of the reviewed studies* (RQ3)—for example, quasi-experimental vs. experimental designs, i.e., a distinction important to account for different evaluation designs with and without control groups, pre-testing, or randomization—the *effect sizes and covariates* (RQ4) in the included meta-analyses (relating to a subsample of 14 reviews only), and the *research gaps* (RQ5) identified by the papers with respect to theory building, methodology, and specific aspects pertaining to the topic addressed.

Owing to the reviews' different topical foci, individual open categories were used to extract certain data (e.g., regarding theoretical assumptions or research gaps). Where possible, a closed, multi-level coding format was used to assess the extent to which a category applied to the synthesized individual studies within a review. Specifically, coding recorded whether a category applied to none of the studies (0 = none), to a single study (1 = yes, one only), to two or more studies (2 = yes, some), or to all studies within the respective review (3 = yes, all). The coding was primarily based on the information provided in the respective review articles, with occasional reference to individually included studies in cases of ambiguity.

Five research assistants and one student assistant (all members of the authorial team) carried out the coding. During the coding scheme's development, the categories were first discussed and then applied by all coders to a single paper to test their applicability. The subsequent comparison of the coding revealed a largely shared understanding of the coding categories. Following the procedure of Kraft et al. (2018), the quality of coding was ensured by critical review: each review article was completely and independently coded by two coders. Coding was then compared and discrepancies discussed until a consensus was reached. In cases of uncertainty, a third coder was consulted, who then reapplied the disputed categories to the paper in question.

5.3.2 Meta-meta-analysis

Related to our RQ4, 14 of the 27 reviews were classified as meta-analyses (Table 3, last column). These meta-analyses report standardized estimates referring to either pre-post comparisons (mean change), comparisons between treatment and control group (mean difference), or differential developments of treatment and control groups (difference in mean change). To derive an overarching estimate of the interventions' effectiveness, we extracted data on the number of primary studies included and the total number of participants within these studies as well as effect sizes, variances, and confidence intervals. For standardization purposes, all reported effects were converted into standardized mean differences (SMD), quantified as Hedges' g , using the formulas provided by Borenstein et al. (2009). Where necessary, effect sizes were adjusted so that positive effects represent desirable outcomes (i.e., reduced burnout, decreased anxiety).

In a meta-meta-analysis, the statistical methods employed in the first-order meta-analysis may be applied at the second-order level (e.g., Türk et al. 2023). To summa-

size the effects, random effects models were estimated based on restricted maximum likelihood (REML) estimation using the R package *metafor* (Viechtbauer 2010). In a random effects model, the true effects are assumed to vary between studies, which is an appropriate assumption for the present purposes, given the different interventions and constructs.

Initially, (1) a comprehensive model was estimated to capture the overall effect size across all meta-analyses. Four further specific models were examined, each focusing on different dimensions of effectiveness: (2) knowledge, (3) attitudes, beliefs, and affect, (4) teachers' behavior (e.g., instructional practices), and (5) student learning. Only one effect per meta-analysis was considered for each of the five models. If multiple effects were reported within a single meta-analysis for a given dimension of effectiveness, the overall effect was included, if available, or the effects were aggregated (weighted by inverse variance, e.g., Borenstein et al. 2009).

To estimate the overall effect, it was necessary to account for the overlap between meta-analyses (i.e., primary studies appearing in multiple meta-analyses). A total of 348 primary studies were included, with 22 included twice, one study three times, and two studies four times, indicating a small overlap of primary studies between the selected meta-analyses. The procedure proposed by Munder et al. (2013) was employed for this purpose. For each primary study, the number of meta-analyses in which it appeared was determined, and the inverse of this number—representing the uniqueness of each study—was summed for each meta-analysis to obtain the adjusted number of primary studies included. This procedure was conducted separately for each metanalytic model ((1) to (5)). The weight of each meta-analysis for the overall effect in each model was then calculated based on the adjusted number of primary studies and the variance of the “true” effect between meta-analyses (detailed in Mingeback et al. 2018; Munder et al. 2013). This approach ensures that each primary study contributes only once to the overall effect.

Effect sizes were interpreted based on Cohen's (1988) classification (≥ 0.2 : small; ≥ 0.5 moderate, ≥ 0.8 : large). Q test and I^2 statistic were used to estimate the heterogeneity. I^2 in the context of a meta-meta-analysis quantifies the proportion of variance that exists between the meta-analyses themselves rather than between the individual studies within each meta-analysis. It may be interpreted as low (25%), moderate (50%), or high (75%) (Higgins et al. 2003). Analyses of publication bias may be found in the Supplementary Material.

6 Findings

6.1 Justifications and theoretical frameworks (RQ1)

In light of our RQ1's concern with *justifications and theoretical frameworks*, we were interested in the reviews' theoretical outlines, such as the justifications for conducting the respective literature reviews. As outlined in Sect. 2.1, at least two major factors appear suited to justifying interventions in teacher education. First, it is necessary to update or innovate the teacher education curriculum. Intervention studies introducing new content and methods may generate the specific knowledge

required for future curriculum design. Second, interventions applying experimental study designs facilitate testing of the causality of relationships previously investigated as mere statistical correlations in cross-sectional or longitudinal survey studies. Randomized control trials have the potential to yield meaningful insights into effects that can only be assumed in correlational survey statistics. Meanwhile, we found that while almost all reviews mentioned or clearly outlined the first reason, less than half of the reviews detailed their reasons for justifying their review to the need for providing causal evidence (see Table 5).

In principle, all reviews assume the existence of an effect chain—namely, that the intervention promotes teacher variables (mostly competence facets), which, in turn, should be positively reflected in teachers' (increased) teaching quality and students' (increased) learning gains. This becomes evident in some reviews—for example, those that explicitly refer to the key components of the frequently cited effect chain model developed by Desimone (2009) to examine the impact of TPD on specific outcomes (e.g., Ansyari et al. 2020, 2022; Basma and Savage 2023; Didion et al. 2020; Kalinowski et al. 2019; Zhou et al. 2023). Other reviews provide outlines that also highlight the authors' embedding their review in the assumption of such an effect chain's existence, whereas only three reviews did not explicitly mention this notion (Brock and Carter 2017; Ning et al. 2022; Zhou et al. 2023).

The authors of almost all reviews include definitions, models, concepts, or theories relevant to the topic at hand. By contrast, few reviews developed their own

Table 5 Findings on justifications and theoretical frameworks (absolute frequencies derived from coding)

Dimension	Analysis category	Short description	Yes	No	Unclear
Reason for intervention studies	Curriculum innovation	Authors justify the necessity of interventions by focusing on a topic that has not been adequately covered in teacher training (curriculum) and/or reflects more recent societal-cultural developments, leading to the need for focusing on this topic in interventions	23	3	1
	Need for causal evidence	Authors justify the necessity of interventions by highlighting that experimental designs are needed to provide causal evidence regarding effects (which might have previously been found in survey studies/non-experimental studies)	11	16	0
Theoretical framework	Effect chain assumption	Authors mention and apply the notion of an effect chain (e.g., teacher education—teacher competence—teaching—student learning) to the review topic	24	3	0
	Theoretical framework referenced	Authors reference definitions, models, concepts, theories relevant for the topic of the review (e.g., TPACK)	26	1	0
	Theoretical framework developed	Authors present a substantial theoretical framework they developed in their paper, reflecting a level more general than a single study's scope	4	23	0

Coding recorded whether a category was fulfilled (*yes*) or not fulfilled (*no*) by the respective review. If there was insufficient information for such coding, the review was coded as *unclear*.

theoretical framework (e.g., “framework for evaluating the effects of data use PDIs on teacher and student outcomes” by Ansyari et al. 2020, 2022; “heuristic for summarizing key findings on ICT integration in lesson plans” by König et al. 2022; “theory of action for teaching coaching” by Kraft et al. 2018). For example, among the reviews focused on digitalization (see Table 3, Area D), only that on ICT integration in lesson planning (König et al. 2022) created a heuristic for summarizing and conceptualizing its overall findings. No further approaches or models were developed for theoretical framing of intervention studies in teacher education digital learning opportunities. The TPACK model, portraying the professional knowledge that teachers should possess (Mishra and Koehler 2006), is by far the most frequently applied and referenced theoretical framework (e.g., Billingsley et al. 2019; König et al. 2022; Ning et al. 2022; Wilson et al. 2020; Young et al. 2013), and various reviews cite a few more models individually. Overall, the theoretical grounding for the reviews is weak, which should be acknowledged as a research gap to be addressed by future research.

6.2 Input and output of the intervention studies (RQ2)

With our RQ2, we focus on the *input* and *output* of the intervention studies—that is, basic characteristics of the interventions and the targeted learning outcomes. Regarding *content*, that is, topical areas as part of the input was already outlined and described in Sect. 5 (see Table 3). Regarding *structure*, the minimum and maximum intervention length was coded for each review. While this information does not allow for an estimate of central tendency, it is possible to illustrate the range or variability on this basis. To facilitate comparison, duration was recoded into the number of weeks. This ranges from one to many weeks to interventions that lasted several months, one year, or even several years. Generally, the interventions varied considerably with respect to duration, making it difficult to draw any reliable conclusions across the various reviews.

Consideration of the methods used (Table 6) as part of *didactic* features reveals that interventions are typically organized as coursework (or a training similar to coursework etc.), since all reviews include some or even all studies comprising coursework—with the exception of only one review, in which we found that just one single study included coursework (Allsopp and Haley 2015). This seems unusual given the dissemination of this intervention format. By contrast, approximately a quarter of the reviews included no interventions with a connection to teaching practice, and only two reviews built on a selection of interventions with the requirement to link the treatment with an examination of occurrences in the classroom (Kraft et al. 2018; Kretlow et al. 2010). The use of digital resources is evident in approximately two thirds of all reviews. Those five reviews in which all interventions included utilized digital resources (see Table 6, column “All”, last row) had their topical focus in the area of digitalization (see Table 3). In contrast to this, six reviews had not and two more reviews had not clearly included intervention studies with reference to digital resources, among them the two reviews in the area of classroom management (McGuire et al. 2024; Whear et al. 2013).

