

Evidence Portal: The NIoT Evidence Toolkit Technical report (February 2026)

Contents

Introduction	2
NIoT Evidence Toolkit methodology	3
Strand development.....	4
What do we mean by impact?	6
Metrics	6
Impact on teachers – individual reviews	7
Impact on teachers – strand level	7
Impact on pupils – individual reviews.....	9
Impact on pupils – strand level.....	9
Reporting quality – individual reviews.....	10
Strength of evidence – strand level	11
Effective use of the Toolkit	14
Limitations	15
Toolkit next steps	15
References	17
Appendices	19

Introduction

Our vision is for a school system that nurtures the talents of teachers and leaders at all stages of their careers, so they can provide children with the excellent education they deserve. But we know the quality of professional learning across England varies greatly, and all too often colleagues who could benefit from great support miss out, and so do their pupils. That is why we have created the NIoT Evidence Portal.

The NIoT Evidence Portal has been created to provide access to the best available evidence on professional learning. A core aspect of our portal is the NIoT Evidence Toolkit, an interactive online resource that distils findings from high quality evidence on a wide range of teacher and leader professional learning approaches.

The NIoT Evidence Toolkit enables those involved in supporting the professional learning of others to make evidence-informed decisions. This includes colleagues in a range of roles, from senior leaders with a professional learning responsibility, to middle leaders responsible for specific subjects or phases, as well as mentors, and all who contribute to developing colleagues informally.

The Toolkit provides concise summaries on a range of professional learning approaches, based on systematic evidence from around the world. Each strand includes insights from experienced teachers and leaders, sharing examples of evidence-informed professional learning from a range of settings. Practitioner insights illustrate the application of evidence in the design and delivery of professional learning, enabling colleagues to learn from great practice and experts to support system-wide improvement.

The Evidence Toolkit was initially launched in beta in October 2025 to provide an opportunity for users to share their feedback and help shape its ongoing development. This technical report details the NIoT Evidence Toolkit development as of February 2026 and will be updated as necessary to reflect any further improvements. We welcome continued feedback on all aspects of the Toolkit (<https://evidenceportal.niot.org.uk/contact/>), including our methodology, as we do not assume to have all the answers. Our hope is that by working together, we can help improve professional learning for everyone.

NIoT Evidence Toolkit methodology

This section provides a detailed account of the design and development of the Toolkit. It explains the 'how' behind the Toolkit, ensuring transparency, reproducibility and demonstrating the rigour of the process.

The Toolkit draws on evidence from a [meta systematic mapping review](#) which was conducted to explore the nature and scope of evidence syntheses related to initial teacher education (ITE) and continued professional development (CPD). The meta review drew on evidence syntheses—summaries of literature that have been done systematically (hereafter, reviews)—sourced from databases including the Web of Science, Scopus, ProQuest, EBSCOHost, Google Scholar, and OpenAlex). The search string specifically targeted terms related to teacher professional development, education sectors, and various evidence synthesis types (see Appendix 1). Comprehensive searches were conducted between January 2024 to July 2025, for literature published between 1 January 2013 to 31 July 2025. Following full text screening within evidence synthesis software EPPI Reviewer (Thomas et al., 2023), additional sources were identified through citation and bibliography searching, bidirectional checking of citations and recommendations within the OpenAlex platform (Priem et al., 2022), and manual searches of relevant organisational websites such as the Education Endowment Foundation and Campbell Collaboration.

Publications were included in the meta review if they were an evidence synthesis with a method section (e.g., systematic review, scoping review, meta-analysis; for an overview of evidence synthesis types, see Sutton et al. 2019), focused on initial teacher education, in-service teachers' continuing professional development, or both, at all educational levels (early years, primary, secondary, college, and higher education), and were published in English between 2013 and July 2025. A total of 442 publications met the inclusion criteria, 65.2% of which were systematic reviews.

Developing the Toolkit from a meta review helped to prevent selective inclusion of only positive or well-known reviews, and ensured that all reviews meeting predefined criteria were considered. Reviews that were rated as low or critically low in reporting quality—as evaluated using the Quality of Evidence Synthesis Assessment Tool (QuEST; Bond et al., 2024)—were not included, as these may contain methodological flaws or biases that compromise trustworthiness of the findings.

