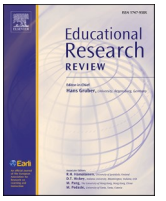
Educational Research Review 49 (2025) 100742

Contents lists available at ScienceDirect

Educational Research Review

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



Review

(When) do teacher professional development interventions improve student Achievement? A meta-analysis of 128 high-quality studies

Adrie J. Visscher [a,*], Natasha Dmoshinskaia [a], Marta Pellegrini [b], Anny Rey-Naizaque [a]

[a] Teacher Development Section, Faculty of Behavioral, Management and Social Sciences (BMS), University of Twente, the Netherlands
[b] Department of Pedagogy, Psychology, Philosophy, University of Cagliari, Italy

ARTICLE INFO

Keywords:
Meta-analysis
Teacher professional development
In-service teachers
Effectiveness
Student achievement

ABSTRACT

Enormous sums of money are spent worldwide on teacher professional development (TPD), seeking to optimize the quality of in-service teachers' teaching. An important question is whether this pays off in terms of improved student achievement, which we investigated in this meta-analysis of 128 (quasi-)experimental, high-quality studies with 356 effect sizes. Results showed an average effect size of 0.09, 95 % CI [0.07, 0.11], considered a medium-size effect on student achievement (Kraft, 2023). Effects varied greatly, with the 95 % prediction interval ranging from −0.15 to 0.32. We used confirmatory and exploratory multiple meta-regression models to examine the effects of potential moderators. The confirmatory model revealed that grade level and publication status were predictors of the effect, while TPD goal did not explain differences in effects on student achievement. The exploratory model seemed to indicate that the number of learning-theory principles applied in designing the TPD (out of these four: teacher performance standards, teacher self-regulation, teacher coaching and teacher cooperation) positively influenced student outcomes. Interestingly, the number of TPD hours and the type of trainer were not statistically significant predictors of TPD effectiveness. We reflect upon the findings and the state-of-the-art of TPD research and formulate recommendations for its further development.

1. Introduction

Teaching encompasses a set of complex professional skills, and pre-service teacher professional development (TPD) can only bring teachers to a starting point as professionals, from which they can further learn and develop professionally. Van de Grift (2010) showed that on average, teachers grow a great deal during the first 5–10 years of their career, and then arrive at a performance plateau (with respect to their pedagogical-didactic skills, and the impact they have on student achievement). They stay at the plateau for many years, until their skills and impact on student learning start to decrease after 20–25 years (Kini & Podolsky, 2016; Leigh, 2010; Visscher, 2017). Not all teachers develop to the same extent with respect to the quality of their teaching and their effectiveness (i.e., their impact on student achievement; Hamre & Pianta, 2005; Wiliam, 2016). About 10–15 % of the variance in student achievement is explained by differences between teachers, for example, differences in the quality of their teaching (Hanushek & Rivkin, 2006; Meelissen & Luyten,

* Corresponding author. University of Twente, PO Box 217, 7500 AE, Enschede, the Netherlands.
E-mail address: a.j.visscher@utwente.nl (A.J. Visscher).

https://doi.org/10.1016/j.edurev.2025.100742
Received 18 December 2023; Received in revised form 22 October 2025; Accepted 3 November 2025
Available online 4 November 2025

2011; Nye et al., 2004). TPD activities for in-service teachers might be a way to improve teaching quality, among other things by supporting teachers in acquiring complex professional teaching competencies that are very hard to learn on your own as a teacher, such as differentiating instruction in the classroom (van Geel et al., 2019).

Enormous sums of money are invested worldwide in in-service TPD interventions, which we define as interventions targeting in-service teachers, aimed at further developing their professional knowledge, skills or attitudes (or a combination of these three) alone, or in conjunction with the introduction of a curriculum or a (digital) tool. For example, in the U.S. this amount was about $18, 000, 000,000 per year in 2014 (Bill & Melinda Gates Foundation, 2014), and now it is probably more. An important question is to what extent all these TPD efforts and resources invested are worthwhile, which then leads to the question, what is the goal of TPD?

Desimone (2009) presented a well-known theory of action for TPD interventions (Fig. 1). Ideally, teachers develop their professional competencies (knowledge, skills, and attitudes, among other things) as a result of participating and learning in TPD interventions, and apply those competencies in the classroom, where students, as a result, become more engaged in the teaching–learning process and learn more (a block that is missing between the two right-hand blocks in Fig. 1), than before the TPD intervention. Students thereafter are expected to perform better on a test measuring their learning achievement. In this study, we answer the question to what extent and under what conditions in-service TPD indeed leads to better student achievement.

1.1. Previous reviews of TPD

We are definitely not the first researchers to find the effectiveness of in-service TPD worth investigating. However, the very influential Desimone (2009) and Timperley et al. (2007) reviews of TPD research were both conducted at a time when very few randomized studies of in-service TPD were available, meaning that the researchers did not have much unbiased information at their disposal about the impact of TPD interventions on student learning. They could not use rigorous study inclusion criteria, as that would have meant that very few studies could have been included (e.g., Yoon et al., 2007).

Table 1, which presents a list of other reviews of TPD research conducted in the last 25 years, shows that the average number of experimental studies included in the reviews of TPD research has increased considerably. Randomized controlled trials (Fletcher-Wood & Zuccollo, 2020; Sims et al., 2023) form the best basis for causal claims regarding TPD effectiveness. It can also be observed that the effect sizes found for TPD effects on student achievement have decreased over the years. The inclusion of many more high-quality experimental studies has probably played a role here.

These prior reviews differ in other respects as well. For example, some were more narrative and did not use meta-analytic methods to combine effect sizes (Darling-Hammond et al., 2017; Kennedy, 2016; van Veen et al., 2012); some included only studies with standardized tests to measure the student achievement effects (Kraft et al., 2018), although most reviews also included researcher-made tests; in some reviews, the control groups in the included studies were not always business-as-usual control groups (Kennedy, 2016); and some reviews focused on only one specific kind of TPD (coaching teachers, Kraft et al., 2018), although most TPD reviews allowed for a large variety of types of TPD.

The differences between the reviews mentioned here led to different effect size estimates of the impact of TPD on student achievement. Study inclusion criteria determine the results that will be obtained. This is evident in the decomposition by Neitzel, Zhang et al. (2022) of the overall results of a meta-analysis by Kulik and Fletcher (2016) of the research into the effectiveness of intelligent tutoring systems. Kulik and Fletcher (2016) found an overall effect of 0.65 across 55 included studies. Neitzel, Lake et al. (2022) found that after excluding the small (fewer than 60 students) or brief (shorter than 12 weeks) studies, the average effect size decreased to 0.39 and 0.40, respectively, while after excluding studies with researcher-made tests for measuring student achievement, the effect size dropped to 0.10.

1.2. Distinguishing features of this meta-analysis

Our meta-analysis differs in several ways from previous TPD reviews. First, our goal was to *minimize methodological bias*. In most previous reviews, study inclusion criteria were not very strict. We wanted to minimize the probability that the effects of TPD interventions were attributable to methodological characteristics of the studies included, instead of to the TPD interventions in those studies. We therefore included only RCTs and quasi-experimental designs (QEDs) with business-as-usual control groups, and excluded studies with researcher-made tests of student achievement. We tried to reduce study bias further by only including studies with small pretest and teacher attrition differences between experimental and control groups.

Second, the review reported here covers a *broader period of research* into TPD than previous reviews of TPD studies; for example, this review covers 31 years (1990–2021) versus 18 years (2002–2020) in the Sims et al. (2023) review.

Third, we wanted to maximize the *practical relevance* of our findings, and therefore excluded studies of classroom interventions implemented in educational practice by others than teachers, interventions involving fewer than 20 teachers per condition, and interventions shorter than 12 weeks long, because such studies do not tell us much about what the TPD intervention effects on student learning will be if the classroom interventions are implemented in educational practice at scale.

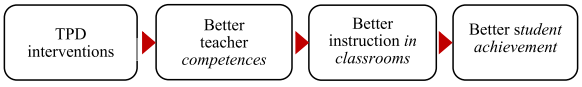


Fig. 1. Theory of action for TPD interventions (Desimone, 2009).

Table 1
Number of studies and experimental studies included and findings of previous reviews of in-service TPD.

Reviews in chronological order	Number of studies	Number of experimental studies	Average effect size
Kennedy (1998)	12	6	0.68; 0.26; 0.57
Yoon et al. (2007)	9	6	0.54
Timperley et al. (2007)	72	9	0.60
van Veen et al. (2012)	34	3	–
Kennedy (2016)	32	22	–
Darling-Hammond et al. (2017)	35	20	–
Basma & Savage (2018)	17	17	0.22
Egert et al. (2018)	9	9	0.14
Kraft et al. (2018)	60	56	0.18
Lynch et al. (2019)	95	86	0.21
Didion et al. (2020)	28	21	0.18
Fletcher-Wood & Zuccollo (2020)	42	42	0.09
Sims et al. (2023)	104	104	0.05

The information in publications about the characteristics of the TPD is often (very) limited and confusing. We therefore used not only the data from *publications* about studies of TPD interventions, but also obtained data about included studies from *interviews* with those who knew what happened in the TPD interventions. In the 115 interviews (covering 90 % of the publications reporting the included studies) that we conducted with researchers/TPD facilitators around the world, we verified the information that we had found in the publication(s), and asked about TPD features that we could not find information about in the publications. Although it was a challenge to find the right people and to obtain their cooperation for the interviews, it led to more valid and complete information about the interventions.

TPD reviews so far have mainly focused on how effective TPD is *in general*, that is, on average across TPD interventions that differ in TPD goals (e.g., Sims et al., 2023; Timperley et al., 2007; Yoon et al., 2007). It might, however, well be that attempts to improve specific aspects of teaching by means of TPD are more effective than others, because those aspects are easier to improve and/or have more impact on student learning. Kennedy (2016) and Pellegrini et al. (2021) distinguished between TPD goals and indeed found differential effects. In this review, we therefore distinguished between *categories of TPD goals*, and investigated how similar or different they are in terms of their effects on student achievement. Below we will elaborate on the TPD goals we investigated.

Finally, in our meta-analysis we investigated TPD interventions from the *perspective of learning theory* (e.g., Schunk, 2019; van Merriënboer & Kirschner, 2018). TPD is about teacher learning, but learning theory has only inspired TPD reviews to a very limited degree. We tried to fill this gap by selecting potential moderators not only from the usual perspective of study characteristics (e.g., sample size, unit of assignment), but also based on what is known about when complex professional tasks are learned most effectively (e.g., teacher scaffolding in acquiring complex skills, by means of classroom coaching or by stimulating teachers to self-regulate their learning processes). We elaborate on this below.