Table 6 Findings on basic characteristics (*didactic* features) of interventions as input (absolute frequencies derived from coding)

Dimension	Analysis category	Short description	All	Some	One	None	Unclear
Methods	Course work	Typical training formats similar to academic settings (e.g., seminars, workshops)	13	13	1	0	0
	Teaching practice	Any practical exploration or application in practical school teaching settings involved	2	18	0	7	0
	Digital resources	Digital resources are a decisive element of providing the intervention	5	13	1	6	2

Coding recorded whether a category applied to *none* of the studies, to *one* single study, to *some* (i.e., two or more studies), or to *all* studies within the respective review. If there was insufficient information for such coding, the review was coded as *unclear*.

Regarding the operationalization of the interventions' targeted learning outcomes, which were assigned to a rough structure that assumed an effect chain of teacher competence—teaching—student learning (Blömeke et al. 2022), the findings presented in Table 7 reveal heterogeneity in the individual reviews' focus: Each analysis category was wholly ignored by approximately one third to two thirds of the reviews (see Table 7, last column “None”: 11 for knowledge, 14 for skills, 12 for attitudes/beliefs, 17 for teaching (with one unclear), 13 for student learning). It appears that when a review focuses on one of the learning outcomes, such as a change in teacher knowledge, other outcomes (e.g., teaching) are not necessarily included. This may

Table 7 Findings on targeted learning outcomes of interventions (absolute frequencies derived from coding)

Dimension	Analysis category	Short description	All	Some	One	None	Unclear
Teacher competence	Change of knowledge	Teacher knowledge (e.g., pedagogical content knowledge) is part of outcome measurement	3	11	2	11	0
	Change of skills	Teacher skills (e.g., lesson planning skill) is part of outcome measurement	1	11	1	14	0
	Change of attitudes/beliefs	Teacher attitudes (e.g., toward inclusive teaching) is part of outcome measurement	5	8	2	12	0
Process	Teaching	Teacher behavior, teacher performance in class, or instructional quality and practice are part of outcome measurement	2	6	1	17	1
Students	Student learning	Student learning data is part of outcome measurement	4	10	0	13	0

Coding recorded whether a category applied to *none* of the studies, to *one* single study, to *some* (i.e., two or more studies), or to *all* studies within the respective review. If there was insufficient information for such coding, the review was coded as *unclear*.

reflect the fact that teacher education's effectiveness may be demonstrated using multidimensional conceptualizations of criteria (König and Blömeke 2020).

6.3 Designs of the reviewed studies (RQ3)

With our third research question, we related our research to the *designs of the reviewed studies*. We included only those studies that clearly applied a study design that would facilitate examination of the intervention's potential effects. Therefore, randomized controlled trials (RCTs) or (quasi-)experimental study designs (QEDs) and pre-test-post-test designs were included while survey studies or qualitative studies were excluded, sometimes resulting in corrections to the number of studies covered by the single review (Table 3). The number of studies was also corrected to indicate whether a subsample of studies was used solely to compute the overall effect size (e.g., Basma and Savage 2023).

On the one hand, the majority of reviews included only—or at least some—intervention studies that implemented a pre-post design (all: 12; some: 10; total: 22), and also a majority of reviews included only or at least some intervention studies with control group (all: 11; some: 9; total: 20). These findings reveal, on the other hand, that the *basic characteristics* of intervention study design (i.e., pre-post, control group) are a consistent component of teacher education effectiveness research, even if not wholly implemented in all intervention studies selected by the reviews. This leaves room for methodological improvement in any future research effort (Hill et al. 2021). More *challenging characteristics* of intervention study design (true experimental design with randomized grouping, multiple treatments) appear to be less common in intervention studies. No review uses these criteria in relation to their specific selection of empirical studies. In particular, multiple treatments—that is, the comparison of different forms of treatment implementation (e.g., comparing different feedback formats or frequencies)—appears to be highly unusual, given that 22 of the 27 reviews either provided the information that none of the intervention studies included had multiple treatments (19 reviews) or remained unclear on this issue (3 reviews).

Regarding assessment quality, we sought to determine how essential approaches such as proficiency testing (e.g., to directly assess teacher knowledge) and (standardized) observation measures (e.g., to capture instructional quality) were for selecting intervention studies based on the reviews. First, more than half of the reviews do not or do not clearly inform the reader about these approaches. Only a small number of reviews indicate that all intervention studies included implemented proficiency testing (Ansyari et al. 2022; Didion et al. 2020; Ning et al. 2022) or standardized observation measures (Kretlow et al. 2010), whereas a greater number of reviews indicated that at least some of the selected studies contained such assessment features. By contrast, self-report instruments (rather than proficiency testing or standardized observation) appeared to be common among the intervention studies included in the reviews (Table 8).

Table 8 Findings on designs of reviewed intervention studies (absolute frequencies derived from coding)

Dimension	Analysis category	Short description	All	Some	One	None	Unclear
Study design	Pre-post design	Data on teacher measures and/or teaching measures measured before and after a treatment, including designs with waiting control group(s)	12	10	0	1	4
	Control group included	Treatment group, which participated in intervention, is compared to a control group	11	9	2	1	4
	True experimental design with randomized grouping	True experimental design (Campbell and Stanley 1963, p. 8) with comparison of two or more groups regarding teacher and/or teaching measurement data, based on randomized grouping	0	17	0	7	3
	Multiple treatments	Distinction between a single treatment or a concurrent treatment	0	4	1	19	3
	Self-report instruments	Interventions effects assessed using self-report measures	8	6	1	8	4
Assessment quality	Proficiency testing	Intervention effects are assessed using proficiency testing	3	10	0	11	3
	(Standardized) observation	Intervention effects are assessed using (standardized) observation measures	1	7	1	13	5

Coding recorded whether a category applied to *none* of the studies, to *one* single study, to *some* (i.e., two or more studies), or to *all* studies within the respective review. If there was insufficient information for such coding, the review was coded as *unclear*.

6.4 Effect sizes and covariates (RQ4)

With RQ4, we examine the *effect sizes and covariates* and carry out a meta-meta-analysis for a subgroup of 14 reviews (Table 3, last column) that exhibit summarized effect sizes. For this purpose, the summarized effects of the individual meta-analyses and their variances were extracted or determined as part of the coding procedure (Table 4). An overview of the meta-analyses and information to the extracted outcomes may be found in Supplementary Material, Table S1. The following results are presented using forest plots, whereby each individual effect is represented by a square accompanied by its confidence intervals. The square's size reflects the weight, determined by the adjusted number of primary studies. The overall effect is illustrated by the so-called “diamond” at the bottom of each plot (Figs. 4, 5, 6, 7 and 8).

Model 1 represents the overall effect, incorporating all 14 meta-analyses into the total (see Fig. 4). Overall, a significant moderate effect of teacher education interventions is observed (SMD 0.55, 95% CI 0.35–0.74, $p < 0.001$). Heterogeneity was very high, which is unsurprising, given the studies' diverse focal topics ($Q[df=13]=413.52$, $p < 0.001$, $I^2=98.75\%$). Consideration of the intervention effects on knowledge (Fig. 5) reveals a significant strong overall effect (SMD 0.89, 95% CI 0.76–1.02, $p < 0.001$). The Q test indicated a relatively homogeneous effect size