Initially, 442 eligible reviews were identified through the meta review. During the quality assessment phase, 68 reviews of low quality and 10 of critically low quality were excluded from inclusion in the Toolkit. In the identification and screening stage, additional reviews were excluded from inclusion in the Toolkit development for three reasons:

- 1). The reviews were too narrowly focused in scope to be applicable beyond its specific context.
- 2). The reviews did not offer actionable findings to inform the design, implementation, or evaluation of professional development of teachers and leaders.
- 3) The reviews did not include either pupil or teacher impact outcomes.

Strand development

This section details the methodology employed in the development of the strands of the Toolkit. The strands focus on different approaches or aspects of teacher professional learning.

When selecting topics for the Toolkit strands, the following questions were considered:

- Is the topic of interest to those supporting the professional learning of others?
- Is there enough high-quality evidence?
- Is there practice relevant information available in the reviews?

The strand content was co-created by colleagues with professional development leadership expertise and education researchers. The integration of expertise from both evidence synthesis colleagues and the content development team ensured the content remained closely aligned with the review findings while simultaneously being crafted into an accessible and practical summary for individuals seeking to adopt an evidence-informed strategy in professional learning.

After each strand was written, several rounds of internal and external peer review were undertaken with a wide range of colleagues, where they were asked to comment on the clarity and language used, the appropriateness of tone for the intended audience, and whether there were any additional concerns. Feedback was taken into consideration and further edits were made.

The information and findings extracted from the reviews were systematically organised in line with the following core strand elements:

Strand element	Description
<i>Strand name</i>	A description of the professional learning approach.
<i>Strand level metrics</i>	Each strand is assessed using the following metrics:

	• Impact on teachers, using <i>promising</i>, <i>mixed</i>, <i>negative</i>, and <i>not reported</i> categories (Appendix 3). • Impact on pupils, using <i>promising</i>, <i>mixed</i>, <i>negative</i>, and <i>not reported</i> categories (Appendix 5). • Strength of evidence, using <i>very strong</i>, <i>strong</i>, <i>moderate</i>, <i>weak</i> and <i>very weak</i> categories.
<i>What is it?</i>	Defines the specific professional learning approach included in the strand, explaining its meaning in an educational context.
<i>Key Findings</i>	Details the observed impacts of the approach being discussed. • Impact on teachers. This subsection describes how the approach impacts teachers' outcomes. • Impact on pupils. This subsection presents the evidence regarding the impact on pupil outcomes.
<i>How effective is the approach?</i>	Assesses the overall effectiveness of the approach, providing a general evaluation of its impact on both teachers and pupils as well as factors that influence its effectiveness (where applicable). It further addresses the limitations in the existing evidence base.
<i>How to implement it well</i>	Outlines key elements for successful implementation. It includes additional subsections listing behaviours, contextual factors, and structures to align with the Education Endowment Foundation's implementation guidance report.
<i>Barriers to effective implementation</i>	Identifies common obstacles that can hinder successful implementation, as reported in the reviews.
<i>Other considerations</i>	Presents additional factors to bear in mind for effective and sustainable implementation.

<i>Summary</i>	Aims to capture the key takeaways and most pertinent insights.
<i>Reviews</i>	Provides a table summarising the reviews used in the strand, including details on the reference, publication years of included research, main focus, number of studies included in the review, and countries covered. The table also presents the scores of each individual review for reporting quality (Appendix 6), impact on teachers (Appendix 2), and impact on pupils (Appendix 4).

Whilst an initial selection of strand topics was developed for the October 2025 beta-launch of the Toolkit, more strands have been added since and will continue to be added on a regular basis. These will be based on the evidence available and the needs and interests of the sector.

What do we mean by impact?

From a research perspective, impact can be defined in various ways. In the meta review and for the Toolkit, we employ an inclusive definition of 'impact' to refer to the usefulness and effectiveness of a PD approach that is supported by a range of quantitative and/or qualitative evidence. It is important to note that the meta review on which the Toolkit is based includes all kinds of reviews that summarise statistical and qualitative evidence. Therefore, where the word 'impact' is used in the Toolkit, we refer to any of the following:

- Correlational impact: evidence that points to a relationship or association between the PD approach and the focused outcome(s);
- Causal impact: evidence indicating the PD approach leads to the change in the focused outcome(s);
- Experiential impact: qualitative evidence that captures participants' perceptions, experiences and reported effects of the PD approach.