1.2.1. TPD goals distinguished

We distinguished between ten types of TPD goals in this review. This is based on various strands of literature. In the TPD literature, improving teachers' subject-matter content knowledge is mentioned as one of the characteristics of effective TPD (e.g., Darling-Hammond et al., 2017; Desimone, 2009). Not surprisingly the goal of many TPD interventions is to expand teachers' *knowledge of the subject matter* they teach, and/or to improve their *pedagogical content knowledge* (PCK; Gess-Newsome, 1999; Shulman, 1986), that is, their knowledge of how to teach specific subject-matter content, and their knowledge of the misconceptions learners may have when learning specific subject matter, along with how to prevent and remedy those misconceptions. So, the first TPD goal we distinguished was *TPD for (P)CK*.

Research (e.g., Danielson, 2013; Grossman et al., 2013; Pianta et al., 2008; van de Grift, 2007, 2010) has pointed to characteristics of effective teaching that can also be targeted in TPD: learning as teachers how to provide a *safe classroom climate* for students, *organize and manage the classroom productively, differentiate instruction* in line with student needs, *activate student engagement*, and *develop students' metacognition* have been found to be important aspects of effective teaching. In this review, improvement in each of these aspects of teaching could also be a TPD goal (5 additional types of TPD goals).

TPD for *comprehensive approaches* is another category of TPD goal we distinguished. In that case teachers learn how to implement approaches that are comprehensive in one of two possible senses. One possibility is that teachers learn how to implement a set of intervention components (e.g., Success for All, which includes cooperative learning, student performance monitoring, student tutoring and cooperation with parents and the school community; Borman et al., 2007). In the other type of comprehensive approach, teachers learn how to use general, instead of subject-specific instructional approaches (e.g., dialogic teaching, which improves the quality of classroom talk in general in order to increase pupils' engagement, learning, and attainment; Alexander, 2015).

TPD for *learning how to teach non-traditional content* was another potential TPD goal we distinguished, as in some TPD interventions teachers learn how to teach content that is not usually part of the primary and/or secondary school curricula, for example, teaching a growth mindset (Rienzo et al., 2015) or teaching philosophy to primary school students (Gorard et al., 2015).

The last two TPD goal categories we included in our review were TPD for the *implementation of a curriculum* (Wolf et al., 2018), and TPD for the *implementation of a (digital) tool* (e.g., Savage et al., 2013). Digital tools can, for example, provide teachers with real-time feedback on the progress of their students, on the basis of which teachers can provide support to students where needed. Teachers

participating in TPD programs for these two goals learn about the background of specific curriculum material or a (digital) tool, their content and about how they can use the curricula or tools in their classrooms.

1.3. Potential TPD moderators

Our main goal was to study six potential moderators related to TPD treatments. Four moderators were drawn from learning theory (Schunk, 2019). So far, what is known from cognitive psychology about how professionals learn effectively (cf., van Merriënboer & Kirschner, 2018) has played a limited role in TPD research (see also Kennedy, 2016; Maandag et al., 2017).

Teaching is a very complex professional task, as is improving it. TPD often encompasses learning professional knowledge, attitudes, and skills as a teacher. The transfer of what is learned to its application in professional practice (in our case, the classroom) is not self-evident at all (Blume et al., 2010; van Merriënboer & Kirschner, 2018). One potential reason for the lack of transfer of training is that teachers are not supported well, and must find out much on their own about how to integrate TPD content in their teaching (Cohen & Ball, 2001). Supporting teachers in their learning is crucial for the quality of their learning (van de Pol et al., 2010; van Merriënboer & Kirschner, 2018) and can be done in various ways. We studied the impact of four forms of support (our four moderators from learning theory).

We expected teacher learning to be more effective when what teachers have to learn and to apply in the classroom is completely clear to them. That is not a given, however. Well-specified teacher *performance standards* can help in this respect (e.g., Hambleton et al., 2000; van Merriënboer & Kirschner, 2018): written standards that define what teachers are supposed to learn and to do in their lessons as a result of the TPD. Without such standards, it is hard for teachers to work towards the desired goals.

Performance standards enable teachers to compare their own performance with the standards (monitoring), and to determine what (parts of the) standards they already meet and where there is still room for improvement, and to plan how they will work on further improvement (control; Nelson & Narens, 1990). The combination of performance standards and *teacher self-regulation* (our second principle and moderator from learning theory; e.g., Endedijk et al., 2012; Zimmerman & Schunk, 2011) may be powerful for TPD. Approaching TPD as a passive, receptive form of learning, in which a TPD trainer transmits knowledge to teachers and is the only one who monitors, corrects and optimizes teacher learning, ignores the potential of teachers' active involvement in the learning process. The time available for joint activities of teachers and trainers will also always be limited, so it is important that teachers actively monitor and regulate their learning.

Teacher coaching (in the classroom or online), is the third principle from learning theory and moderator that we investigated. TPD often includes summer institutes or workshops in which TPD trainers transmit knowledge to teachers. The expectation is that teachers will thereafter apply what they have learned in their classrooms. However, integrating newly learned competencies with old ones and changing classroom behavior in the intended way is complex and hard to do as a teacher on your own (e.g., Sims et al., 2023). More and more, teachers are being supported in implementing the TPD content in the classroom, through receiving feedback (face-to-face or online) on how well they are implementing the TPD content, and on where and how they still can improve (e.g., Darling-Hammond et al., 2017; Kraft et al., 2018). Kraft et al. showed that teacher coaching can be very effective.

The final moderator from learning theory that we studied is teacher *cooperation*. Teachers cooperating in an intervention can exchange knowledge and feedback, and in that way learn from each other (e.g., Maandag et al., 2017; Timperley et al., 2007). Cooperation between the teachers at a school (for example, the whole school, or all teachers from a department or a grade) and/or teachers outside their schools (professional learning communities of teachers across the schools involved in an TPD intervention) can also be important for staying motivated, and for persevering when change and improvement prove to be hard (Jansen in de Wal, 2016).

We expected the four principles from learning theory (performance standards, teacher self-regulation, classroom coaching and teacher cooperation) to be especially effective if they occurred *together* in TPD interventions. We therefore deliberately investigated these four principles as *combinations* and not each of them separately, as combinations and interactions of variables have important meanings that are not found when studying individual variables (Cohen et al., 2003; Hiebert et al., 2005).

We investigated two additional moderators. The *number of TPD hours* during which teachers learn and develop in a TPD intervention was our fifth moderator. As the teacher competencies to be learned and applied in TPD programs are often complex, and as learning is not a linear process (people forget, for example; Timperley et al., 2007), it takes time to fully grasp the competencies taught in TPD programs. TPD time has been mentioned in many publications (e.g., Basma & Savage, 2018; Darling-Hammond et al., 2017; Desimone, 2009; Yoon et al., 2007); however, there is no strong evidence of its precise impact on TPD effectiveness. We investigated whether larger numbers of hours spent on TPD interventions enabled teachers to learn and implement the complex competencies better, and thus be more effective than TPD programs with fewer TPD hours.

Finally, we expected that it may also matter for the effects of TPD who the *TPD trainer* (our sixth moderator) was (Cheung & Slavin, 2016; Kennedy, 2016). Is it the developer/researcher behind the intervention who implements the TPD themself, or is it an independent trainer who does not have the strong commitment ("super realization", Slavin & Smith, 2009; Wolf et al., 2020) to the TPD intervention that developers/researchers have? TPD interventions with developers/researchers as trainers might yield results that are not indicative of the effects achieved when a TPD program is implemented at scale.

Based on the previous, we addressed the following research questions:

RQ1. What is the average effect of TPD interventions for in-service teachers on students' academic achievement?
RQ2. What is the average effect per TPD goal on students' academic achievement?
RQ3. What factors moderate the effect of TPD interventions on students' academic achievement?

2. Method

This meta-analysis was carried out following the guidelines for high-quality systematic reviews as described by Page et al. (2021), and Pigott and Polanin (2020).

2.1. Search procedures

We used a systematic and comprehensive search strategy in order to capture all relevant studies published in the period 1990–2021. We started by locating previous reviews and searching the reference lists to identify studies included in previous syntheses. A total of 30 reviews on TPD or educational programs in K-12 that also included TPD studies were screened (Section 1 of the online supplemental materials provides a complete list). In addition, we searched for new primary studies that had not been included in reviews yet. We used a combination of keywords for the database search (see Table S2.1) in ERIC, Education Source, PsycInfo, and SCOPUS. To locate unpublished studies, we also conducted a search on the first 100 hits in Google Scholar, using different combinations of the selected keywords, as well as the following relevant websites of institutions and conferences: What Works Clearinghouse (WWC), Institute of Education Sciences (IES), Best Evidence Encyclopedia (BEE), and Education Endowment Foundation (EEF). In our interviews with the researchers in the included studies, we also asked them for other (un)published studies into in-service TPD conducted by either themselves or others.

2.2. Eligibility criteria

The inclusion criteria were adapted from the What Works Clearinghouse (2022) and recent reviews from the Best Evidence Encyclopedia (Neitzel, Lake et al., 2022; Pellegrini et al., 2021), and were as follows:

- Randomized studies and quasi-experimental designs with treatment and control group matching occurring prior to the intervention.
- Pre-K to grade 12 studies in regular schools, excluding interventions targeting special education students.
- A sample of at least 20 teachers/classes per condition. Small studies tend to affect effect sizes (Cheung & Slavin, 2016). We considered the sample of teachers instead of students, because of the focus of our review.
- Studies on the effects of TPD interventions. If the TPD included practice with or delivery of a program in class, the TPD had to have been implemented by teachers. We thus excluded supplemental approaches delivered by researchers or teaching assistants.

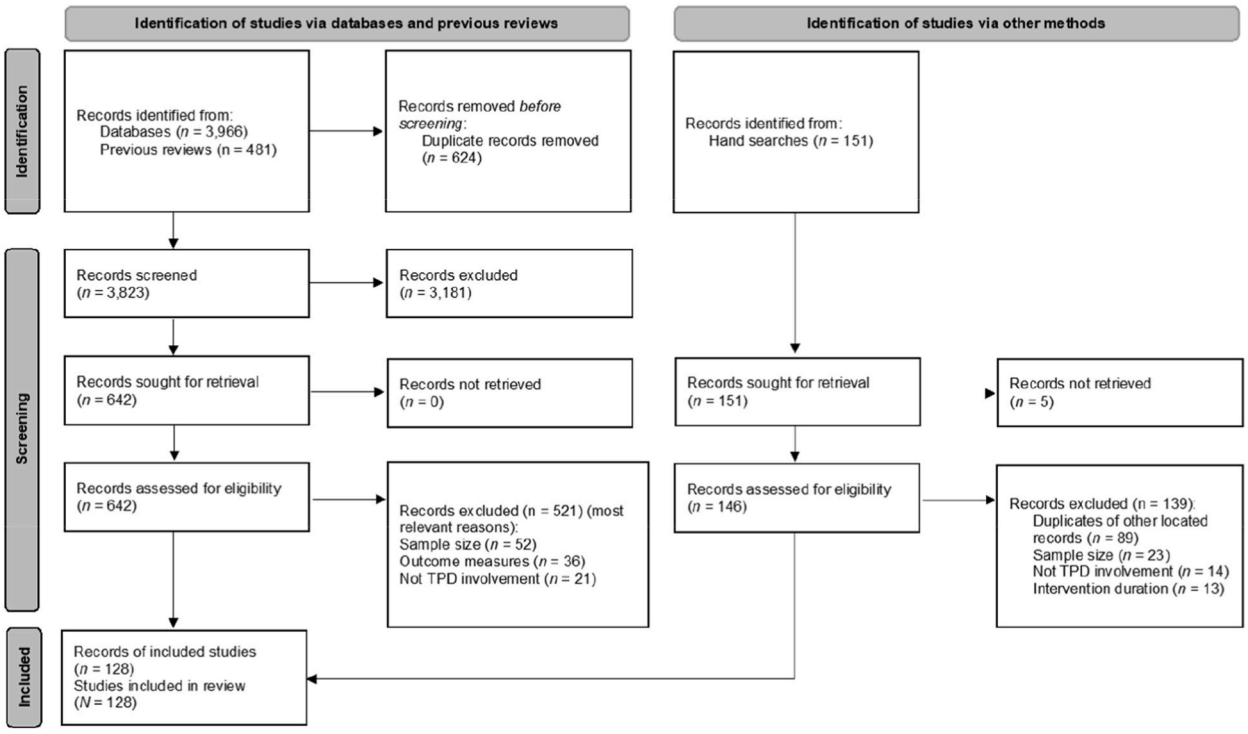


Fig. 2. Prisma selection diagram (adapted from Page et al., 2021) ($N = 128$).