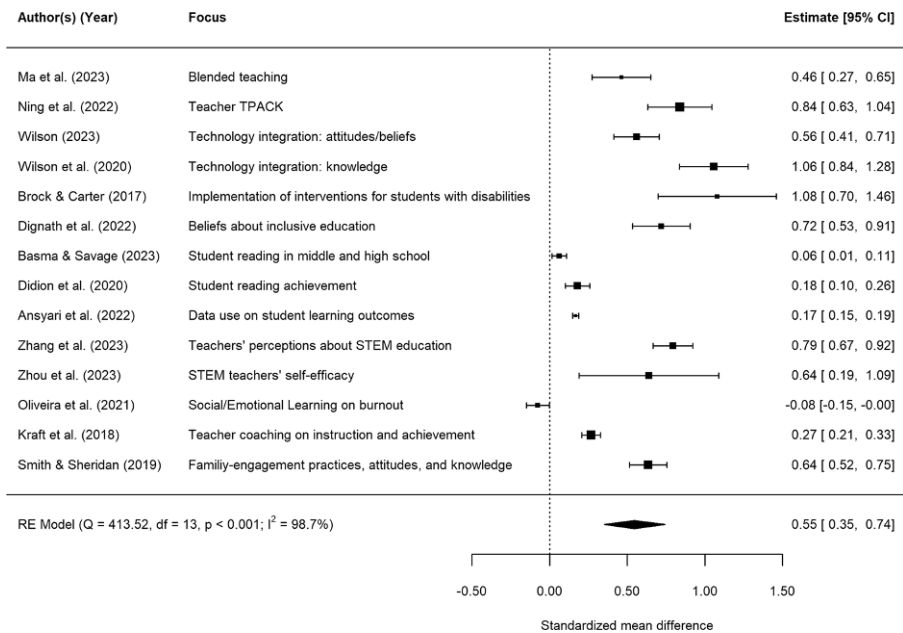


Fig. 4 Forest plot of overall intervention effects

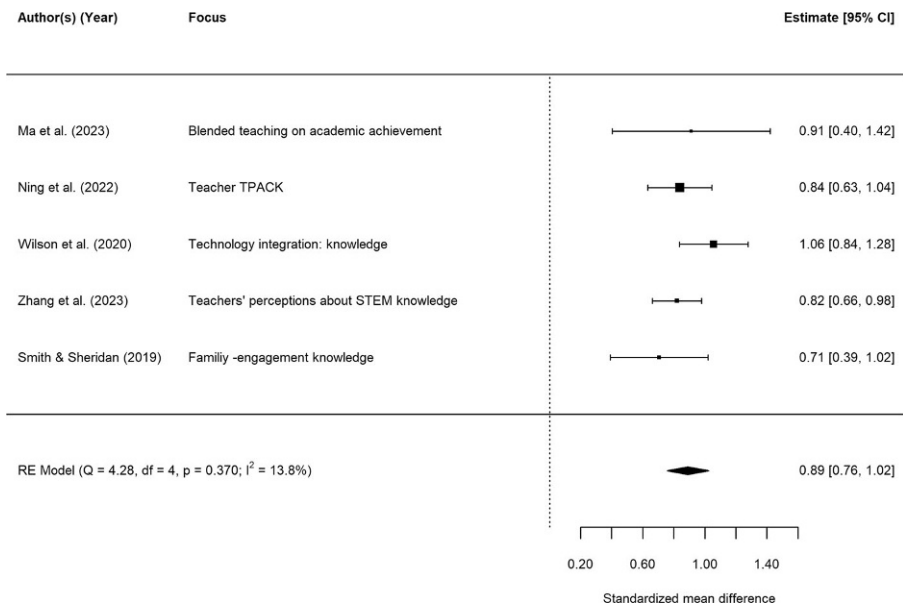


Fig. 5 Forest plot of intervention effects on knowledge of teachers

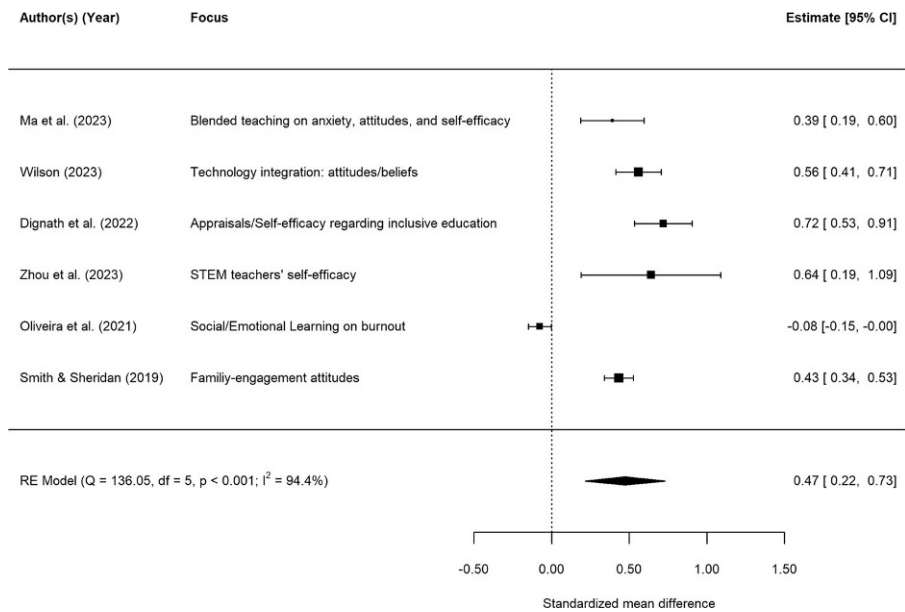


Fig. 6 Forest plot of intervention effects on teachers' beliefs, attitudes, and affect

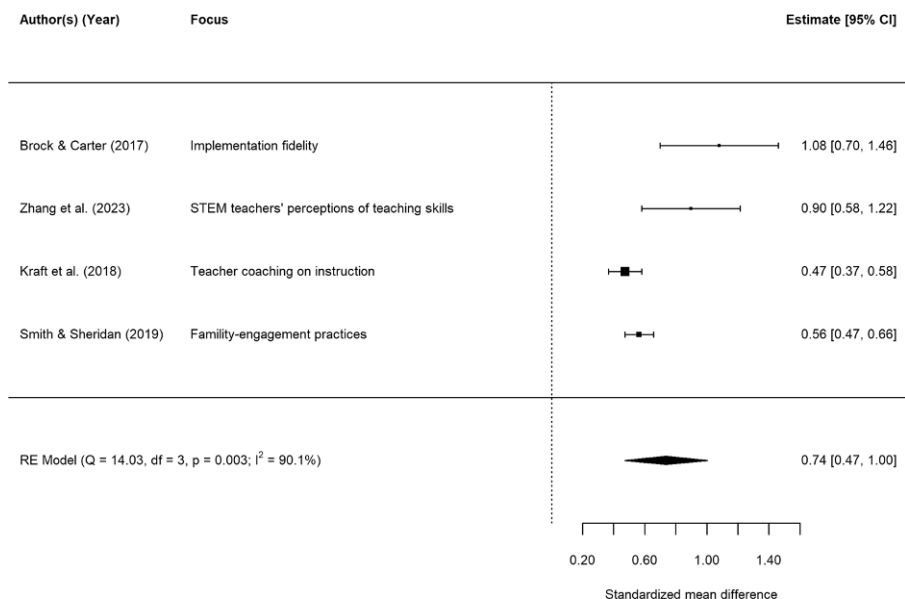


Fig. 7 Forest plot of intervention effects on teachers' behavior, instructional quality or practice

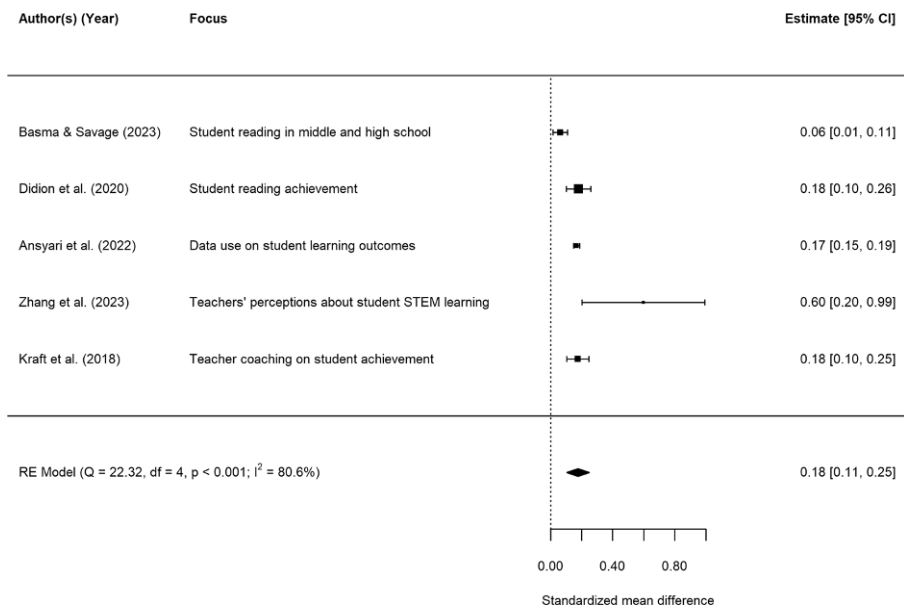


Fig. 8 Forest plot of intervention effects on student learning

across studies ($Q[df=4]=4.28$, $p>0.05$, $I^2=13.82\%$). The effect on beliefs, attitudes, and affect (Fig. 6) is notably smaller but remains significant (SMD 0.47, 95% CI 0.22–0.73, $p<0.001$). In contrast to knowledge, the heterogeneity is very high ($Q[df=5]=136.05$, $p<0.001$, $I^2=94.40\%$).

Regarding behavioral measures, such as teachers' behavior, instructional quality, or practices (Fig. 7), a significant effect is observed (SMD 0.74, 95% CI 0.47–1.00, $p<0.001$), with considerable variation across the meta-analyses ($Q[df=3]=14.028$, $p<0.01$, $I^2=90.13\%$). By contrast, the effects on student learning (Fig. 8)—a more distal measure of effectiveness—are considerably smaller (SMD 0.18, 95% CI 0.11–0.25, $p<0.001$) but nonetheless significant. Again, heterogeneity is very high ($Q[df=4]=22.32$, $p<0.001$, $I^2=80.55\%$).

Potential explanations for the observed heterogeneity in effect sizes may initially be found in the moderator analyses of the individual meta-analyses. The moderators considered commonly include the interventions' specific characteristics, particularly their duration, participant characteristics (e.g., demographic factors), and aspects of study quality (e.g., research design, measurement quality, and overall rigor).

Table 9 presents a synthesis of moderator effects examined across multiple meta-analyses. Several publications focus on the (methodological) quality of the studies included in these reviews. While significant effects are rarely observed, initial evidence suggests that true experimental designs tend to yield stronger effects than do quasi-experimental designs (Didion et al. 2020; Ning et al. 2022). Considering the participant characteristics and the interventions' features, it is generally noteworthy that moderator analyses often do not reveal significant effects. For instance, effect sizes show no clear dependence on the participating teachers' teaching experience.

Table 9 Synthesis of moderator analyses conducted in the 14 reviews that included meta-analytic results

Moderator	Effect	Ambiguous	No effect	Number of reviews
Study quality/Methodological quality				
<i>Overall</i> (higher quality > lower quality)	1	1	3	5
<i>Research design</i> (experimental > quasi-experimental)	2	0	1	3
<i>Measurement quality</i> (higher quality/standardization > lower quality/ standardization)	1	1	3	5
Participants' characteristics				
<i>Career level</i> (In-service > pre-service)	1	0	2	3
<i>Teaching experience</i> (in years; larger effect for higher experience)	0	0	2	2
<i>School level</i> (e.g., differences between high school, middle school, elementary school)	1	0	2	3
Intervention features				
<i>Duration</i> (larger effects for higher duration)	1	2	6	9
<i>Mode</i> (in person > online)	0	0	2	2
<i>Field experience</i> (field experience > no field experience)	1	0	2	3
<i>Coaching</i> (coaching > no coaching)	0	0	3	3

Ambiguous means that interpretability remains unclear. For example, the effect is only significant for one of several constructs. "Number of reviews" refers to the total number of meta-analyses in which the respective moderator was examined.