Both the meta review and the Evidence Toolkit were guided by external Advisory Boards throughout the course of development. The Advisory Boards included representatives from a range of organisations including the Department for Education, the Education Endowment Foundation alongside respected colleagues from schools and academic institutions. They provided guidance on the methodology, content of the strands and the usability of the resources developed.

Metrics

Metrics are used to evaluate research in two ways:

1. At the level of **individual** reviews – these appear alongside each review at the bottom of strand pages.
2. At the **strand** level – these are visible at the top of each research strand and on the Toolkit homepage, providing an aggregate assessment based on all included reviews.

Impact on teachers – individual reviews

At the **level of individual reviews**, the impact of professional learning approaches on teaching practice and teacher outcomes is assessed using four distinct metrics: positive, mixed, negative, and not reported. These categories reflect the evidence presented in each review:

- **Positive:** the review provides clear and consistent evidence of beneficial effects on teachers. This includes improvements in teachers' knowledge, skills, attitudes, and well-being, which contribute to enhanced teaching performance.
- **Mixed:** the review suggests a potential influence on teacher outcomes, but the evidence is variable or inconsistent. Some findings point to positive effects, while others are negative, inconclusive or limited in scope. Overall, the impact cannot be confidently confirmed and may depend on specific contexts or implementation factors.
- **Negative:** the review identifies consistently detrimental effects of the approach, such as diminished teaching effectiveness, increased stress, or negative impacts on professional relationships.
- **Not reported:** the review does not report teacher outcomes and the approach's impact cannot be determined.

Impact on teachers – strand level

At the **strand level**, the analysis synthesises findings across multiple reviews to evaluate the overall impact on teachers for each professional learning approach. Each category was assigned a score, on a scale from 1-3 (as below) and an average calculated. Reviews not including teacher impacts, i.e. those with the 'not reported' rating, were not included.

Impact on teachers from individual reviews	Score assigned
Positive	3
Mixed	2
Negative	1

For example, the impact on teachers for 'Teacher collaboration' calculation looks like this:

Review	Impact on teachers	Score
Camili (2024)	Mixed	2
de Jong (2022)	Mixed	2
De la Iglesia (2024)	Mixed	2
DeLuca (2015)	Mixed	2
Fernandez-Rio (2022)	Positive	3
Fry (2025)	Mixed	2
Holmqvist (2021)	Positive	3
Jin (2024)	Mixed	2
Ramos (2022)	Mixed	2
Rytivaara (2024)	Positive	3
Vangrieken (2015)	Mixed	2
Overall score	<i>Mixed</i>	<i>2.3</i>

Broadly, the following descriptions describe each category level:

- **Promising:** the approach consistently demonstrates a clear and beneficial impact across reviews. Evidence shows meaningful and desirable changes in teaching practice and teacher outcomes, with a strong and repeatable pattern.
- **Mixed:** the impact varies across contexts or reviews. While some findings indicate positive effects, others may show negative, limited or conflicting results. The effectiveness of the approach may depend on specific conditions such as implementation quality, school setting, or teacher characteristics.
- **Negative:** the approach is associated with consistently detrimental effects. Evidence points to undesirable changes in teaching practice or teacher outcomes, such as reduced effectiveness or increased stress.
- **Not reported:** there is insufficient evidence to determine the overall impact. This may result from a lack of relevant meta-analyses or limited data within other types of reviews that prevents clear conclusions.

Impact on pupils – individual reviews

To understand how professional learning approaches affect pupils, each individual review was assessed for its reported impact on pupil outcomes. These outcomes include academic achievement, personal development, and overall well-being. The findings are categorised into four evidence-based metrics: positive, mixed, negative, and not reported.

- **Positive:** the review presents clear evidence of a beneficial effect on pupils. This includes gains in their academic performance, improved attitudes toward learning, enhanced skills, or better emotional and social well-being.
- **Mixed:** the review suggests a potential influence on pupil outcomes, but the evidence is variable or inconsistent. Some findings point to positive effects, while others are negative, inconclusive or limited in scope. Overall, the impact cannot be confidently confirmed and may depend on specific contexts or implementation factors.
- **Negative:** the review identifies consistently detrimental or unintended consequences for pupils. These may include reduced engagement, lower academic performance, or negative effects on emotional or social development.
- **Not reported:** the review does not report pupil outcomes and the approach's impact cannot be determined.