- Standard practice already in place in schools (business-as-usual) for the control group, excluding studies in which the control group received an alternative intervention.
- Student academic achievement as the dependent variable, with independent measures, that is, not developed by the evaluators or developers of the TPD program (Cheung & Slavin, 2016; Wolf & Harbatkin, 2023).
- A minimum duration of 12 weeks from the beginning of the intervention to the posttest, in order to have findings of practical value. Very brief studies have been found to inflate effect sizes (e.g., Gersten et al., 2014; Nelson & McMaster, 2019).
- Baseline equivalence demonstration between the two conditions, namely, pretest differences smaller than 0.25 *SD* for the analytical sample (What Works Clearinghouse, 2022).
- Differential attrition during the intervention between the two conditions smaller than 15 % at the level of assignment (e.g., school, class) or at the student level.
- Sufficient data to calculate a standardized mean difference (*SMD*) effect size (Hedges' *g*).
- Studies published between January 1990 and December 2021 in English.

2.3. Selection process

We started the study selection process by locating studies included in previous syntheses ($n = 481$ studies). The database searching yielded a total of 3966 records (i.e., articles and reports). We were left with 3823 records after removing duplicates (see Fig. 2 for the complete PRISMA selection diagram). The titles and abstracts of these records were screened using *Covidence* software, which supports the systematic review process of screening, full-text review and coding. The screeners met regularly to create and refine the screening tool that guided the selection process. We then retrieved the full texts of the 642 records that passed the first screening. Two independent reviewers examined the full texts by applying the eligibility criteria in *Covidence*. Another 146 records were identified by hand searching. After removing 89 duplicates, the remaining records were also double-reviewed. We contacted the study authors if their publication(s) did not include all the required data to make a decision, asking for additional information with respect to our inclusion criteria. The interrater reliability (IRR) for this phase was initially 93 %. When disagreements occurred between the two reviewers on the inclusion of a study, these were resolved through a discussion and the involvement of a third reviewer, reaching 100 % IRR. In total, 128 records with 128 studies (356 effect sizes) met our inclusion criteria.

2.4. Study coding

We coded study characteristics using MUTOS (Methods, Units, Treatments, Outcomes, Settings), a framework initially developed by Cronbach (1982) as UTOS for experimental studies, to assess the extent to which study results can be generalized to the population. U labels unit characteristics, including to whom (e.g., grade level, type of population) the results can be generalized. T labels treatment features, indicating how the intervention was meant to be delivered. O refers to outcome characteristics and is about the intervention effects that were measured. S refers to setting features, indicating to what context (e.g., a country, teacher characteristics, school type (s)) the results can be generalized. The UTOS framework was expanded for meta-analytic purposes by adding the Methods component (Aloe & Becker, 2009), to consider the impact of study design characteristics on intervention results.

Before beginning data extraction, a codebook (see Table 2) was drafted and revised after jointly coding a sample of studies, in order to reach alignment between coders. One of the authors coded the studies for Methods, Units, Outcomes, and Settings in a spreadsheet. A random sample of 10 % of the studies were then independently coded by a second author. We assessed the coders' IRR for relevant characteristics coded using Cohen's kappa. The results per variable ranged from 0.70 to 0.94, showing substantial (0.61 - 0.80) or perfect (0.81–1.00) agreement (Landis & Koch, 1977). Discrepancies were discussed and resolved between the coders, reaching 100 % agreement.

As stated, information on TPD treatment characteristics was obtained from the publications and from additional interviews with those who were knowledgeable about the TPD (authors and/or TPD trainers). For this purpose, TPD characteristics were first coded by a researcher before the interview. The results were then sent to the interviewees, to enable them to think about the questions and answers before the interview. In the interview, we checked our results and also asked for missing information. The interviews were done by two researchers, who reached agreement about the codes after the interview, at which point codes were added and corrected if necessary, based on the extra information obtained in the interview. A total of 115 interviews were held with researchers and trainers in many parts of the world. We were unable to conduct an interview for only 10 % ($n = 13$) of the included records. For those records, two researchers coded the treatment characteristics independently based on the publication, and then compared their codes.

Quality appraisal. As we used strict inclusion criteria in order to select only high-quality studies, it was not necessary to perform a quality appraisal of the included studies. Nonetheless, we coded methodological characteristics including study design, assignment level, and publication status.

Missing data. During the data extraction phase, we screened data for missingness. If data about moderators were missing in studies, we coded them as "not reported". We then searched for the information in other reports about the same study or contacted authors, asking for the missing data. After this process, data related to SES and (some) TPD principles were still missing for some studies. For SES, we considered studies with unreported SES information as having been conducted in schools with average/high SES, assuming that low-SES contexts would have been reported. If information about a TPD principle was still missing after the interview, we considered that principle not to be included in that TPD intervention.

Table 2
Codebook of moderators, levels, and explanations.

Moderator	Levels and explanations
Methods	
Study design	(a) RCT = randomized controlled trials (b) QED = quasi-experimental designs Cheung and Slavin (2016) found that the type of research design affects the intervention impact, with larger effect sizes for QEDs compared to RCTs. In our review, only five out of 128 studies were quasi-experiments. We therefore did not use study design as a moderator in our analyses.
Assignment unit	(a) School = school-level assignment (b) Teacher/class = class- or teacher-level assignment (c) Student = individual assignment Assigning groups of participants to condition instead of individuals is often used to minimize intervention contamination between treatment and control groups. Following previous research (e.g., Torgerson, 2001; Williams et al., 2022), we wanted to take into account the potential confounding effect of assignment level, but only nine studies with student-level assignment were included. We therefore did not use assignment unit as a moderator in our analyses.
Publication status	(a) Published = published in journals, as books, or book chapters (b) Unpublished = issued as dissertations, conference abstracts, or reports from organizations (e.g., IES, EEF) The issue of publication bias and selection bias is well-known in meta-analysis. Polanin et al. (2016), for example, found a larger effect size of 0.18 for published studies than for unpublished studies in their analysis of 383 meta-analyses.
Sample size	(a) fewer than 250 students (b) more than 250 students (based on Cheung & Slavin, 2016) Previous studies found an association between student sample size and effect sizes (Cheung & Slavin, 2016; Kraft et al., 2018). Because of our inclusion criteria (at least 20 teachers per condition), student sample size in our studies ranged between 105 and 29,995 students, with only nine studies with fewer than 250 students. We therefore did not include sample size in the analysis.
Units	
Socio-economic status	(a) Low SES = 60 % or more of the students receiving a free lunch (b) Average/High SES = fewer than 60 % of the students receiving a free lunch For countries other than the US and UK (very few countries in our data set), SES often was not reported as such, and we therefore used other measures such as family income. Van Ewijk and Sleegers (2010) found a differential effect on student performance at different SES levels. Studies with missing information were included in the average/high SES category.
Targeted population	(a) All = interventions targeted at all students in class (b) Specific = interventions for a specific group of students (i.e., low achievers, English language learners - ELLs). Research has shown that studies conducted with homogeneous student groups lead to larger effects compared to those with heterogeneous groups (Bakker et al., 2019; Kraft, 2020). In our review, only 17 out of 128 studies focused on specific populations of students. Thus, we did not include this variable in the moderator analysis (see Model building process for details):
Treatments	
TPD goal	We initially coded nine different TPD goals (see the "TPD goals distinguished" section), but we only found enough studies for our analyses for the following categories: (a) (P)CK = TPD for (pedagogical) content knowledge (b) TPD for comprehensive approaches = interventions in which teachers either learn how they can implement and use a set of related intervention components in an integrated way, or how to use general, instead of subject-specific instructional approaches (c) TPD for curricula = TPD to support the implementation of a curriculum (d) TPD for (digital) tools = TPD to support the implementation of a (digital) tool (e) Other = TPD intervention not included in the previous categories
Number of TPD hours	We coded this as a continuous moderator expressed as the number of hours of workshops and/or coaching received by teachers.
TPD trainer	(a) External trainer = TPD facilitators who were independent from the TPD program and the evaluation of the TPD program (b) Researchers/developers = the developers of the program or the researchers who conducted the study delivered the TPD program See the "Potential moderators" section for an explanation of the levels.
TPD principles	We coded each of the four identified TPD principles (teacher coaching, cooperation, performance standards, and self-regulation) as a dummy variable: 0 = studies that did not include this principle 1 = studies that included this principle
Number of principles	(a) Zero = none of the principles was included in the TPD (b) One = one of the principles was included in the TPD (c) Two = two of the principles were included in the TPD (d) Three = three of the principles were included in the TI (e) Four = four of the principles were included in the TPD See the "Potential moderators" section for an explanation of the levels.
Outcomes	
Tested subject	(a) Reading = outcome measures in the reading domain (b) STEM = outcome measures in the mathematics and science domains (c) Other subject = outcome measures in one or more other subject domain(s) Previous research suggested diverse levels of effectiveness based on the subject tested (de Boer et al., 2014; Kraft, 2020). We first coded the specific subject (reading, mathematics, science, writing, language) and then combined writing and language in the category "Other subjects", and mathematics and science in the category "STEM", because of the small number of studies in which writing, language, and science were tested at the student level.
Settings	
Grade level	(a) Elementary = pre-K to grade 6 (b) Secondary = grades 7 to 12

(*continued on next page*)

Table 2 (*continued*)

Moderator	Levels and explanations
	(c) Mixed = a combination of students in different grades Bloom et al. (2008) showed that annual learning gains are large in the lower grades (K, grade 1) and gradually decrease thereafter. In line with this, researchers have studied differential intervention effects for school levels (Cheung & Slavin, 2016; Kraft, 2018) and groups of grades (early years, primary years; Sims et al. 2023). They did not find significant differences.
Country	(a) US = studies conducted in the US (b) UK = studies conducted in the UK (c) Other countries = studies conducted in other countries After coding, 90 % of the UK studies were unpublished and 100 % of the studies in other countries were published. We therefore decided to include Publication status as a potential moderator and exclude Country as a potential moderator from the analyses.