Similarly, studies focusing on more specific features of individual interventions, which are not reiterated here, frequently fail to identify significant effects (Didion et al. 2020; Wilson 2023; Wilson et al. 2020). However, significant effects may be more likely to be detected when the moderators are closely aligned with the meta-analysis' core focus (e.g., coaching: Kraft et al. 2018; STEM education: Zhou et al. 2023).

The most frequently examined overarching factor is the duration of professional development, which is the focus of nine meta-analyses. Ma et al. (2023) demonstrated the positive effects of longer professional development programs, although this was only observed for academic performance and not for attitudes or self-efficacy. Similarly, Basma and Savage (2023) found positive effects regarding student learning progress as a result of teachers' participation in longer training. However, Zhou et al. (2023) reported a negative effect of extended intervention duration but highlighted an important interaction: the negative impact of duration only emerged when the intervention did not focus on STEM-specific pedagogy. This suggests that content-related issues in interventions may exert a particularly negative effect with increased contact time. The remaining six meta-analyses detected no significant ef-

fect of duration, suggesting that other factors may play a more significant role in the effectiveness of professional development programs (see also Brock and Carter 2017).

6.5 Research gaps (RQ5)

Research gaps relate to the further development of theory, the consideration of research methodological standards of the individual intervention studies as well as the need to improve the methodological quality of literature reviews (e.g., the reporting of effect sizes) and the need for further investigations in the area of the topical focus.

Regarding theory development, few reviews offer recommendations for future research avenues. For example, Wilson (2023; Wilson et al. 2020) highlights the need for a more comprehensive conceptualization of the teacher knowledge base (Wilson et al. 2020, p. 12) and a more focused framework of technology integration and the impact of specific course features (Wilson 2023, p. 274). Young et al. (2013, p. 167) emphasize the need for theories that will help to explain the findings they report, and, similarly, Lautenbach and Heyder (2019, p. 246) advocate for theories aimed at better understanding how attitudes toward inclusion may be changed in the context of university education. Additionally, the gaps between policy and practice are stated as a problem that theory development should be targeted at (Dignath et al. 2022; Whear et al. 2013). However, most reviews do not discuss future theory development as a specific research gap.

The authors of several reviews conclude that future studies should fulfill high-quality methodological standards. Regarding the design of the single empirical intervention studies, pre-post studies should be combined with control groups and variables moderating the effects of teacher education interventions should be captured. As Ning et al. (2022, p. 17), for example, state, “this will improve our understanding of the relationship between teacher education interventions and their pre- and post-variables.” Examination of the combinations of intervention features and their interaction is also required. Furthermore, many empirical intervention studies fail to consider the long-term effects. As several reviews note, a total operationalization of the effect chain is required—for example, longitudinal analysis of how changes in teacher outcome measures affect further teaching, the transition into practice, connecting the participants’ scores of an intervention to their actual practice and the student learning progress. Those reviews providing meta-analyses criticize the fact that single empirical intervention studies often lack decisive statistical information. Therefore, authors recommend improved reporting, with more consistent reporting of results so that future meta-analyses procedures are supported. Important statistical information is sometimes missing from studies. Another problem is the quality of assessment of constructs. For example, self-report measures may be biased due to social desirability. In general, rigorous measurement of constructs are needed—not least because psychometric problems were observed for specific scales. Researchers should reach a robust consensus on common outcomes and implement common measures to facilitate future meta-analyses. Finally, some reviews highlight the need

for improved measures of fidelity of implementation to ensure that TPD goals are clearly outlined and followed through as intended.

Regarding the reviews' topical focus, many argue that future research should probe the topic at hand in depth and generally expand the research in the specific area. For example, in their review of blended learning intervention studies, Ma et al. (2023) highlight the need for comparisons of online teaching and blended teaching to consolidate the topical focus. Generally, authors recommend further research to produce detailed findings or elucidate (highly) specific aspects that are underrepresented in current research. Other reviews argue that further contexts should be accounted for, such as the influence of culture and policy on teachers' attitudes toward inclusion (Khamzina et al. 2024; Lautenbach and Heyder 2019) or the economic framing of interventions in classroom management in school settings (Whear et al. 2013). Few reviews suggest increasing the number of intervention studies among the target group of pre-service teachers and teachers of alternative certification programs (e.g., Smith and Sheridan 2019; Zhou et al. 2023).

7 Discussion

The aim of this review was to provide an overview of research on the effectiveness of teacher education using intervention studies. Teacher education was understood in a broad sense—that is, including both ITE and TPD. Given the existence of not only numerous intervention studies but also an increasing number of literature reviews that summarize topical selections of individual studies, we performed a systematic literature search to identify such reviews, comprising systematic reviews, scoping reviews, and meta-analyses. Our intention was to adopt an overarching perspective that, to date, has rarely been elaborated in the literature—with the purpose to generate and discuss overview findings on (1) justifications and theoretical frameworks, (2) input and output of the intervention studies, (3) study design, (4) effect sizes, and (5) research gaps.

In this way, we identified 27 reviews published within the last 15 years. The majority related to topics that have attained particular prominence in the educational context in recent years: inclusive education and digitalization. However, other key subject areas, such as language education and STEM, also featured prominently in the reviews were identified. This confirmed that research into the effectiveness of teacher training and professional development is definitely following the thematic development of central educational science discourses.

With RQ1, we asked which frameworks exist and are used for intervention studies in teacher education and to support the characteristics of the specific interventions with reasons (e.g., curriculum innovation). Our findings revealed that almost all reviews mentioned the need to update or innovate the teacher education curriculum as a justification for intervention studies. By contrast, less than half of the reviews related their reasons for justifying their review to the need to furnish causal evidence. With the exception of three reviews, all publications clearly assumed the existence of an effect chain—either with reference to well-known frameworks (e.g., Desimone 2009) or described by the authors. In almost all reviews, the authors in-

clude definitions, models, concepts, or theories that are relevant to the topic at hand. By contrast, few reviews actually developed their own theoretical frameworks.

Based on these findings, one may conclude that intervention studies in teacher education generally focus on examining effectiveness rather than using empirical evidence for theory development. The fact that theories are not foregrounded does not appear to be particularly problematic, but actually testing *whether* something works should also be related to theory development, enriching, for example, the scientific discourse surrounding testing *why* something works. One might assume that this also reflects the focus on testing innovations as opposed to testing causal assumptions. Theories should seek to explain rather than describe. This theoretical deficit raises the question of relevant theories that explain pre-service and in-service teachers' professional development that may be expected being referred to by the conceptualization of intervention studies and can be tested using experimental designs.

Surprisingly, whereas specific frameworks (e.g., Desimone 2009) were cited, other existing frameworks of teacher education effectiveness were not. For example, the multi-level framework developed by TEDS-M (Tatto et al. 2008) was not applied by the reviews, for example, to place intervention studies into a larger context of empirical findings on teacher education effectiveness. The question arises whether scholars sufficiently relate interventions in teacher education to survey studies aiming at an examination of teacher education effectiveness on the level of longitudinal or cross-sectional study design (König and Blömeke 2020), for example, to evaluate causal effects across different data sets (e.g., Colnet et al. 2024). Although authors of the selected reviews, when considering the notion of having an effect chain, related their work to the potential connection between teacher education and teaching, the well-known overall framework by Scheerens and Blömeke (2016), which may serve as a prominent example of a specific theoretical framework for the effect chain, was not cited by the reviews. This might show that the notion of having an effect chain possibly lacks further systematic approaches, which might explain why, as we found with findings related to RQ3, there was hardly any comprehensive operationalization of the effect chain's components. Again, these findings generally show how heterogeneous the approaches are that are applied in the field of teacher education effectiveness research.

RQ2 concerned the *input* and *output* of the intervention studies, that means, basic characteristics of the interventions and the targeted outcomes. Regarding the input, we were interested in content, structure, and didactic elements (Kalinowski et al. 2019). As part of content, the topical focus was analyzed and classified into areas (Table 3), showing the dominance of topics in digitalization and inclusive education. Descriptions regarding the duration of the interventions as a structural element of the interventions' input were difficult to summarize, given that the available information prohibits any conclusion as to the extent to which this information serves as a representative measure. The considerable variation in duration made it difficult to draw any reliable conclusion across the reviews, showing that duration might not be easy to rely on (e.g., Kalinowski et al. 2019; Lipowsky and Rzejak 2015). Looking at the didactic elements, the findings clearly show that interventions are typically organized as coursework (or a training similar to coursework etc.). Interventions connected to

teaching practice were missing in approximately one quarter of reviews, and only two reviews built on a selection of interventions with the requirement to link the treatment to an examination of what happens in the classroom. This demonstrates that teaching practice is a substantial component in the characteristics of many intervention studies, but at the same time is not completely operationalized throughout the research. This corresponds to findings reported by previous reviews (e.g., König et al. 2024; Scheerens and Blömeke 2016). Digital resources were part of about two third of intervention studies, predominantly presented by reviews in the area of digitalization, but absent in other areas such as classroom management.

Operationalization of the interventions' targeted learning outcomes, which were assigned to a rough structure based on the assumption of an effect chain as proposed for example by Desimone (2009), appeared to be highly heterogeneous. Clearly, various approaches to investigating the notion of the effect chain exist without any claim of comprehensive measurement. Consequently, when a review focuses on one of the learning outcomes (e.g., change in teacher knowledge), other outcomes may be excluded. Various ways of prioritizing certain outcomes while neglecting others make it difficult to compare findings across the single literature reviews. Differences in operationalization may also result from the lack of a systematic basis, as hardly any reviews developed their own theoretical framework. That the selection of outcomes seems to be heterogeneous might show that there is still the need for developing clear expectations of what ITE and TPD should target. Obviously, further agreement on the outcomes of teacher education has to be reached (Tatto 2021). This might contribute to the ongoing discussion about "what teachers should learn and be able to do" (Darling-Hammond and Bransford 2007) and may support determining the essential role of teacher education in sustaining the teaching profession (Cochran-Smith et al. 2012).