Impact on pupils – strand level

At the **strand level**, the analysis synthesises findings across multiple reviews to evaluate the overall impact on pupils for each professional learning approach. Each category was assigned a score, on a scale from 1-3 (as below) and an average calculated. Reviews not including pupil impacts, i.e. those with the 'not reported' rating, were not included.

Impact on pupils from individual reviews	Score assigned
Positive	3
Mixed	2
Negative	1

For example, the impact on pupils for 'Data-informed professional learning' calculation looks like this:

Review	Impact on pupils	Score
Ansyari (2022)	Positive	3
Salas-Pilco (2022)	Positive	3
Shanahan (2025)	Mixed	2
Overall score	Promising	<i>2.7</i>

Broadly, the following descriptions describe each category level:

- **Promising:** the approach shows a consistent and positive influence on pupil outcomes across a range of reviews. The evidence points to meaningful and desirable changes, such as improved academic achievement, motivation, or well-being, with a strong and repeatable pattern.
- **Mixed:** the approach shows varied or inconsistent outcomes. Some reviews report positive effects, while others show negative, minimal or conflicting results. This suggests that the impact may depend on specific factors such as context, implementation, or student demographics.
- **Negative:** the approach is associated with consistently negative outcomes for pupils. Evidence indicates undesirable effects, such as reduced learning gains or negative emotional or behavioural impacts.
- **Not reported:** there is insufficient evidence to determine the overall impact on pupils. This may result from a lack of relevant meta-analyses or limited data that prevents clear conclusions.

Reporting quality – individual reviews

At the individual review level, the reporting quality was evaluated using the Quality of Evidence Synthesis Assessment Tool (QuEST; Bond et al., 2024), adapted from AMSTAR 2 (Shea et al., 2017) and the DARE criteria (Centre for Reviews and Dissemination, 1995; Lai & Bower, 2020). Each review received an overall score out of 10 (see Appendix 6) and was categorised as critically low (0-2.5), low (3-4.5), medium (5-7), high (7.5-8.5), or excellent (9-10), consistent with prior reviews (e.g., Urdaneta-Ponte et al., 2021).

Low and critically low-quality reviews were excluded from the development of the Toolkit to ensure the reliability and credibility of the information provided. Reviews assessed as low or critically low can suffer from methodological flaws, biases, or inconsistent findings, which could lead to misleading representation of the evidence (Lunny et al., 2024).

Strength of evidence – strand level

This section outlines the methodological approach used to systematically assess the strength of the evidence for each strand in the Toolkit. The purpose of this method is to provide a transparent and objective score that reflects both the quality and number of the underlying reviews within each strand, whilst also considering the strength and quality of the evidence in the primary studies within the reviews where possible.

To calculate the strength of evidence of a strand, three factors were taken into consideration. Firstly, the average reporting quality of the included reviews, secondly the number of reviews in the strand, and finally the robustness of primary studies within the reviews. By considering these three factors, we have attempted to systematically evaluate the strength of evidence for each strand, thereby gaining a more nuanced overview of the evidence.

Factor one: Average reporting quality

An average reporting quality score was calculated using all the reporting quality scores with each strand. These average scores then equated to a raw score, which was then weighted from lowest to highest (as below; 0-1 scale) to be used as one of three factors in the strand strength of evidence assessment.

Average reporting quality score	Raw Score	Weighted Score
0-2.5	1	0.0
3-4.5	2	0.25
5-7	3	0.5
7.5-8.5	4	0.75
9-10	5	1.0

Factor two: Number of reviews

The number of reviews included in a strand were categorised into <5, 5–10, and >10 based on methodological principles that balanced the breadth of evidence with the reliability of synthesis. This approach follows methodological guidance from frameworks such as 'A Measurement Tool to Assess Systematic Reviews' (AMSTAR 2; Shea et al., 2017) and the Cochrane Handbook for Systematic Reviews of Interventions (Higgins et al., 2024).

More than 10 reviews may indicate a substantial evidence base. A larger pool of reviews enhances the robustness of synthesis, allowing for more comprehensive analysis, detection of heterogeneity, and assessment of publication bias (Shea et al., 2017).

Between 5 and 10 reviews may provide a moderate evidence base. This range allows for comparative synthesis and trend identification across multiple review findings. It supports meta-analytical techniques, enabling a more reliable assessment of consistency and quality. Fewer than 5 reviews may represent a limited evidence base. When only a small number of reviews are available, it becomes difficult to assess consistency across findings, and the risk of bias or overinterpretation increases. According to AMSTAR 2 (Shea et al., 2017), such limited synthesis contributes to imprecision and lower confidence in conclusions. The Cochrane Handbook (Higgins et al., 2024) also notes that sparse review data may not support meaningful aggregation or generalisability.