2.5. Effect size calculation

Hedges' *g* was used as the effect size measure, to correct for small sample bias. We calculated effect sizes adjusted for pretest and other baseline covariates included in the primary studies, using the procedures described by Taylor et al. (2022), and Lipsey and Wilson (2001) and based on the statistics provided in the primary studies (e.g., *F*-tests, regression coefficients, means and standard deviations). When studies used clusters as the unit of assignment, effect sizes and variances were adjusted for clustering based on Hedges (2007).

2.6. Meta-analytic models

We used a random-effects model to estimate the average effect size and explore heterogeneity in TPD effects, as we did not expect that studies from different settings and TPD interventions would share a true underlying effect. Several included studies reported more than one outcome for the same sample (resulting in a correlated effects structure), and/or included multiple subsamples (e.g., multiple treatment and control conditions), resulting in a hierarchical effects structure. We had intended to use a correlated hierarchical effects (CHE) model with robust variance estimation (RVE) to take into account both types of dependencies (Pustejovsky & Tipton, 2022). However, the use of the CHE model was precluded by the large imbalance in the number of effect sizes between the included studies. We therefore used the correlated effects (CE) model with RVE that permits accounting for dependencies among effects within studies without requiring knowledge of the covariance structure, assuming no within-study variation, and producing robust variance estimates (Hedges et al., 2010; Pustejovsky & Tipton, 2022). Based on Pustejovsky and Tipton (2022), we assumed a correlation of $\rho = 0.80$ for effect sizes nested in studies to estimate the CE model. We used the *robumeta* R package (Fisher et al., 2023) that also includes small sample corrections to fit the CE model. We reported the model estimates as unreliable if the Satterthwaite degrees of freedom were below four (Fisher & Tipton, 2015).

We quantified heterogeneity of study effects by calculating a 95 % prediction interval (PI), which is the range of values in which we expect that 95 % of the effect sizes from comparable populations will fall (Borenstein et al., 2017). As suggested in previous reviews (e.g., Williams et al., 2022), we calculated the estimated percentage of true effect sizes that were larger than 0, larger than 0.05 and larger than 0.20, respectively, the latter two being the thresholds indicated by Kraft (2020) for medium and large effects. We used cluster-bootstrapped confidence intervals in the *MetaUtility* R package (Mathur et al., 2021). We used multiple meta-regression to explore potential moderators of the effects. After modeling heterogeneity, we computed conditional means for categorical moderators, adjusted for all covariates included in the meta-regression model.

2.7. Model building process

We first estimated a null model to determine the average effect size (RQ1) controlling for publication status as a methodological confounder. The plan was to also control for assignment unit and study design; however, only nine studies used student-level assignment and only five studies were QEDs. For RQ2 and RQ3, we examined heterogeneity among studies performing a moderator analysis. We centered all moderators to facilitate the interpretation of the intercept as the average adjusted effect size. We distinguished between a confirmatory meta-regression model and an exploratory model guided by our learning theory perspective. As presented in the introduction, we expected a differential impact for TPD goals, TPD trainer, number of TPD hours, and the number of learning theory principles applied in the design of the intervention. Based on MUTOS, we also considered other study characteristics as potential moderators. We intended to include in the meta-regression model the variables targeting population, SES, tested subject, grade level, and country (see Table 2). However, after coding the studies it proved to be impossible to include targeted population and country because not all moderator values were well enough represented in our body of studies. For example, only 17 studies targeted specific populations, while 111 included all students. Thus, the confirmatory meta-regression model included grade level, SES, tested subject and TPD goals central to RQ2.

Number of TPD hours, TPD trainer, and number of learning theory principles were not included as variables in the confirmatory model because those factors were strongly interrelated. For example, the (P)CK and TPD for curricula goals had a balanced number of studies for each level of the variable "number of learning theory principles", while the category Other TPD goals had few studies for each level. We tested these additional TPD moderators in an exploratory way, including all moderators from the confirmatory model and adding the number of TPD hours, TPD trainer, and the number of learning theory principles. For moderators with multiple

categories, we conducted tests of multiple-constraint hypotheses using the Wald test function provided in *clubSandwich* (Pustejovsky, 2023).

2.8. Sensitivity analysis

In the CE model, we tested the assumption of the correlation between effect sizes with sensitivity analysis, assessing results for $\rho =$ 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. Results showed consistent estimates of coefficients and standard errors (see Table S4.1). Thus, we used the $\rho = 0.08$ as suggested by Pustejovsky and Tipton (2022).

A sensitivity analysis was performed to test the influence of outliers. Outliers tend to affect the average effect and heterogeneity, leading to biased conclusions (Viechtbauer & Cheung, 2010). We screened the data for potential outliers using methods used in other meta-analyses (Myers et al., 2022, 2023): computing skewness and kurtosis statistics, and visual examination of the box plot. Values of skewness and kurtosis between ±1.96 were considered to be acceptable (Doane & Seward, 2011). The skewness statistic (1.09) was within acceptable values, while the kurtosis statistic (5.60) was outside the range. Visual analysis of the box plot confirmed the presence of potential outliers. We considered an effect size to be an outlier if it was more than 1.5 times the interquartile range above the 75th percentile, or below the 25th percentile. To prevent data loss, we winsorized the 15 identified outliers to 1.5 times the interquartile range above the 75th percentile and the interquartile range below the 25th percentile. We conducted separate analyses for the full sample of studies with the original and winsorized data, in order to evaluate our results' sensitivity to outliers (Harwell & Maeda, 2008; Myers et al., 2023).

Table 3
Characteristics of the included studies ($N = 128$).

Moderator	Level	*N* (%)
Method characteristics		
Assignment unit	School	93 (65.0)
	Student	9 (6.3)
	Teacher/Class	41 (28.7)
Research design	RCT	138 (96.5)
	QED	5 (3.5)
Publication status	Published	75 (52.4)
	Unpublished	68 (47.6)
Unit characteristics		
Targeted population	All	126 (88.1)
	Specific	17 (11.9)
SES	Low SES	55 (38.5)
	Average/High SES	60 (42.0)
	Not reported	28 (19.5)
Treatment characteristics		
TPD Goal	(P)CK	40 (28.0)
	Comprehensive approaches	31 (21.7)
	TPD for curricula	51 (35.7)
	TPD for (digital) tools	12 (8.4)
	Other	9 (6.3)
TPD trainer	Researchers/developers	102 (71.3)
	External trainer	41 (28.7)
Number of TPD principles	Zero	19 (13.3)
	One	40 (28.0)
	Two	39 (27.3)
	Three	32 (22.4)
	Four	13 (9.0)
		M (SD)
Number of TPD hours		48.37 (43.83)
Outcome characteristics		
Tested subjects	STEM	99[a] (27.1)
	Reading	234[a] (65.3)
	Other subjects	23[a] (7.6)
Setting characteristics		
Grade level	Primary	86 (60.1)
	Secondary	29 (20.3)
	Mixed	28 (19.6)
Country	U.S.	106 (74.1)
	U.K.	30 (21.0)
	Other countries	7 (4.9)
Total studies (*N*)		143[b]
Total effect sizes		356

Note. N = number of studies, [a] Number of effect sizes instead of studies. [b] Because 15 reports reported more than one intervention compared to the same control group, the number of studies displayed in this table is 143. However, in the analyses we considered these samples as dependent, for a total number of 128 studies.

In categorizing SES, we assumed that studies with unreported SES information had been conducted in schools with average/high SES students. Since this is a strong assumption, we ran another meta-regression model to check whether the results changed if three categories (i.e., Low SES, Average/high SES, Not Reported) were used instead of two (Low SES, Average/high SES).

2.9. Selective reporting bias

Although we used search strategies aimed at finding as many unpublished studies as possible, our findings might be influenced by selective reporting bias (i.e., publication bias). The potential for bias in the selection of publications can threaten the validity of the findings, as it often results in an overestimation of the average effect size. We used two methods to assess the robustness of our review to selective reporting bias. First, we included publication status (unpublished vs. published) as a predictor in the meta-regression model. Second, we used selection modeling, which involves using a weight-function model to estimate the probability of selection based on the *p*-value in random-effects meta-analyses, providing an estimate of the degree to which selective reporting bias may be present (Citkowicz & Vevea, 2017). To apply the weight-function model, we used the *weightr* R package (Coburn & Vevea, 2022), adding several cut points (see Table S4.4).

3. Results

3.1. Descriptive characteristics of the included studies

We included 128 studies with 356 effect sizes in which the effects of TPD interventions for in-service teachers on student academic achievement were evaluated. In Table 3, we provide the descriptives for the MUTOS characteristics that were coded. More detailed characteristics are presented per study in Table S3.1 and Table S3.2 (in the Supplemental materials).

3.2. Overall treatment effect

The results of the null model (RQ1) showed a statistically significant, average weighted effect size of 0.09 ($SE = 0.01$, 95 % CI = [0.07, 0.11], $p < .001$). According to the benchmarks for effects measured by means of standardized tests proposed by Kraft, 2020, 2023, this effect size can be interpreted as a medium effect of TPD interventions on student academic achievement. The 95 % prediction interval ranged from −0.15 to 0.32, indicating substantial heterogeneity between studies. The probability that a TPD intervention had an effect larger than 0 was 85 % (95 % CI [77 %, 93 %] whereas the probability that the effect was larger than 0.05 and smaller than 0.20 was 65 % (95 % CI [56 %, 71 %] and 12 % [0 %, 17 %], respectively. Fig. 3 shows the distribution of the observed effect sizes with the 95 % prediction interval (gray section), and the distribution of the empirical Bayes effect size predictions, generated using the full meta-regression model. Fig. 3 visualizes the finding that TPD interventions do not guarantee positive effects on

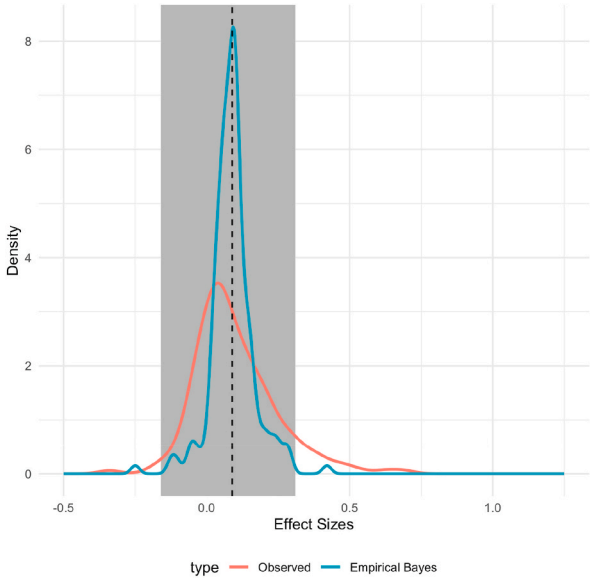


Fig. 3. Distributions of Observed Effect Sizes and Empirical Bayes Effect Size Predictions
Note. The red line shows the distribution of the observed effect sizes. The blue line shows the distribution of the empirical Bayes effect size predictions. The dashed line shows the average effect size adjusted for the moderators included in the full meta-regression model. The shaded box shows the 95 % prediction intervals. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

student outcomes and can also be counterproductive. The challenge is therefore to explain why some TPD interventions work and others do not.