With our RQ3, we wanted to analyze the study designs' methodological rigor. The findings demonstrate that *basic characteristics* of intervention study design (pre-post, control group) are a consistent component of teacher education effectiveness research, even if not completely implemented in all intervention studies selected by the reviews. More *challenging characteristics* of intervention study design (true experimental design with randomized grouping, multiple treatments) appear less common in intervention studies, though. Randomized grouping in educational settings might be difficult to implement (König 2013). Multiple treatments are challenging, since several aspects of an intervention study become more complex, e.g., creating and providing more than one treatment and underlying further treatments with specific hypotheses (e.g., competing hypotheses) testing, increasing sample size, and extending measurement instruments. Compared with single treatment designs, intervention study designs comprising multiple treatments may better contribute to theory development (e.g., Hanushek and Jackson 2013). Therefore, it seems there is room for methodological improvement in future research efforts (Hill et al. 2021). Similarly, assessment approaches are mostly based on self-report instruments whereas more challenging methods, such as proficiency testing (e.g., to directly assess teacher knowledge) and (standardized) observation measures (e.g., to capture instructional quality), were less frequently applied.

Over the last decade, the quality and methods in assessing outcomes in higher education has significantly increased (Zlatkin-Troitschanskaia et al. 2020). The reviews selected for the present synthesis built on intervention studies that were conducted, on average, in the overall time frame between 2000 (min. 1970, max. 2015) and 2017 (min. 2010, max. 2022; see Table 3). Therefore, it is clear that the latest methodological advances in assessing teacher competence or instructional quality (e.g., Yang and Kaiser 2022) could only find limited entry into the text corpus analyzed and synthesized in the present review. This shows how the continuous and potentially systematic accumulation of educational research findings may be delayed—underlining the importance of literature reviews that provide helpful orientation and relevant directions for any future research efforts.

With RQ4, we examined the *effect sizes and covariates* and conducted a meta-meta-analysis for a subgroup of 14 reviews that exhibit summarized effect sizes. It transpired, first, that there was a significant moderate effect of teacher education interventions (SMD 0.55, see Fig. 4). To the best of our knowledge, such an estimation of the interventions in various domains on both teacher and student outcome variables has hardly been reported in educational research. Regarding TPD at least, in his meta-meta-analyses, Hattie (2009, 2023) reported overall effect sizes in the mid-range ($d=0.62$ in 2009, p. 120; $d=0.44$ in 2023, p. 233) that related to both teacher and student learning, which thus might be partly comparable to the overall effect we could find here.

On a differentiated level, which allowed us to examine different constructs, the findings revealed that strong overall effects could be reported for teacher knowledge and teachers' behavior, instructional quality and practices. By contrast, the effects on student learning are considerably smaller, presumably because it serves as a more distal outcome of interventions in teacher education. Attitudes, beliefs, and affects showed moderate effects, indicating that these teacher competence facets may be more difficult to change than, for example, teacher knowledge or instructional practice (König 2012). Finally, moderator effects examined across multiple meta-analyses were synthesized. Surprisingly, few significant effects were observed. However, initial evidence suggests that true experimental designs tend to yield stronger effects than quasi-experimental designs, emphasizing the need for methodological rigor in intervention studies (Hill et al. 2021).

Findings from our meta-meta-analysis of effects are promising, as they show from a more general perspective that interventions in teacher education actually do achieve certain targets and fulfill expectations towards the professional learning of pre-service and in-service teachers. Moreover, especially outcomes such as teacher knowledge and instructional practice seem to be components that can be substantially changed through intervention. This shows not only the potential of practical relevance that interventions and treatments may have, but it supports the assumption that teacher education effectiveness research is valuable in “translating” the specific learning progress among teachers into reportable statistics. However, although theoretically obvious, student outcomes are more distal to any intervention targeting the learning of teachers, thus the effects on student learning are, on average, small, with just too many factors influencing student learning (Hattie 2023; Scheerens and Blömeke 2016).

With our final question, RQ5, we sought to synthesize the findings on research gaps that, according to the authors of the 27 reviews, future intervention studies should assess. Very few reviews addressed theory development, which is surprising, given the findings of the present synthesis (see RQ1). By contrast, the authors of several reviews conclude that in future research, studies should meet high-quality methodological standards. Study design should be enhanced by more complex elaboration—for example, by including control groups, accounting for operationalization of moderators, measuring multiple outcomes, and implementing the analysis of long-term effects. Moreover, further intervention studies that comprise full operationalization of the effect chain are required. To facilitate future meta-analyses, publications should provide decisive statistical information, and consistent reporting of intervention study results should be encouraged. The topical focus should, according to the authors of the 27 reviews, further be addressed by future research, whereby then the particular topic should be looked at in depth by future studies and it should be generally expanded.

Although there are substantial research gaps showing clear limitations of the current state of research, it is promising that these gaps can be identified to a remarkably detailed level, thus extending the discussion on methodological limitations of intervention studies in teacher education that had already been outlined by Hill et al. (2021). If future intervention studies in teacher education use such recommendations to improve methodological approaches, there is good reason to assume that teacher education effectiveness research, which has been identified as an emerging research paradigm (König et al. 2024), can further be profiled. In any case, the number of literature reviews summarizing the findings from intervention studies in teacher education will further increase over the upcoming years, making necessary further meta-meta-analyses as provided in the present synthesis.

8 Limitations and outlook

Our approach to synthesize literature reviews and summarizing the results of meta-analyses is not without limitations. First, the 27 reviews are heterogeneous publications and thus do not necessarily permit comparison across their findings (e.g., Borenstein et al. 2009). This became most evident when the aspect of duration of interventions was coded from the reviews, but it was not possible to reach a standardized measure that could be adequately interpreted across the 27 reviews. Another limitation is the reviews' lack of statistical information, resulting, for example, in a subsample of only 14 reviews out of 27 that could be used for meta-meta-analysis. Besides, since there were too few studies with skills accounted for in the reviews, it was not possible to report specific effects on teacher skills apart from teacher knowledge in the meta-analyses. Future research would benefit from more meta-analyses, since, for example, this would allow the analysis of both knowledge and skills as well as moderator effects on increased outcomes in meta-meta-analyses. To this end, we were obliged to rely on a descriptive synthesis of what the authors had reported in their reviews, which is, of course, a clear limitation in our approach.

Because the 27 reviews relate to intervention studies that were conducted in various countries, it was not possible for us to account for country-specific analyses or draw specific conclusions for an individual teacher education system (e.g., in Germany). This is another limitation of our approach. Future research might be interested to analyze intervention studies on such a country-specific level or to compare groups of intervention studies conducted in different countries.

As the number of intervention studies in teacher education has increased remarkably in recent decades, we anticipate the publication of further literature reviews synthesizing these studies in the near future. Meta-meta-analyses seem to be an appropriate means of informing scholars, educational policy, and practitioners in the interest of remaining up to date (Hattie 2023). Moreover, the effort invested in intervention studies should not be limited to single publications of these studies. Rather, they should contribute to the development of evidence-based teacher education. Literature reviews, meta-analyses, and meta-meta-analyses thus appear appropriate for supporting the evolution of teacher education effectiveness research being toward a higher level of educational science.

9 Recommendations for future teacher education interventions

Literature reviews do not only serve to show what is already known about a phenomenon, but they also serve for “making claims (...) about what new research we need to undertake to address questions that are unanswered” (Newman and Gough 2020, p. 4). Therefore, to outline what we can learn from the findings of the present synthesis and meta-meta-analysis, five recommendations for teacher interventions to be carried out in the future are provided in the following.

1. Concluding from the findings related to RQ1, it seems reasonable that future reviews of intervention studies, but also intervention studies themselves, should be encouraged to *contribute to theory development* in the field more strongly rather than reporting on concrete empirical findings only. It seems to be crucial examining whether there is a solid theoretical basis that needs to be expanded and confirmed or whether new theories may even need to be generated (e.g., Friedrichs 1990).
2. Regarding findings on input variables (RQ2), it seems to be striking that too few interventions prioritize combining training and practice, although this is one of the most demonstrably effective design elements (e.g. Desimone 2009; Lipowsky and Rzejak 2015). Future studies and reviews could start here to find out why this has been neglected (e.g., for pragmatic reasons) and develop strategies to *facilitate the link between theory and practice*.
3. Regarding output variables (RQ2), it became evident that heterogeneous approaches resulted in inconsistent operationalization, and the components proposed by Desimone (2009) have not been assessed comprehensively yet. This impeded comparability of findings, also when conducting meta-analyses. It seems that *developing guidelines for conducting interventions in teacher education* could be helpful to solve this problem of lack of comparability.

4. Moreover, there is the need for *improving the quality of assessing the dependent variables*. The development and use of new, more objective test instruments could improve the quality of intervention research. In any case, significantly greater efforts must be made to develop objective and standardized measurement methods for different learning objectives that are also conceptually compatible with promising theoretical models (e.g., Desimone 2009). In the context of concrete interventions in the classroom, particular effort should be made to achieve the best possible constructive alignment between learning goals, intervention activities, and assessment formats.
5. Concluding from the findings related to RQ3, it is crucial *getting access to the field* for conducting empirical educational research at high quality. As findings related to RQ3 and RQ4 showed, true experimental designs with randomized grouping and multiple treatments appear to be less common in intervention studies, but they are important in increasing the quality of studies to be included in meta-meta-analyses. A fundamental problem in effectiveness research is implementing RCTs in the context of existing curricula (König 2013). This makes a randomized allocation to the intervention and control groups rarely possible and thus also negates the great advantage of real experiments, which is that all potential confounding variables (known and unknown) would be distributed comparably across the groups (given a sufficiently large sample size) and thus maximize the internal validity of the study. Before carrying out interventions, it thus should be clarified whether random assignment of participants to the groups is possible from a legal, administrative, and educational perspective. Regarding intervention studies, it is important that stakeholders and institutions show commitment in permitting and supporting research efforts and encourage pre-service and in-service teachers as well as their students to participate in intervention studies and specific measures of evaluating intervention studies. This will increase the overall quality of research designs, data, analyses, and findings.