In order to contribute to the overall strength of evidence rating, the number of reviews were assigned a weighted score as follows:

Number of reviews included in the strand	Weighted score
<5	0.3
5-10	0.6
>10	1

Factor Three: Robustness of primary study evidence

In order to consider the strength of evidence at the strand level, it was also important to consider the robustness of the primary studies synthesised within the reviews. The robustness was assessed taking the following factors into consideration:

- The number of primary studies on which the evidence of impact on teachers and/or pupils is based; and
- The methodological quality of primary studies that report evidence of impact.

Each review was given a score from 1-5 for *both* the robustness of primary study evidence for impact on teachers and the robustness of evidence for impact on pupils, with an average score for the strand used in the final strength of evidence calculation. However, if a review received an initial score of 'Strong' as it had many studies (e.g. 45), but the methodological quality of those primary studies was reported as low, then the robustness rating would be lowered one (in this case to 'Moderate').

Robustness of primary study evidence	Score	Description
---	--------------	--------------------

Very strong	5	A very large amount of evidence (65+ studies).
Strong	4	A large amount of evidence (41-64 studies).
Moderate	3	A reasonable amount of evidence (26-40 primary studies).
Low	2	A limited amount of evidence (11-25 primary studies).
Very low	1	A very limited amount of evidence (10 or fewer primary studies).

Strength of evidence calculation

When calculating the overall strength of evidence of a strand, the three factors are taken into consideration and weighted as follows:

- Average reporting quality of reviews (40%)
- Number of reviews in the strand (20%)
- Robustness of primary study evidence (40%)

The number of reviews was only weighted at 20% to account for any duplication of primary studies between reviews, and because the reporting quality and robustness of evidence are of higher importance.

A final score was then calculated from these values (on a scale from 0-1) and an overall strength of evidence level was applied as follows:

Score	Strength of evidence	Description
0.85 – 1.0	Very strong	The reporting quality across reviews is excellent, and the overall robustness of the evidence base is very strong.
0.75 – 0.84	Strong	The reporting quality across reviews is high, and the overall robustness of the evidence base is moderate to very strong, although some minor limitations might be present.

0.5 – 0.74	Moderate	The reviews are well-reported, with moderate confidence in the evidence base due to limitations affecting robustness.
0.3 – 0.49	Weak	The reporting quality across reviews is good, but the evidence base is limited and/or the robustness of evidence is weak.
0.0 – 0.29	Very weak	The reporting quality across reviews is good, but the evidence base is very limited, and the robustness of evidence is very weak.

For example, the Inclusion strand final strength of evidence level calculation looks like this:

Average reporting quality	Weighted score (0-1)	Number of reviews	Weighted score (0-1)	Robustness of primary studies	Weighted score (0-1)	Final score	Final strength of evidence
Medium (3)	0.5	>10	1	Weak	0.3	<i>0.52</i>	<i>Moderate</i>

Effective use of the Toolkit

The NIoT Evidence Toolkit is designed to make global evidence on teacher training and development accessible and practical. It is hoped that it will be a key resource for anyone involved in designing or delivering professional learning, from PD leads and mentors to school leaders and policymakers. Its purpose is to help you make evidence-informed decisions that directly impact teacher effectiveness.

Key principles for effective use:

Blending evidence with practical expertise

The Toolkit combines synthesised evidence with practical insights from practicing teachers and leaders. The Toolkit can be used to ground professional learning in a strong evidence base, while also bringing in the practical, tacit knowledge of experienced teacher educators. This ensures that the training being designed is not only research-informed but also relevant and actionable for the target audience.

Applying evidence with professional judgement

A Toolkit can provide a wealth of global evidence, but it is not a prescriptive guide. The key is to apply professional judgement, interpreting the evidence in light of one's specific

needs and unique context. This thoughtful process is what turns broad research findings into professional learning that is both relevant and impactful.

Looking beyond the stated impact

The impact metrics for an approach can hide a wide range of outcomes. The Toolkit provides details on the effectiveness of the approach in various contexts, how to implement it well, and potential barriers to effective implementation. By considering these nuances, rather than just the stated impact metrics, a more informed decision can be made. This helps ensure that the selected strategy is most likely to be effective within a specific setting.