3.3. Moderators of treatment effectiveness

To address RQ2 and RQ3, we used CE models distinguishing between confirmatory and exploratory analysis. In Table 4 we present the results for the confirmatory model (complete meta-regression results are given in Table S5.1). For categorical moderators, the coefficients (i.e., β) are the average effect sizes for each moderator level (conditional means), controlling for all other factors included in the meta-regression model. For continuous moderators, β is the regression coefficient rather than the conditional mean. The Satterthwaite degrees of freedom were greater than 4 for all parameters, indicating that the results were reliable. For RQ2, no statistically significant difference in students' academic achievement related to TPD goals was found. TPD for (P)CK and TPD for curricula had an average effect of 0.10, followed by the Other TPD goal category ($ES = 0.09$). Comprehensive approaches and TPD for (digital) tools had an average effect size of 0.08 and 0.04, respectively.

As far as the method, unit, outcome, and setting characteristics are concerned, the results for RQ3 showed a statistically significant larger effect in primary grades compared to mixed grades ($p = .046$). However, the *F*-test for the multigroup moderator did not reach statistical significance. Table 4 shows similar effect sizes for primary ($ES = 0.10$) and secondary ($ES = 0.09$) school grades, and a smaller effect for mixed grades ($ES = 0.06$). A statistically significant difference ($p = .003$) was found for published ($ES = 0.10$) versus unpublished studies ($ES = 0.05$), in favor of the first group. None of the other study characteristics examined was statistically significantly associated with treatment outcomes.

For the TPD program treatment characteristics, we fitted an exploratory model that added to the previous model TPD trainer, number of TPD hours, and the number of learning theory principles (see Table 5). Considering the exploratory nature of this analysis, we reported the regression coefficients rather than the conditional means. Statistical significance of these moderators should be read with caution. The results of the exploratory model showed consistent results for the moderators included in the confirmatory model, and a positive trend in the estimates for the number of learning theory principles, suggesting that the incorporation of a larger number of learning theory principles in TPD interventions goes together with a greater impact on student learning. Table 6 presents the number of interventions for each principle in different combinations, providing an overview of those appearing most frequently across the different combinations. Nineteen interventions did not include any of the four principles. Considerable variability was observed in the use of only a single principle, with *cooperation* never represented and *coaching* showing the highest frequency. In two-principle combinations, *self-regulation* and *coaching* were the most frequently occurring principles. When three principles were combined, *self-regulation* was consistently present, followed by *performance standards*, and *coaching*. Finally, 13 interventions included all four learning principles.

3.3.1. Sensitivity analysis

The sensitivity analysis for the full sample of studies showed no meaningful differences in the estimates of the null model and the confirmatory model derived from the data with outliers and the data with winsorized values. The weighted average effect size in this

Table 4
Results for the confirmatory meta-regression model ($N = 128$).

Moderator	Mean or β	*SE*	*t*	*df*	*p*	95 % CI
Intercept	0.09	0.01	7.42	72.4	<0.001	[0.06, 0.11]
Publication status			3.07	91.9	0.003	
Published	0.10	0.02				[0.08, 0.14]
Unpublished	0.05	0.01				[0.02, 0.08]
SES			−1.39	58.6	0.169	
Low SES	0.07	0.01				[0.05, 0.10]
Average/High SES	0.10	0.02				[0.07, 0.13]
Grade level			2.28	36.3	0.117	
Primary	0.10	0.02				[0.07, 0.13]
Secondary	0.09	0.02				[0.06, 0.12]
Mixed	0.06	0.02				[0.03, 0.09]
Tested subject			1.15	35.2	0.328	
Reading	0.09	0.02				[0.06, 0.13]
STEM	0.07	0.01				[0.05, 0.10]
Other subjects	0.13	0.04				[0.04, 0.21]
TPD goals			1.25	37.7	0.306	
(P)CK	0.10	0.03				[0.04, 0.16]
Comp. Appr.	0.08	0.01				[0.05, 0.10]
Curricula	0.10	0.02				[0.06, 0.13]
Digital tools	0.04	0.02				[0.00, 0.09]
Other	0.09	0.04				[0.01, 0.17]

Note. The first column reports conditional means for categorical moderators (Mean) and regression coefficients (β) for continuous moderators (i.e., the number of TPD hours). *SE* = standard error; *df* = Satterthwaite degrees of freedom; CI = confidence interval. The *p* values assess the statistical significance of moderators, not whether each conditional mean for categorical moderators differed from zero.

Table 5
Results for the exploratory meta-regression model ($N = 128$).

Moderator	β	*SE*	*t*	*df*	*p*	95 % CI
Intercept	0.09	0.01	7.71	60.9	<0.001	[0.07, 0.12]
Published	0.05	0.02	2.73	72.8	0.008	[0.01, 0.09]
Low SES	−0.02	0.02	−1.40	47.8	0.169	[-0.05, 0.01]
Primary	0.04	0.02	1.91	28.2	0.067	[-0.00, 0.07]
Secondary	0.04	0.02	1.62	26.4	0.117	[-0.01, 0.09]
Reading	−0.03	0.04	−0.73	19.5	0.475	[-0.12, 0.06]
STEM	−0.05	0.04	−1.41	20.8	0.172	[-0.13, 0.03]
(P)CK	0.01	0.05	0.28	12.0	0.786	[-0.09, 0.11]
Comp. Approaches	−0.04	0.04	−1.07	15.1	0.302	[-0.11, 0.04]
TPD for curricula	0.02	0.04	0.54	11.3	0.600	[-0.06, 0.10]
TPD for digital tools	−0.04	0.04	−0.91	17.4	0.378	[-0.14, 0.05]
Number of TPD hours	0.00	0.00	0.96	9.2	0.361	[-0.00, 0.00]
Trainer: Res/Dev	0.00	0.03	0.12	58.8	0.909	[-0.06, 0.06]
One principle	0.03	0.02	1.09	29.6	0.287	[-0.02, 0.07]
Two principles	0.03	0.02	1.16	32.2	0.257	[-0.02, 0.07]
Three principles	0.05	0.03	2.05	30.1	0.049	[-0.00, 0.10]
Four principles	0.13	0.09	1.41	23.4	0.171	[-0.06, 0.33]

Note. The first column reports regression coefficients (β) for both categorical and continuous moderators; *SE* = standard error; *df* = Satterthwaite degrees of freedom; CI = confidence interval. The *p* values assess the statistical significance of moderators.

Table 6
Number of interventions using one, two, three, or four principles, per principle.

	One principle ($n = 40$)	Two principles ($n = 39$)	Three principles ($n = 32$)	Four principles ($n = 13$)
Coaching	24	27	21	13
Cooperation	0	11	14	13
Performance Standards	6	10	29	13
Self-regulation	10	30	32	13

Note. $N = 143$, Nineteen interventions did not report using any principles.

second null model was 0.09 (*SE* = 0.01, 95 % CI = [0.07, 0.11], $p < .001$). In the confirmatory model with winsorized effect sizes, the characteristics that were significant moderators in the main analysis remained significant or marginally significant. Since little variation was found after winsorizing the values, we present the results that were obtained using the original dataset as the results of the main analysis in the article, and show the results of the sensitivity analysis in Table S4.2.

A sensitivity analysis on SES revealed no statistically significant differences in the estimates between low SES and average/high SES, showing similar results for the model in which studies with missing information were categorized as "Not reported" (Table S4.3) as compared to the main analysis that we presented here.

3.4. Selective reporting bias

Of the studies included in our review, 48 % came from the gray literature, which is a high percentage. When including publication status as a predictor in the confirmatory model, published studies showed twice the effect size of unpublished studies, suggesting a significant difference. The selection model approach using cut points presented in Table S4.4 (Coburn & Vevea, 2022) returned a mean effect estimate upwardly adjusted at each step (e.g., *ES* = 0.14 for the first cut point) compared to the unadjusted estimate, counter-intuitively indicating that effect sizes with large *p*-values ($p > .05$) were more likely to survive selection than statistically significant positive effects ($p < .05$). The results were consistent when controlling for moderators (Table S4.4).

4. Discussion

The present meta-analysis aimed at investigating what impact TPD interventions have on student outcomes and what characteristics of TPD programs moderate this impact. The first research question is based on the fact that TPD activities are not a goal in themselves, but are undertaken to improve teacher competence, classroom practice and student outcomes. Given the enormous sums of money invested in this, it is important to investigate if that goal is accomplished or not, which is likely the reason that our first research question has also been addressed in many other TPD meta-analyses. Answering that question was also a prerequisite for quantifying and investigating the heterogeneity of TPD effects (Pigott & Polanin, 2020). We found a positive average effect of TPD, but also large variability in TPD effects (ranging from -0.15 to 0.32), so there was no guarantee of TPD effectiveness. This means that answering the second and third research questions about the program-related moderators of TPD impact becomes essential.

We consider that we brought the TPD field a step further by conducting the most comprehensive review thus far of TPD research. No other review of TPD research has covered such a long period or included so many high-quality TPD studies. The high number of unpublished studies that we found (about half of all included studies), was crucial for obtaining balanced, not-overestimated findings.

Moreover, no other review has involved additional interviews (115 in total) conducted to check precisely what happened in the TPD programs. Doing this resulted in more valid findings than when only the often very limited data reported in primary studies are used (as usually done in meta-analyses). In summary, the collected data present the best evidence for conclusions about the effects of TPD on student outcomes and the program-related moderators of these TPD effects. Although other reviews of TPD research have addressed these same research questions, those reviews did not achieve the same level of thoroughness and evidentiary value.

4.1. Average treatment effect (RQ1)

We found a statistically significant, medium, average weighted effect of 0.09 of TPD interventions for in-service teachers on students' academic achievement, as measured by means of standardized tests (Kraft, 2020). Kraft (2020) distinguished between small (smaller than 0.05), medium (0.05 and larger, but smaller than 0.20) and large (0.20 and larger) effect sizes, based on analysis of 1942 effect sizes that were found for RCTs in which a standardized student achievement test was used to measure the impact of an educational intervention. Across the 1942 effect sizes, Kraft (2023) found an empirical weighted average effect size of 0.06 for all sorts of interventions to improve student performance (not only TPD interventions) for both reading and mathematics, and stressed that interventions with enormous effects are rare, but that small effects are important, as they can support the incremental improvement of educational practice. Moreover, if thousands or even millions of students can benefit from such effective TPD interventions, when they are implemented at large scale, then the impact will be enormous. The comparison with studies in which standardized, independent tests were used is important (comparing like with like; cf. Bakker et al., 2019), as we only included such studies. Research has shown that considerably larger effects are found for studies in which researcher- or developer-made tests are used (2.3–2.7 times larger than in studies that use independent broad tests, Wolf & Harbatkin, 2023). The reason for this probably being that researcher- or developer-made tests are over aligned with the intervention content. We, however, do not know to what extent the standardized tests used in our studies aligned with the intervention content of the studies (i.e., with the aspect of student learning that the interventionists attempted to improve). It could, for example, be that some of them were to some extent under aligned, and it may, thus, be that the characteristics of the standardized tests have influenced our findings. This is a topic worth investigating in future research, but difficult as researchers do not always report these matters well.