Supplementary Information The online version of this article (<https://doi.org/10.1007/s11618-025-01319-x>) contains supplementary material, which is available to authorized users.

Funding This work was supported by the Excellent Research Support Program, University of Cologne, Germany, financed with funds from the university allowance (DFG).

Funding Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest J. König, S. Heine, J. Weyers, C.-G. Arntz, C. Kramer, S. Strauß, K. Gerhard, J. Mühle, T. Engartner, J. Großschedl, C. Hanisch, P. Hanke, T. Hennemann, J. Jost, K. Kaspar and B. Rott declare that they have no competing interests.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden. Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen. Weitere

Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

References¹

- *Alexander, J. L., Ayres, K. M., & Smith, K. A. (2015). Training teachers in evidence-based practice for individuals with autism spectrum disorder: a review of the literature. *Teacher Education and Special Education*, 38(1), 13–27. <https://doi.org/10.1177/0888406414544551>.
- *Allsopp, D. H., & Haley, K. C. (2015). A synthesis of research on teacher education, mathematics, and students with learning disabilities. *Learning Disabilities: A Contemporary Journal*, 13(2), 177–206. <http://www.ldw-ldcj.org>.
- *Ansyari, M. F., Groot, W., & De Witte, K. (2020). Tracking the process of data use professional development interventions for instructional improvement: A systematic literature review. *Educational Research Review*, 31, 100362. <https://doi.org/10.1016/j.edurev.2020.100362>.
- *Ansyari, M. F., Groot, W., & De Witte, K. (2022). A systematic review and meta-analysis of data use professional development interventions. *Journal of Professional Capital and Community*, 7(3), 256–289. <https://doi.org/10.1108/jpcc-09-2021-005>.
- *Basma, B., & Savage, R. (2023). Teacher professional development and student reading in middle and high school: a systematic review and meta-analysis. *Journal of Teacher Education*, 74(3), 214–228. <https://doi.org/10.1177/00224871231153084>.
- *Billingsley, G., Smith, S., Smith, S., & Meritt, J. (2019). A systematic literature review of using Immersive virtual reality technology in teacher education. *Journal of Interactive Learning Research*, 30(1), 65–90. <https://www.learntechlib.org/primary/p/176261/>.
- *Brock, M. E., & Carter, E. W. (2017). A meta-analysis of educator training to improve implementation of interventions for students with disabilities. *Remedial and Special Education*, 38(3), 131–144. <https://doi.org/10.1177/0741932516653477>.
- *Didion, L., Toste, J. R., & Filderman, M. (2020). Teacher professional development and student reading achievement: a meta-analytic review of the effects. *Journal of Research on Educational Effectiveness*, 13(1), 29–66. <https://doi.org/10.1080/19345747.2019.1670884>.
- *Dignath, C., Rimm-Kaufman, S., van Ewijk, R., & Kunter, M. (2022). Teachers' beliefs about inclusive education and insights on what contributes to those beliefs: a meta-analytical study. *Educational Psychology Review*, 34, 2609–2660. <https://doi.org/10.1007/s10648-022-09695-0>.
- *Kalinowski, E., Gronostaj, A., & Vock, M. (2019). Effective professional development for teachers to foster students' academic language proficiency across the curriculum: a systematic review. *AERA Open*, 5(1), 1–23. <https://doi.org/10.1177/2332858419828691>.
- *Khamzina, K., Stanczak, A., Brasselet, C., Desombre, C., Legrain, C., Rossi, S., Guirimand, N., & Cilia, F. (2024). Designing effective pre-service teacher training in inclusive education: a narrative review of the effects of duration and content delivery mode on teachers' attitudes toward inclusive education. *Educational Psychology Review*. <https://doi.org/10.1007/s10648-024-09851-8>.
- *König, J., Heine, S., Jäger-Biela, D., & Rothland, M. (2022). ICT integration in teachers' lesson plans: a scoping review of empirical studies. *European Journal of Teacher Education*. <https://doi.org/10.1080/02619768.2022.2138323>.
- *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>.
- *Kretlow, A., Graves, B., & Christina, C. (2010). Using coaching to improve the fidelity of evidence-based practices: a review of studies. *Teacher Education and Special Education*, 33(4), 279–299. <https://doi.org/10.1177/0888406410371643>.
- *Kurniawati, F., De Boer, A. A., Minnaert, A. E. M. G., & Mangunsong, F. (2014). Characteristics of primary teacher training programmes on inclusion: a literature focus. *Educational Research*, 56(3), 310–326. <https://doi.org/10.1080/00131881.2014.934555>.
- *Lautenbach, F., & Heyder, A. (2019). Changing attitudes to inclusion in preservice teacher education: a systematic review. *Educational Research*, 61(2), 231–253. <https://doi.org/10.1080/02619768.2022.2138323>.

¹ Titles included into the review marked by an asterix.

- *Ma, K., Zhang, J., Chutiyami, M., Liang, L. Y., & Dong, J. J. (2023). Effectiveness of blended teaching in preservice teacher education: a meta-analysis. *Distance Education*, 44(4), 1–22. <https://doi.org/10.1080/01587919.2022.2155617>.
- *McGuire, S. N., Meadan, H., & Folkerts, R. (2024). Classroom and behavior management training needs and perceptions: a systematic review of the literature. *Child Youth Care Forum*, 53(2024), 117–139. <https://doi.org/10.1007/s10566-023-09750-z>.
- *Ning, Y. M., Zhou, Y., Wijaya, T. T., & Chen, J. H. (2022). Teacher education interventions on teacher TPACK: a meta-analysis study. *Sustainability*, 14(18), 11791. <https://doi.org/10.3390/su141811791>.
- *Oliveira, S., Roberto, M. S., Veiga-Simao, A. M., & Marques-Pinto, A. (2021). A meta-analysis of the impact of social and emotional learning interventions on teachers' burnout symptoms. *Educational Psychology Review*, 33(4), 1779–1808. <https://doi.org/10.1007/s10648-021-09612-x>.
- *Smith, T. E., & Sheridan, S. M. (2019). The effects of teacher training on teachers' Family—engagement practices, attitudes, and knowledge: a meta-analysis. *Journal of Education and Psychological Consultation*, 29(2), 128–157. <https://doi.org/10.1080/10474412.2018.1460725>.
- *Whear, R., Thompson-Coon, J., Boddy, K., Ford, T., Racey, D., & Stein, K. (2013). The effect of teacher-led interventions on social and emotional behaviour in primary school children: a systematic review. *British Educational Research Journal*, 39(2), 383–420. <https://doi.org/10.1080/01411926.2011.650680>.
- *Wilson, M. L. (2023). The impact of technology integration courses on preservice teacher attitudes and beliefs: a meta-analysis of teacher education research from 2007–2017. *Journal of Research on Technology in Education*, 55(2), 252–280. <https://doi.org/10.1080/15391523.2021.1950085>.
- *Wilson, M. L., Ritzhaupt, A. D., & Cheng, L. (2020). The impact of teacher education courses for technology integration on pre-service teacher knowledge: a meta-analysis study. *Computers & Education*, 156, 103941. <https://doi.org/10.1016/j.compedu.2020.103941>.
- *Young, J. R., Young, J. L., & Hamilton, C. (2013). The use of confidence intervals as a meta-analytic lens to summarize the effects of teacher education technology courses on preservice teacher TPAC. *Journal of Research on Technology in Education*, 46(2), 149–172. <https://doi.org/10.1080/15391523.2013.10782617>.
- *Zhang, J., Zhou, M. Z., & Xiaotian, Z. (2023). Interventions to promote teachers' perceptions about STEM education: a meta-analysis. *Education and Information Technologies*, 28(6), 7355–7390. <https://doi.org/10.1007/s10639-022-11492-9>.
- *Zhou, X., Shu, L., Xu, Z., & Padrón, Y. (2023). The effect of professional development on in-service STEM teachers' self-efficacy: a meta-analysis of experimental studies. *International Journal of STEM Education*, 10(37). <https://doi.org/10.1186/s40594-023-00422-x>.
- Accurso, K., & Gebhard, M. (2021). SFL praxis in the US teacher education: a critical literature review. *Language and Education*, 35(5), 402–428. <https://doi.org/10.1080/09500782.2020.1781880>.
- Barrio, B. L., Lindo, E. J., Combes, B. H., & Hovey, K. A. (2015). Ten years of response to intervention: implications for general education teacher preparation programs. *Action in Teacher Education*, 37(2), 190–204. <https://doi.org/10.1080/01626620.2015.1004603>.
- Blömeke, S., Felbrich, A., Müller, C., Kaiser, G., & Lehmann, R. (2008). Effectiveness of teacher education: State of research, measurement issues and consequences for future studies. *ZDM*, 40, 719–734. <https://doi.org/10.1007/s11858-008-0096-x>.
- Blömeke, S., Suhl, U., Kaiser, G., & Döhrmann, M. (2012). Family background, entry selectivity and opportunities to learn: what matters in primary teacher education? An international comparison of fifteen countries. *Teaching and Teacher Education*, 28(1), 44–55. <https://doi.org/10.1016/j.tate.2011.08.006>.
- Blömeke, S., Jentsch, A., Ross, N., Kaiser, G., & König, J. (2022). Opening up the black box: Teacher competence, instructional quality, and students' learning progression. *Learning and Instruction*, 79. <https://doi.org/10.1016/j.learninstruc.2022.101600>.
- Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. <https://doi.org/10.1002/9780470743386>.
- Bottiani, J. H., Larson, K. E., Debnam, K. J., Bischoff, C. M., & Bradshaw, C. P. (2017). Promoting educators' use of culturally responsive practices: a systematic review of in-service interventions. *Journal of Teacher Education*, 69(4), 1–19. <https://doi.org/10.1177/0022487117722553>.
- Brevik, L. M., Gudmundsdottir, G. B., Lund, A., & Strømme, T. A. (2019). Transformative agency in teacher education: fostering professional digital competence. *Teaching and Teacher Education*, 86, 102875. <https://doi.org/10.1016/j.tate.2019.07.005>.
- Brock, M. E., Huber, H. B., Carter, E. W., Juarez, A. P., & Warren, Z. E. (2014). Statewide assessment of professional development needs related to educating students with autism spectrum disorder.