Selection modeling analysis showed that effect sizes with large *p*-values ($p > .05$) were more likely to survive study selection than statistically significant positive effects ($p < .05$). This may be due to the fact that a large proportion of the unpublished studies that were included were funded by the Institute of Educational Studies (IES) in the United States and the Education Endowment Foundation (EEF) in the United Kingdom. These organizations use very strict requirements that more often lead to studies with small effect sizes, which are thus often statistically non-significant, than studies that do not meet these strict criteria.

Our effect size is in line with the recent meta-analyses of TPD research conducted by Fletcher-Wood and Zuccollo (2020) and Sims et al. (2023), who found effect sizes of 0.09 and 0.05 respectively. It is remarkable that our average effect size is even somewhat higher than the 0.05 in Sims et al. (2023) and equal to the 0.09 in Fletcher-Wood and Zuccollo (2020), given that our inclusion criteria were much stricter than the criteria used in both reviews in order to avoid methodological bias. The fact that we included more and different studies than in the other reviews might have played a role here.

We also found that the heterogeneity of treatment effects was large, with some TPD interventions having negative effects on student achievement (about 15 % of the TPD effects are 0 or negative). The fact that effect sizes varied is not surprising, as our included studies were about TPD interventions that differed considerably in terms of their goals (e.g., teachers' ability to teach dialogically versus teachers' ability to use a specific curriculum) and TPD approaches. Zooming in on the specific characteristics of TPD programs can help to clarify why some forms of TPD are more effective than others. As mentioned previously, we conducted that analysis from a learning theory perspective. We present the results of the analysis below.

4.2. Effects per TPD goal (RQ2)

TPD goal did not prove to be a statistically significant moderator of student outcomes in our set of studies. Most goals had an effect size ranging between 0.07 and 0.10; however, TPD for the implementation of (digital) tools had an effect size of 0.04. Based on our data, one thus could conclude that it may not matter much which of the goal categories TPD focuses on. They may all be similarly effective, not just TPD for (P)CK, as it has been claimed frequently in the TPD literature (e.g., Garet et al., 2008, 2010, 2011, 2016; Desimone, 2009; Maandag et al., 2017). However, we do not know much yet about the effectiveness of many other important TPD goals (Kaplan et al., 2020), because of the lack of high-quality studies addressing them. More in general, our findings for RQ2 (and RQ3) should be interpreted with caution as they examine the influence of moderators on the relationship between TPD and student learning, rather than the causal effects.

4.3. Moderating effect of study and TPD characteristics (RQ3)

Several study characteristics were found to moderate the effect of TPD interventions on student achievement. Effects for primary school grades were statistically significantly higher ($ES = 0.10$) than effects for mixed grade levels ($ES = 0.06$), and we found the same average effect for secondary school grades, but that was not statistically significant when compared with mixed effects (fewer studies).

We found that published TPD studies had larger effect sizes than unpublished ones (0.10 versus 0.05), which reflects the general publication bias problem (Polanin et al., 2016) and should urge researchers into TPD to search for and include as many unpublished reports in their reviews as possible to present a valid picture of TPD effects. None of the other study characteristics investigated (target

population, SES, the outcome subject) proved to be statistically significantly associated with TPD effects. It is remarkable that we did not find that the target population had a moderating effect in our data, as it has been argued in the literature (Simpson, 2018) that it is easier to obtain positive intervention results with restricted (i.e., more homogeneous) samples. Only 12 % of our studies were conducted with restricted samples, which might explain this finding. Another noteworthy finding is that we did not observe a moderating effect of SES, as, in general, there is a strong relationship between students' SES and their performance (van Ewijk & Sleegers, 2010). This may have been caused by the fact that SES was not reported in 20 % of the studies and by the fact that we had to combine average and high-SES students into one group (less discriminative).

As far as the outcome subject is concerned, it is important to know that 65 % of the outcomes were related to reading, 27 % to STEM subjects and the rest to all other subjects combined. This implies that we do not yet know the impact of TPD for other subjects.

An important finding is that if principles from learning theory are applied in combination in TPD interventions, then students seem to perform better. TPD research has generally studied the explanatory power of separate factors. Our cluster approach was recently also used by Sims et al. (2023), who also found indications that TPD designs including more mechanisms from the field of behavior change were more effective. This finding seems to indicate that there is not a single silver bullet for improving such a complex skill as teaching, and that combinations of several principles may be required. That might especially be the case for moderators that together act as powerful forms of teacher support and strengthen each other. Our data showed that of the four principles, teacher self-regulation and teacher coaching were applied most often in TPD designs (in 72 interventions each), performance standards (45 interventions) and teacher cooperation (25 interventions) somewhat less often. In 13 % of the TPD interventions, none of these four principles was applied, and in 9 % of them all four principles were applied. Thus, there is much room for improvement in terms of the degree to which the four principles are used; for example, in about 80 (!) of our studies there was no clear picture showing teachers what the desired teacher behavior looked like (i.e., performance standards), which then makes it hard to learn that behavior as a teacher.

Darling-Hammond et al. (2017) argued that effective TPD is of "sustained duration". Van Veen et al. (2012) have made claims for effective TPD lasting at least a semester. TPD researchers and interventionists often argue that their intervention did not last enough to be effective. It is interesting that TPD interventions with more TPD hours were not necessarily more effective in our high-quality studies. It could be that in interventions that last longer, after a while teachers become less motivated, as a result of which TPD effects could decrease.

Having TPD trainers be developers/researchers versus external trainers did not seem to matter for student achievement. In most of our interventions, the trainers were developers/researchers (70 %). We expected external trainers to be less committed to making the intervention work (as it is not their "baby") and maybe also to be somewhat less knowledgeable about the intervention content, and less effective as a consequence. The lack of a larger set of really small studies in the set we reviewed may have played a role for our finding, as in small studies developers/researchers might be especially able to do everything they can to make the intervention work ("super realization", Slavin & Smith, 2009). It could also be that external trainers are more experienced with providing training in general in schools, and in that way compensate for their potentially lower levels of commitment. The finding that the two types of trainers did not differ in effectiveness is a positive finding, as external trainers will always be required for scaling up effective TPD interventions in educational practice.

4.4. The contribution of the study to TPD research and theory

The findings of our study show that even when very strict study inclusion criteria are used to minimize methodological bias and to maximize ecological validity (findings that are representative of what might be expected in educational practice), TPD effects on student achievement are on average considerable, which is an important finding for the TPD knowledge base. As said, effects also varied considerably, which calls for a scientific explanation of the variability.

4.4.1. The importance of the learning theory perspective for TPD research and theory

With this study, we highlight the relevance of the learning theory perspective for achieving a better understanding of why students learn a lot in some TPD programs, but little to nothing in others. The core of TPD is *learning complex professional skills*, and it is striking how little attention so far has been paid in TPD research to the learning theory perspective (and how much attention to other types of moderators). *Learning* complex professional skills is hard, and the forms of support offered to teachers (scaffolding) are therefore crucial. We studied four such forms of support for teacher learning. First of all, a clear definition of what "good" looks like is important. For example, when teachers participate in a TPD program for classroom differentiation, then the TPD facilitators ideally make it clear to participating teachers what a teacher who differentiates well does, and why that teacher does those things (their mental models of how things work). We found that such standards for "good" performance were often lacking in TPD programs, which makes learning difficult, as teachers then cannot work well on achieving that type of performance. This first learning theory principle influences the second one, teacher self-regulation. If what counts as "good" performance is clear, then teachers can compare their own competence with "good" performance, and plan how to improve it further. Classroom coaching and teacher cooperation (as a means of mutual support and knowledge exchange) were the other learning theory principles studied.

In our exploratory regression analysis, we saw that the more these four learning principles were applied in TPD programs, the larger the positive TPD effect on student outcomes: TPD effects going from 0.03 to 0.13. In particular, TPD programs using all four learning theory principles had a larger effect on student learning, which is very interesting from a theoretical point of view, and makes this worthy of being studied in greater depth in TPD research.

Our deliberate approach of studying these four principles *as a set* instead of as separate variables was chosen because we think that it is not enough only to have, for example, a clear definition of "good." For TPD effectiveness, we also consider classroom coaching,

teacher self-regulation and teacher cooperation to be necessary. It appears that studying *combinations of variables* can help us explain more variance in TPD effects. Cohen et al. (2003) and Hiebert et al. (2005), based on their research in other contexts, also argued that combinations and interactions of variables have important meanings that are not found when studying individual variables separately. We therefore did not consider whether one of these four principles or a particular subset of them moderated TPD effectiveness more than the others. This would also have been impossible to do, as there were too many potential combinations and some combinations had very few studies. However, Table 6 presents the number of interventions for each principle in different combinations, providing an overview of the most prevalent ones. The table indicates that there is a lot of room for improving TPD programs.

In our view, learning theory in general deserves much more attention in TPD research. There are several other features of effective instruction in the context of learning professional skills that likely also play a role in TPD. For example, it could matter whether *TPD content* (e.g., only declarative or procedural knowledge or, in the case of more complex TPD goals, combinations of knowledge, skills and attitudes) and *TPD method* (e.g., prior knowledge articulation, whole tasks representative of what teachers have to be able to do in the classroom, spaced practice with cognitive feedback, interleaving and differentiation in line with differences between teachers) match the teacher competences required for TPD goal accomplishment (e.g., classroom management, classroom differentiation). As TPD researchers, there is much for us to be learned from the already existing knowledge base on learning complex professional skills effectively in general (i.e., in professions other than teaching; e.g., van Merriënboer & Kirschner, 2018).

4.4.2. Other moderators of TPD effectiveness

In addition to the learning theory principles, we found that TPD seems to be differentially effective, being more effective in primary grades than in mixed grade levels, which may be related to the fact that teachers in general simply have more impact on student learning in the lower grades than in higher grades (Bloom et al., 2008), because students then still know relatively little subject-matter content. This finding should be taken into account when comparing the results of TPD studies conducted at different levels of the educational system.

Moreover, in our data, we did not find evidence that TPD is differentially effective for different TPD goals, although in many TPD studies and meta-analyses it has been claimed that TPD is especially effective when it focuses on teachers' (pedagogical) content knowledge (e.g., Desimone, 2009; Garet et al., 2008, 2010; Lynch et al. 2019; Maandag et al., 2017), on TPD for classroom management (Pellegrini et al., 2021), or on the implementation of new curricula and a better understanding of how students learn (Lynch et al., 2019). There may be many different ways to improve teaching and student learning, which is good news.

Similarly, based on this meta-analysis, we now have strong evidence that a factor that is often mentioned in the literature as crucial for TPD effectiveness, the number of TPD hours, did not seem to matter for TPD effectiveness in our data. It may be more important what you do in a TPD intervention than how many hours you work on TPD in a program. However, the measurement of *long-term TPD effects* could change this picture. In our body of studies, such measurements were rare. In most cases, we only knew the effects of TPD measured immediately after the TPD intervention. It could be that teachers need much more time to master new teaching skills. We did include quite a few long studies; specifically, 48 included studies lasted for more than 1 year, with four of them lasting 3 years. Longitudinal studies are expensive and not easily conducted, but without them knowledge about how TPD effects develop in the long run and under what particular conditions remains out of reach.

4.4.3. Better information for a better understanding of TPD

This study also made it clear that more and better information on several factors is needed from TPD researchers when they report on TPD interventions, so that researchers doing meta-analyses can decide to leave out studies with a high risk of bias and can use as much information as possible to draw nuanced conclusions about TPD effects and their moderators. The more the reporting of TPD research improves along these lines, the more the quality of theories on TPD effectiveness is likely to improve.

For example, strikingly little information was reported in TPD publications on the *evidence base* for the design of a specific TPD program. Why did TPD designers expect that teacher and student learning would improve when teachers participated in their TPD program? What was their theory of improvement (Desimone, 2009)? And was it based only on ideas and expectations, or on solid empirical evidence for the effectiveness of a specific TPD approach? From those TPD programs that provided information on this, we learned that the interventions were at different stages of development. Some programs were the first attempt to improve education on the basis of a specific idea (thus, there was much uncertainty about their effectiveness), whereas others built on previous program versions and evaluations, and as such there was (much) more evidence about their effectiveness. This could be a factor worth investigating further: Are TPD programs with a strong evidence base on average more effective than interventions with little evidence? One would expect that interventions become more effective as a result of trying them out and optimizing them based on research findings, and as such, it may be better to distinguish between categories of evidence bases for TPD interventions.

Drawing nuanced conclusions about TPD effectiveness was also challenged by the fact that the *fidelity of implementation* of TPD programs (e.g., how often teachers attended the TPD meetings, and how well and how long they applied the TPD program content) was investigated only in a minority of studies. We learned that all of the four Desimone (2009) links in Fig. 1 were studied only in 12 of our TPD programs. Thus, in most cases we did not know how much teachers learned in a TPD program, or to what extent and how they applied in the classroom what they had been taught in the TPD program. This black box needs to be opened. Namely, information on Desimone's (2009) Blocks 2 and 3 is necessary in order to learn how they mediate the effects of TPD interventions.

Evaluations of TPD interventions generally only report *average effects* (Bryan et al., 2021; Gelman et al., 2023). Such average treatment effects may be small or zero, but treatments may nonetheless be (very) significant for one or more of the student and teacher groups. If more is learned about which subgroups of teachers and students (e.g., low-versus high-performers) benefit from what kinds of TPD interventions, and which students experience negative effects from what TPD interventions, TPD interventions can be targeted

better, and potential negative effects of basing policy-making and practice on average treatment effects can be prevented.

4.5. Limitations of this study and recommendations for future TPD research

The majority of the studies in our data set were conducted in the US or the UK (95 %). We must therefore be cautious about generalizing the findings, as some TPD interventions may be more effective in some countries than in others. It is unknown whether our findings apply to all school levels (67 % of our included studies were conducted in primary schools) and other school subjects (65 % of our studies were interventions targeted at reading, 27 % at STEM). We studied TPD for several goals, but we did not determine how effective TPD is for other TPD goals, such as improving classroom management, instructional differentiation, and teaching meta-cognitive skills. More high-quality studies are needed to further expand such knowledge with respect to TPD.

We had to deal with missing data for some moderators we aimed to test based on our learning theory perspective. More than one-fourth (28 %) of the studies did not report information on SES, and in those cases we assumed the SES to be average SES. Acknowledging this limitation, we also conducted a sensitivity analysis, which did not reveal statistically significant differences in the estimates between low SES and average/high SES when including studies that did not report SES as a separate category.

4.6. Implications for practice

TPD designers, TPD facilitators, and practitioners in schools and at higher levels of the educational system can benefit from our findings when designing, implementing and selecting TPD interventions. The most important finding, in our view, is that TPD *on average* proves to have a positive, medium effect on student achievement (for the school subjects we could study), that categories of TPD goals did not differ substantially in effectiveness of TPD in our data. At the same time, the considerable heterogeneity in TPD effects implies that there are no guarantees of improved student learning as a result of TPD. TPD effectiveness seems to be related to the characteristics of the TPD program and the contexts in which it is implemented. For example, we found larger average effects for primary education than for mixed grade levels, and in our findings we saw a trend suggesting that the incorporation of a larger number of learning theory principles in TPD programs goes together with a greater impact on student learning, however, this should be investigated further. We hope our findings inspire designers, facilitators and consumers of TPD programs in their complex but valuable TPD activities.

Author statement

Adrie J. Visscher – conceptualization, data collection, writing (original draft, review and editing).
Natasha Dmoshinskaia – conceptualization, data collection, data curation, writing (original draft, review and editing).
Marta Pelligrini – data collection, methodology, data analysis, writing (original draft, review and editing).
Anny Rey-Naizaque – data collection, data curation.

Conflicts of interest

We have no known conflicts of interest to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.edurev.2025.100742.

Data availability

Data will be made available on request.

References

Alexander, R. J. (2015). *Towards dialogic teaching: Rethinking classroom talk* (4th ed.). Dialogos.
Aloe, A., & Becker, B. (2009). Modeling heterogeneity in meta-analysis: Generalizing using Cronbach's (M)UTOS framework and meta-analytic data. *Paper presented at the annual meeting of the American educational Research Association*. New York, NY.
Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics, 102*, 1–8. https://doi.org/10.1007/s10649-019-09908-4
Basma, B., & Savage, R. (2018). Teacher professional development and student literacy growth: A systematic review and meta-analysis. *Educational Psychology Review, 30*, 457–481. https://doi.org/10.1007/s10648-017-9416-4, 2018.
Bill & Melinda Gates Foundation. (2014). *Teachers know best: Teachers' views on professional development*. ERIC Clearinghouse.
Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289–328. https://doi.org/10.1080/19345740802400072
Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*(4), 1065–1105. https://doi.org/10.1177/0149206309352880

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta- analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for all. *American Educational Research Journal, 44*(3), 701–731. https://doi.org/10.3102/0002831207306743

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour, 5*, 980–989. https://doi.org/10.1038/s41562-021-01143-3

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

Citkowicz, M., & Vevea, J. L. (2017). A parsimonious weight function for modeling publication bias. *Psychological Methods, 22*(1), 28–41. https://doi.org/10.1037/met0000119

Coburn, K. M., & Vevea, J. L. (2022). Weightr: Estimating weight-function models for publication bias (R package version 2.0.2) [computer software]. https://cran.r-project.org/web/packages/weightr/weightr.pdf.

Cohen, D. K., & Ball, D. L. (2001). Making change: Instruction and its improvement. *Phi Delta Kappan, 83*(1), 73–77. https://doi.org/10.1177/003172170108300115

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*, 119–142. https://doi.org/10.3102/01623737025002119

Cronbach, L. J. (1982). *The evaluation of educational and social programs*. Jossey-Bass.

Danielson, C. (2013). *The framework for teaching evaluation instrument. Danielson Group*.

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

de Boer, H., Donker, A. S., & van der Werf, M. P. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research, 84*(4), 509–545. https://doi.org/10.3102/0034654314540006

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

Didion, L., Toste, J. R., & Filderman, M. J. (2020). Teacher professional development and student reading achievement: A meta-analytic review of the effect. *Journal of Research on Educational Effectiveness, 13*(1), 29–66. https://doi.org/10.1080/19345747.2019.1670884

Doane, D. P., & Seward, L. E. (2011). Measuring skewness: A forgotten statistic? *Journal of Statistics Education, 19*(2). https://doi.org/10.1080/10691898.2011.11889611

Egert, F., Fukkink, R. G., & Eckhardt, A. G. (2018). Impact of in-service professional development programs for early childhood teachers on quality ratings and child outcomes: A meta-analysis. *Review of Educational Research, 88*(3), 401–433. https://doi.org/10.3102/0034654317751918

Endedijk, M. D., Vermunt, J., Verloop, N., & Brekelmans, M. (2012). The nature of student teachers' regulation of learning in teacher education. *British Journal of Educational Psychology, 82*(3), 469–491. https://doi.org/10.1111/j.2044-8279.2011.02040.x

Fisher, Z., & Tipton, E. (2015). Robumeta: An R-package for robust variance estimation in meta-analysis. https://arxiv.org/pdf/1503.02220.pdf.

Fisher, Z., Tipton, E., & Zhipeng, H. (2023). Robumeta: Robust variance meta-regression [Computer software] Version R package 2.1. https://cran.r-project.org/web/packages/robumeta/index.html.

Fletcher-Wood, H., & Zuccollo, J. (2020). *The effects of high quality teacher professional development on teachers and students. A rapid review and meta-analysis*. *The Education Policy Institute*. https://epi.org.uk/publications-and-research/effects-high-quality-professional-development/.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., … Sztejnberg, L. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4030). *National Center for Education Evaluation and Regional Assistance*. Institute of Education Sciences, U.S. Department of Education https://ies.ed.gov/ncee/pdf/20084030.pdf.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., … Borman, G. D. (2016). *Focusing on mathematical knowledge: The impact of content-intensive teacher professional development* (NCEE 2016-4010). *National Center for Education Evaluation and Regional Assistance*. Institute of Education Sciences, U.S. Department of Education https://files.eric.ed.gov/fulltext/ED569154.pdf.

Garet, M. S., Wayne, A., Stancavage, F., Taylor, J., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., & Doolittle, F. (2010). *Middle school mathematics professional development impact study: Findings after the first year of implementation (NCEE 2010-4009)*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://ies.ed.gov/ncee/pubs/20104009/pdf/20104009.pdf.

Garet, M. S., Wayne, A., Stancavage, F., Taylor, J., Walters, K., Song, M., … Doolittle, F. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation* (NCEE 2011-4024). *National Center for Education Evaluation and Regional Assistance*. Institute of Education Sciences, U.S. Department of Education https://files.eric.ed.gov/fulltext/ED519922.pdf.

Gelman, A., Hullman, J., & Kennedy, L. (2023). Causal quartets: Different ways to attain the same average treatment effect. *The American Statistician, 78*(3), 267–272. https://doi.org/10.1080/00031305.2023.2267597

Gersten, R., Taylor, M. J., Keys, T. D., Rolfhus, E., & Newman-Gonchar, R. (2014). Summary of research on the effectiveness of math professional development approaches. *U.S. department of education, institute of education sciences, National Center for education evaluation and regional assistance, regional educational laboratory Southeast*. http://ies.ed.gov/ncee/edlabs.