- der. *Focus on Autism and Other Developmental Disabilities*, 29, 67–79. <https://doi.org/10.1177/1088357614522290>.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Ravenna Books.
- Cochran-Smith, M. (2001). Constructing outcomes in teacher education: policy, practice, and pitfalls. *Education Policy Analysis Archives*. <https://doi.org/10.14507/epaa.v9n11.2001>.
- Cochran-Smith, M., & Villegas, A. M. (2015). Framing teacher preparation research: an overview of the field, part 1. *Journal of Teacher Education*, 66(1), 7–20. <https://doi.org/10.1177/0022487114549072>.
- Cochran-Smith, M., Cannady, M., McEachern, K. P., Mitchell, K., Piazza, P., Power, C., & Ryan, A. (2012). Teachers' education and outcomes: mapping the research terrain. *Teachers College Record*, 114(10), 1–49. <https://doi.org/10.1177/016146811211401006>.
- Coe, R., Waring, M., Hedges, L. V., & Ashley, L. D. (Eds.). (2021). *Research methods and methodologies in education* (3rd edn.). SAGE.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn.). Laurence Erlbaum.
- Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., & Yang, S. (2024). Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science*, 39(1), 165–191. <https://doi.org/10.48550/arXiv.2011.08047>.
- Cook, B. G., & Cook, S. C. (2013). Unraveling evidence-based practices in special education. *The Journal of Special Education*, 47, 71–82. <https://doi.org/10.1177/0022466911420877>.
- Cramer, C., König, J., Rothland, M., & Blömeke, S. (Eds.). (2020). *Handbuch Lehrerinnen- und Lehrerbildung*. UTB Klinkhardt. <https://doi.org/10.36198/9783838554730>.
- Cramer, C., Brown, C., & Aldridge, D. (2023). Meta-reflexivity and teacher professionalism: facilitating multiparadigmatic teacher education to achieve a future-proof profession. *Journal of Teacher Education*. <https://doi.org/10.1177/00224871231162295>.
- Cummings, R., Harlow, S., & Maddux, C. D. (2007). Moral reasoning of in-service and pre-service teachers: a review of the research. *Journal of Moral Education*, 36(1), 67–78. <https://doi.org/10.1080/03057240601185471>.
- Darling-Hammond, L., & Bransford, J. (Eds.). (2007). *Preparing teachers for a changing world: what teachers should learn and be able to do*. Wiley.
- 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>.
- Dunst, C. J., Hamby, D. W., Howse, R. B., Wilkie, H., & Annas, K. (2020). Research synthesis of meta-analyses of preservice teacher preparation practices in higher education. *Higher Education Studies*, 10(1), 29–47. <https://doi.org/10.5539/hes.v10n1p29>.
- Ellis, V., Correia, C., Turvey, K., Childs, A., Andon, N., Harrison, C., Jones, J., & Hayati, N. (2023). Re-definition/redirection and incremental change: a systematic review of innovation in teacher education research. *Teaching and Teacher Education*. <https://doi.org/10.1016/j.tate.2022.103918>.
- ERIC (2024). Education Resources Information Center: Search the Thesaurus. <http://www.eric.ed.gov>.
- Fernández-Batanero, J. M., Montenegro-Rueda, M., Fernández-Cerero, J., & García-Martínez, I. (2022). Digital competences for teacher professional development. Systematic review. *European Journal of Teacher Education*, 45(4), 513–531. <https://doi.org/10.1080/02619768.2020.1827389>.
- Friedrichs, J. (1990). *Methoden empirischer Sozialforschung*. Springer.
- Galluzzo, G. R., & Craig, J. R. (1990). Evaluation of preservice teacher education programs. In W. R. Houston (Ed.), *Handbook of research on teacher education* (pp. 599–616). Macmillan.
- Gerhard, K., Jäger-Biela, D., & König, J. (2023). Opportunities to learn, technological pedagogical knowledge, and personal factors of pre-service teachers: Understanding the link between teacher education program characteristics and student teacher learning outcomes in times of digitalization. *Zeitschrift für Erziehungswissenschaft*, 26(3), 653–676. <https://doi.org/10.1007/s11618-023-01162-y>.
- Hanushek, E. A., & Jackson, J. E. (2013). *Statistical methods for social scientists*. Academic Press.
- Hascher, T. (2014). Forschung zur Wirksamkeit der Lehrerbildung. In E. Terhart, H. Bennewitz & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (2nd edn., pp. 542–571). Waxmann.
- Hattie, J. (2009). *Visible Learning: a synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Hattie, J. (2012). *Visible learning for teachers: maximizing impact on learning*. Routledge.
- Hattie, J. (2023). Visible learning: the sequel: a synthesis of over 2,100 meta-analyses relating to achievement. Routledge. <https://doi.org/10.4324/9781003380542>.
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327(7414), 557–560. <https://doi.org/10.1136/bmj.327.7414.557>.

- Hikida, M., Chamberlain, K., Tily, S., Daly-Lesch, A., Warner, J. R., & Schallert, D. L. (2019). Reviewing how preservice teachers are prepared to teach reading processes: what the literature suggests and over-looks. *Journal of Literacy Research*, 51(2), 177–195. <https://doi.org/10.1177/1086296X19833297>.
- Hill, H., Mancenido, Z., & Loeb, S. (2021). *Effectiveness research for teacher education* (EdWorkingPaper, pp. 21–252). <https://doi.org/10.26300/zhhb-j781>.
- Kaiser, G. (2024). The TEDS research program. Its origin and development (in memoriam Sigrid Blömeke). In J. König, C. Hanisch, P. Hanke, T. Hennemann, K. Kaspar, M. Martens & S. Strauß (Eds.), *Auf die Lehrperson und ihren Unterricht kommt es an. Zehn Jahre empirische Professions- und Unterrichtsforschung im IZef der Universität zu Köln* (pp. 261–287). Waxmann. <https://doi.org/10.31244/9783830998761>.
- Kaiser, G., & König, J. (2019). Competence measurement in (mathematics) teacher education and beyond: implications for policy. *Higher Education Policy*, 32, 597–615. <https://doi.org/10.1057/s41307-019-00139-z>.
- Kauper, T., Bernholt, A., Möller, J., & Köller, O. (Eds.). (2023). *PaLea: Professionelle Kompetenzen und Studienstrukturen im Lehramtsstudium*. Waxmann.
- Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*, 86(4), 945–980. <https://doi.org/10.3102/0034654315626800>.
- König, J. (Ed.). (2012). *Teachers' pedagogical beliefs. Definition and operationalisation—connections to knowledge and performance—development and change*. Waxmann. <https://doi.org/10.25656/01:21030>.
- König, J. (2013). Standpunkt: Methoden und Designs der empirischen Forschung im edukativen Bereich – Bestandsaufnahme und Barrieren für experimentelle Forschung. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 107(1), 19–22. <https://doi.org/10.1016/j.zefq.2012.12.020>.
- König, J., & Blömeke, S. (2020). Wirksamkeits-Ansatz in der Lehrerinnen- und Lehrerbildung. In C. Cramer, J. König, M. Rothland & S. Blömeke (Eds.), *Handbuch Lehrerinnen- und Lehrerbildung* (pp. 172–178). Klinkhardt UTB. <https://doi.org/10.36198/9783838554730>.
- König, J., Rothland, M., & Schaper, N. (Eds.). (2018). *Learning to Practice, Learning to Reflect? Ergebnisse aus der Längsschnittstudie LtP zur Nutzung und Wirkung des Praxissemesters in der Lehrerbildung*. Springer VS. <https://doi.org/10.1007/978-3-658-19536-6>.
- König, J., Heine, S., Kramer, C., Weyers, J., Becker-Mrotzek, M., Großschedl, J., Hanisch, C., Hanke, P., Hennemann, T., Jost, J., Kaspar, K., Rott, B., & Strauß, S. (2024). Teacher education effectiveness as an emerging research paradigm: a synthesis of reviews of empirical studies published over three decades (1993–2023). *Journal of Curriculum Studies*, 56(4), 371–391. <https://doi.org/10.1080/00220272.2023.2268702>.
- Leutner, D. (2013). Perspektiven pädagogischer Interventionsforschung. In E. Severing & R. Weiß (Eds.), *Qualitätsentwicklung in der Berufsbildungsforschung* (pp. 17–28). Bertelsmann. <https://doi.org/10.25656/01:7983>.
- Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review*, 22, 215–243. <https://doi.org/10.1007/s10648-010-9125-8>.
- Lieberman, A., & Pointer Mace, D. H. (2008). Teacher learning: the key to educational reform. *Journal of Teacher Education*, 59(3), 226–234. <https://doi.org/10.1177/0022487108317020>.
- Lipowsky, F. (2006). Auf den Lehrer kommt es an. Empirische Evidenzen für Zusammenhänge zwischen Lehrerkompetenzen, Lehrerhandeln und dem Lernen der Schüler. *Zeitschrift für Pädagogik*. <https://doi.org/10.25656/01:7370>.
- Lipowsky, F., & Rzejak, D. (2015). Key features of effective professional development programmes for teachers. *Ricercazione. Six-monthly Journal on Learning, Research and Innovation in Education*, 27–51.
- Little, J. W. (1993). Teachers' professional development in a climate of educational reform. *Educational evaluation and policy analysis*, 15(2), 129–151. <https://doi.org/10.3102/01623737015002129>.
- Mahler, D., Bock, D., Schaubert, S., & Harms, U. (2024). Using longitudinal models to describe preservice science teachers' development of content knowledge and pedagogical content knowledge. *Teaching and Teacher Education*, 144, 104583. <https://doi.org/10.1016/j.tate.2024.104583>.
- McBride, N. (2016). *Intervention research: A practical guide for developing evidence-based school prevention programmes*. Springer. <https://doi.org/10.1007/978-981-10-1011-8>.
- Mingebach, T., Kamp-Becker, I., Christiansen, H., & Weber, L. (2018). Meta-meta-analysis on the effectiveness of parent-based interventions for the treatment of child externalizing behavior problems. *PloS one*, 13(9), e202855. <https://doi.org/10.1371/journal.pone.0202855>.