Gess-Newsome, J. (1999). Pedagogical content knowledge: An introduction and orientation. In J. Gess-Newsome, & N. G. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 3–17). Springer. https://doi.org/10.1007/0-306-47217-1_1.

Gorard, S., Siddiqui, N., & See, B. H. (2015). *Philosophy for children. Education Endowment Foundation*. https://files.eric.ed.gov/fulltext/ED581147.pdf.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445–470. https://doi.org/10.1086/669901

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*(4), 355–366. https://doi.org/10.1177/01466210022031804

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949–967. https://doi.org/10.1111/j.1467-8624.2005.00889.x

Hanushek, E. A., & Rivkin, S. G. (2006). Teaching quality. In E. A. Hanushek, & F. Welch (Eds.), *Handbook of the economics of education* (Vol. 2, pp. 1052–1078). Elsevier.

Harwell, M., & Maeda, Y. (2008). Deficiencies of reporting in meta-analyses and some remedies. *The Journal of Experimental Education, 76*(4), 403–430. https://doi.org/10.3200/JEXE.76.4.403-430

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta- regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., Hollingsworth, H., Manaster, A., Wearne, D., & Gallimore, R. (2005). Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 Video Study. *Educational Evaluation and Policy Analysis, 27*(2), 111–132. https://doi.org/10.3102/01623737027002111

Jansen in de Wal, J. (2016). *Secondary school teachers' motivation for professional learning*. Open University [Doctoral dissertation].

Kaplan, A., Cromley, J., Perez, T., Dai, T., Mara, K., & Balsai, M. (2020). The role of context in educational RCT findings: A call to redefine "evidence-based practice". *Educational Researcher, 49*(4), 285–288. https://doi.org/10.3102/0013189X20921862

Kennedy, M. M. (1998). *Form and substance in inservice teacher education*. Madison: University of Wisconsin National Institute for Science Education. https://www.msu.edu/~mkennedy/publications/valuePD.html.

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

Kini, T., & Podolsky, A. (2016). *Does teaching experience increase teacher effectiveness? A review of the research. Learning Policy Institute.* https://eric.ed.gov/?id=ED606426.
Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253. https://doi.org/10.3102/0013189X20912798
Kraft, M. A. (2023). The effect-size benchmark that matters most: Education interventions often fail. *Educational Researcher, 52*(3), 183–187. https://doi.org/10.3102/0013189X231155154
Kraft, M., 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
Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42–78. https://doi.org/10.3102/0034654315581420
Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*(2), 363–374. https://www.jstor.org/stable/2529786.
Leigh, A. (2010). Estimating teacher effectiveness from two-year changes in students' test scores. *Economics of Education Review, 29*(3), 480–488. https://doi.org/10.1016/j.econedurev.2009.10.010
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE.
Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis, 41*(3), 260–293. https://doi.org/10.3102/0162373719849044
Maandag, D., Helms-Lorenz, M., Lugthart, E., Verkade, A., & van Veen, K. (2017). *Features of effective professional development interventions in different stages of teachers' careers: NRO report*. Teacher education department of the University of Groningen.
Mathur, M. B., Wang, R., & VanderWeele, T. J. (2021). MetaUtility: Utility functions for conducting and interpreting meta-analyses. https://cran.r-project.org/web/packages/MetaUtility/index.html.
Meelissen, M. R. M., & Luyten, H. (2011). School effectiviteit en prestatieniveau natuuronderwijs in groep 6: secundaire analyses op TIMSS-2007 data [School effectiveness and student performance levels for science in grade 4: Secondary analysis of TIMSS-2007 data]. *Pedagogische Studies, 88*(5), 309–322. https://pedagogischestudien.nl/article/view/14340.
Myers, J. A., Hughes, E. M., Witzel, B. S., Anderson, R. D., & Owens, J. (2023). A meta-analysis of mathematical interventions for increasing the word problem solving performance of upper elementary and secondary students with mathematics difficulties. *Journal of Research on Educational Effectiveness, 16*(1), 1–35. https://doi.org/10.1080/19345747.2022.2080131
Myers, J. A., Witzel, B. S., Powell, S. R., Li, H., Pigott, T. D., Xin, Y. P., & Hughes, E. M. (2022). A meta-analysis of mathematics word-problem solving interventions for elementary students who evidence mathematics difficulties. *Review of Educational Research, 92*(5), 695–742. https://doi.org/10.3102/00346543211070049
Neitzel, A. J., Lake, C., Pellegrini, M., & Slavin, R. E. (2022). A synthesis of quantitative research on programs for struggling readers in elementary schools. *Reading Research Quarterly, 57*(1), 149–179. https://doi.org/10.1002/rrq.379
Neitzel, A., Zhang, Q., & Slavin, R. (2022). *Effects of varying inclusion criteria: Two case studies*. EdarXiv. https://doi.org/10.35542/osf.io/h258x
Nelson, G., & McMaster, K. L. (2019). The effects of early numeracy interventions for students in preschool and early elementary: A meta-analysis. *Journal of Educational Psychology, 111*(6), 1001–1022. https://doi.org/10.1037/edu0000334
Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26*, 125–141. https://doi.org/10.1016/S0079-7421(08)60053-5
Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257. https://doi.org/10.3102/01623737026003237
Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, 71. https://doi.org/10.1136/bmj.n71
Pellegrini, M., Lake, C., Neitzel, A., & Slavin, R. E. (2021). Effective programs in elementary mathematics: A meta-analysis. *AERA Open, 7*(1), 1–29. https://doi.org/10.1177/2332858420986211
Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system: Manual K-3*. Paul H. Brookes.
Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta- analysis in a systematic review. *Review of Educational Research, 90*(1), 24–46. https://doi.org/10.3102/0034654319877153
Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research, 86*(1), 207–236. https://doi.org/10.3102/0034654315582067
Pustejovsky, J. E. (2023). clubSandwich: Cluster-robust (Sandwich) variance estimators with small-sample corrections [Computer software] Version R package 0.5.10. https://cran.r-project.org/web/packages/clubSandwich/index.html.
Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science, 23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3
Rienzo, C., Rolfe, H., & Wilkinson, D. (2015). *Changing mindsets: Evaluation report and executive summary*. *Education Endowment Foundation*.
Savage, R., Abrami, P. C., Piquette, N., Wood, E., Deleveaux, G., Sanghera-Sidhu, S., & Burgos, G. (2013). A (Pan-Canadian) cluster randomized control effectiveness trial of the ABRACADABRA web-based literacy program. *Journal of Educational Psychology, 105*(2), 310–328. https://doi.org/10.1037/a0031025
Schunk, D. H. (2019). *Learning theories: An educational perspective*. Pearson.
Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. https://doi.org/10.3102/0013189X015002004
Simpson, A. (2018). Princesses are bigger than elephants: Effect size as a category error in evidence-based education. *British Educational Research Journal, 44*(5), 897–913. https://doi.org/10.1002/berj.3474
Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., Van Herwegen, J., & Anders, J. (2023). Effective teacher professional development: New theory and a meta-analytic test. *Review of Educational Research, 0*(0). https://doi.org/10.3102/00346543231217480
Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis, 31*(4), 500–506. https://doi.org/10.3102/0162373709352369
Taylor, J. A., Pigott, T., & Williams, R. (2022). Promoting knowledge accumulation about intervention effects: Exploring strategies for standardizing statistical approaches and effect size reporting. *Educational Researcher, 51*(1), 72–80. https://doi.org/10.3102/0013189X211051319
Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development: Best evidence synthesis iteration* [BES]. *New Zealand Ministry of Education*.
Torgerson, D. J. (2001). Contamination in trials: Is cluster randomisation the answer? *BMJ, 322*(7282), 355–357. https://doi.org/10.1136/bmj.322.7282.355
van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research, 49*(2), 127–152. https://doi.org/10.1080/00131880701369651
van de Grift, W. (2010). *Ontwikkeling in de beroepsvaardigheden van leraren [Inaugural lecture]*. Rijksuniversiteit Groningen.
van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*, 271–296. https://doi.org/10.1007/s10648-010-9127-6
van Ewijk, R., & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review, 5*(2), 134–150. https://doi.org/10.1016/j.edurev.2010.02.001
van Geel, M., Keuning, T., Frèrejean, J., Dolmans, D., van Merriënboer, J., & Visscher, A. (2019). Capturing the complexity of differentiated instruction. *School Effectiveness and School Improvement, 30*(1), 51–67. https://doi.org/10.1080/09243453.2018.1539013
van Merriënboer, J. G., & Kirschner, P. A. (2018). *Ten steps to complex learning: A systematic approach to four-component instructional design*. Routledge.
van Veen, K., Zwart, R., & Meirink, J. (2012). What makes teacher professional development effective? A literature review. In M. Kooy, & K. van Veen (Eds.), *Teacher learning that matters* (pp. 23–41). Routledge.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Visscher, A. J. (2017). *Gericht ontwikkelen van leerkrachtkwaliteiten [developing teacher qualities deliberately] [inaugural lecture]*. University of Twente.

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0.* U.S. Department of Education, Institute of Education Sciences. *National Center for Education Evaluation and Regional Assistance (NCEE)*.

Wiliam, D. (2016). Teacher learning: The key to improving the world. Retrieved on January 31, 2023 from http://dylanwiliam.org/Dylan_Wiliams_website/Presentations.html.

Williams, R., Citkowicz, M., Miller, D. I., Lindsay, J., & Walters, K. (2022). Heterogeneity in mathematics intervention effects: Evidence from a meta-analysis of 191 randomized experiments. *Journal of Research on Educational Effectiveness, 15*(3), 584–634. https://doi.org/10.1080/19345747.2021.2009072

Wolf, B., & Harbatkin, E. (2023). Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type. *Journal of Research on Educational Effectiveness, 16*(1), 134–161. https://doi.org/10.1080/19345747.2022.2071364

Wolf, B., Latham, G., Armstrong, C., Ross, S., Laurenzano, M., Daniels, C., Eisenger, J., & Reilly, J. (2018). *English language and literacy acquisition - Validation i3 Evaluation (Valid 22) final report.* Center for Research and Reform in Education. https://jscholarship.library.jhu.edu/handle/1774.2/62371.

Wolf, R., Morrison, J., Inns, A., Slavin, R., & Risman, K. (2020). Average effect sizes in developer-commissioned and independent evaluations. *Journal of Research on Educational Effectiveness, 13*(2), 428–447. https://doi.org/10.1080/19345747.2020.1726537

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007–No. 033). *U.S. Department of education, Institute of education Sciences, National Center for education evaluation and regional assistance, regional educational laboratory Southwest.* http://ies.ed.gov/ncee/edlabs.

Zimmerman, B. J., & Schunk, D. H. (Eds.). (2011). *Handbook of self-regulation of learning and performance.* Routledge.