- Mishra, P., & Koehler, M.J. (2006). Technological pedagogical content knowledge: a framework for teacher knowledge. *Teachers College Record*, 108(6), 1017–1054. <https://doi.org/10.1111/j.1467-9620.2006.00684.x>.
- Munder, T., Brüttsch, O., Leonhart, R., Gerger, H., & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: an overview of reviews. *Clinical psychology review*, 33(4), 501–511. <https://doi.org/10.1016/j.cpr.2013.02.002>.
- Newman, M., & Gough, D. (2020). Systematic reviews in educational research: Methodology, perspectives and application. *Systematic reviews in educational research: Methodology, perspectives and application*, 3–22. https://doi.org/10.1007/978-3-658-27602-7_1.
- Österling, L., & Christiansen, I. (2022). Whom do they become? A systematic review of research on the impact of practicum on student teachers' affect, beliefs, and identities. *International Electronic Journal of Mathematics Education*, 17(4), em710. <https://doi.org/10.29333/iejme/12380>.
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hrobjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *PLOS Medicine*, 18(3), e1003583. <https://doi.org/10.1371/journal.pmed.1003583>.
- Parsons A. W., Parsons S. A., Morewood A., & Ankrum J. W. (2016). Barriers to change: Findings from three literacy professional learning initiatives. *Literacy Research and Instruction*, 55(4), 331–352. <https://doi.org/10.1080/19388071.2016.1193575>.
- Pisaneschi, J. (2020). How do ITE programmes support pre-service primary teachers to become competent and confident teachers of mathematics? A systematic review of interventions trialled in Scotland and the Republic of Ireland. *Scottish Educational Review*, 52(1), 73–95. <https://doi.org/10.1163/27730840-05201006>.
- Rouse, M. (2017). A role for teachers and teacher education in developing inclusive practice. In M. Etherington (Ed.), *What teachers need to know: topics in diversity and inclusion* (pp. 19–35). Wipf & Stock.
- Rowan, L., Mayer, D., Kline, J., Kostogriz, A., & Walker-Gibbs, B. (2015). Investigating the effectiveness of teacher education for early career teachers in diverse settings: the longitudinal research we have to have. *The Australian Educational Researcher*, 42(3), 273–298. <https://doi.org/10.1007/s13384-014-0163-y>.
- Scheerens, J., & Blömeke, S. (2016). Integrating teacher education effectiveness research into educational effectiveness models. *Educational Research Review*, 18, 70–87. <https://doi.org/10.1016/j.edurev.2016.03.002>.
- Schleicher, A. (2012). *Preparing teachers and developing school leaders for the 21st century: lessons from around the world*. OECD Publishing.
- Schleicher, A. (2016). *Teaching excellence through professional learning and policy reform*. OECD Publishing. <https://doi.org/10.1787/9789264252059-en>.
- Schmidt, F.L., Oh, I.S., & Hayes, T.L. (2009). Fixed- versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, 62(1), 97–128. <https://doi.org/10.1348/000711007X255327>.
- Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Slavin, R.E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational researcher*, 31(7), 15–21. <https://doi.org/10.3102/0013189X031007015>.
- Sperling, K., Stenberg, C.J., McGrath, C., Åkerfeldt, A., Heintz, F., & Stenliden, L. (2024). In search of artificial intelligence (AI) literacy in teacher education: a scoping review. *Computers and Education Open*. <https://doi.org/10.1016/j.caeo.2024.100169>.
- Starkey, L. (2020). A review of research exploring teacher preparation for the digital age. *Cambridge Journal of Education*, 50(1), 37–56. <https://doi.org/10.1080/0305764X.2019.1625867>.
- Sung, Y.T., Yang, J.M., & Lee, H.Y. (2017). The effects of mobile-computer-supported collaborative learning: meta-analysis and critical synthesis. *Review of Educational Research*, 87(4), 768–805. <https://doi.org/10.3102/0034654317704307>.
- Tatto, M.T. (2021). Professionalism in teaching and the role of teacher education. *European Journal of Teacher Education*, 44(1), 20–44. <https://doi.org/10.1080/02619768.2020.1849130>.
- Tatto, M.T., & Senk, S. (2011). The mathematics education of future primary and secondary teachers: methods and findings from the teacher education and development study in mathematics. *Journal of Teacher Education*, 62(2), 121–137. <https://doi.org/10.1177/0022487110391807>.

- Tatto, M. T., Ingvarson, L., Schwille, J., Peck, R., Senk, S. L., & Rowley, G. (2008). *Teacher education and development study in mathematics (TEDS-M): policy, practice, and readiness to teach primary and secondary mathematics—conceptual framework*. IEA.
- Thomas, C. N., & Rieth, H. J. (2011). A research synthesis of the literature on multimedia anchored instruction in preservice teacher education. *Journal of Special Education Technology*, 26(2), 1–22. <https://doi.org/10.1177/016264341102600201>.
- Tristani, L., & Bassett-Gunter, R. (2020). Making the grade: teacher training for inclusive education: a systematic review. *Journal of Research in Special Educational Needs*, 20(3), 246–264. <https://doi.org/10.1111/1471-3802.12483>.
- Türk, S., Korfmacher, A.-K., Gerger, H., van der Oord, S., & Christiansen, H. (2023). Interventions for ADHD in childhood and adolescence: a systematic umbrella review and meta-meta-analysis. *Clinical Psychology Review*, 102, 102271. <https://doi.org/10.1016/j.cpr.2023.102271>.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of statistical software*, 36, 1–48.
- Wang, J., Odell, S. J., Klecka, C. L., Spalding, E., & Lin, E. (2010). Understanding teacher education reform. *Journal of Teacher Education*, 61(5), 395–402. <https://doi.org/10.1177/0022487110384219>.
- Weyers, J., Ligtvoet, R., & 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. *Journal of Curriculum Studies*, 56(4), 448–467. <https://doi.org/10.1080/00220272.2024.2355923>.
- Yang, X., & Kaiser, G. (2022). The impact of mathematics teachers' professional competence on instructional quality and students' mathematics learning outcomes. *Current Opinion in Behavioral Sciences*, 48, 101225. <https://doi.org/10.1016/j.cobeha.2022.101225>.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Toepper, M., & Lautenbach, C. (2020). *Student learning in German higher education. Innovative measurement approaches and research results*. Springer. <https://doi.org/10.1007/978-3-658-27886-1>.

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

Affiliations

✉ Prof. Dr. Johannes König

Empirical School Research, University of Cologne, Gronewaldstr. 2, 50931 Cologne, Germany

E-Mail: johannes.koenig@uni-koeln.de

Sandra Heine · Dr. Jonas Weyers · Carl-Gereon Arntz · Dr. Charlotte Kramer · Dr. Sarah Strauß · Kristina Gerhard · Janine Mühle · Prof. Dr. Tim Engartner · Prof. Dr. Jörg Großschedl · Prof. Dr. Charlotte Hanisch · Prof. Dr. Petra Hanke · Prof. Dr. Thomas Hennemann · Prof. Dr. Jörg Jost · Prof. Dr. Dr. Kai Kaspar · Prof. Dr. Benjamin Rott

University of Cologne, Cologne, Germany

Sandra Heine

E-Mail: s.heine@uni-koeln.de

Dr. Jonas Weyers

E-Mail: jonas.weyers@uni-koeln.de

Carl-Gereon Arntz

E-Mail: carl-gereon.arntz@uni-koeln.de

Dr. Charlotte Kramer

E-Mail: Charlotte.Kramer@uni-koeln.de

Dr. Sarah Strauß

E-Mail: sarah.strauss@uni-koeln.de

Kristina Gerhard

E-Mail: kristina.gerhard@uni-koeln.de

Janine Mühle

E-Mail: janine.muehle@uni-koeln.de

Prof. Dr. Tim Engartner

E-Mail: tim.engartner@uni-koeln.de

Prof. Dr. Jörg Großschedl

E-Mail: j.grossschedl@uni-koeln.de

Prof. Dr. Charlotte Hanisch

E-Mail: charlotte.hanisch@uni-koeln.de

Prof. Dr. Petra Hanke

E-Mail: petra.hanke@uni-koeln.de

Prof. Dr. Thomas Hennemann

E-Mail: thennema@uni-koeln.de

Prof. Dr. Jörg Jost

E-Mail: joerg.jost@uni-koeln.de

Prof. Dr. Dr. Kai Kaspar

E-Mail: kkaspar@uni-koeln.de

Prof. Dr. Benjamin Rott

E-Mail: brott@uni-koeln.de