Limitations

Whilst we believe that the NIoT Evidence Portal represents a significant advancement in evidence-informed professional learning, several limitations should be acknowledged:

Primary research. The reviews that the Toolkit synthesises are based on primary research studies for which there are limitations. Firstly, the quality of the primary research is variable and, in some cases, unknown; secondly, for a lot of the primary research the professional learning design was not the main object of the inquiry, but rather a pre-condition for studying another object of interest (see Asterhan and Lefstein, 2024).

Evidence scope and coverage. The Toolkit relies on reviews published between 2013 and 2025. Therefore, emerging practices or innovations not yet captured in secondary research will not yet be included.

Exclusion of low and critically low reviews. Although reviews rated as low or critically low were excluded to maintain methodological rigour, this may have resulted in the omission of contextually rich and practical insights from primary research studies they included.

Strength of evidence. The strength of evidence calculation is based on three weighted factors; the average reporting quality, the number of included reviews and the robustness of the primary studies contained within the reviews.

Interpretation of evidence. Senior researchers with expertise in evidence synthesis methodologies, worked alongside professional learning content developers to write the content and verify adherence to the review evidence. However, we acknowledge that there is bound to be a level of subjectivity and interpretation in the process of reporting what information/evidence supports the strands.

Toolkit next steps

The NIoT Evidence Toolkit will continue to grow as new strands are added, alongside further exemplification and additional supporting resources. It is intended that each July

a search will take place to assess whether there are additional reviews to include in the Toolkit. We welcome continued feedback on all aspects of the Toolkit (<https://evidenceportal.niot.org.uk/contact/>) as we do not assume to have all the answers. Our hope is that by working together, we can help improve professional learning for everyone.

References

- Asterhan, C. S., & Lefstein, A. (2024). The search for evidence-based features of effective teacher professional development: A critical analysis of the literature. *Professional Development in Education*, 50(1), 11–23. <https://doi.org/10.1080/19415257.2023.2283437>
- Bianchi, J., Hirt, J., Vogt, M., & Vetsch, J. (2025). Data extractions using a large language model (Elicit) and human reviewers in randomized controlled trials: A systematic comparison. *Cochrane Evidence Synthesis and Methods*, 3, e70033. <https://doi.org/10.1002/cesm.70033>
- Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta-systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21(1), Article 4. <https://doi.org/10.1186/s41239-023-00436-z>
- Bulut, S. (2009). The effect of cooperative learning techniques on pre-service teachers' self-esteem, achievement, and absences. *Suvremena Psihologija*, 12(1), 23–41. <https://hrcak.srce.hr/82655>
- Buntins, K., Bedenlier, S., Marín, V., Händel, M., & Bond, M. (2023). Methodological approaches to evidence synthesis in educational technology: A tertiary systematic mapping review. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, 54, 167–191. <https://doi.org/10.21240/mpaed/54/2023.12.20.X>
- Centre for Reviews and Dissemination. (1995). DARE criteria. In *Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews*. University of York. <https://www.crd.york.ac.uk/crdweb/ShowRecord.asp?ID=32004000332>
- Fernández-Río, J., Cecchini, J. A., & Méndez-Giménez, A. (2014). Effects of cooperative learning on perceived competence, motivation, social goals, effort, and boredom in prospective primary education teachers. *Infancia y Aprendizaje*, 37(1), 57–89. <https://doi.org/10.1080/02103702.2014.881650>
- Gander, M., & Dane, A. V. (2023). Bug-in-ear coaching in teacher professional development: A scoping review. *Review of Educational Research*, 93(4), 789–823. <https://doi.org/10.3102/00346543231123456>
- Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2024). *Cochrane handbook for systematic reviews of interventions* (version 6.5). Cochrane. <https://www.cochrane.org/handbook>

Lai, J. W. M., & Bower, M. (2020). Evaluation of technology use in education: Findings from a critical analysis of systematic literature reviews. *Journal of Computer Assisted Learning*, 36(3), 241–259. <https://doi.org/10.1111/jcal.12412>

Lunny, C., Kanji, S., Thabet, P., Haidich, A. B., Bougioukas, K. I., & Pieper, D. (2024). Assessing the methodological quality and risk of bias of systematic reviews: Primer for authors of overviews of systematic reviews. *BMJ Medicine*, 3(1). <https://doi.org/10.1136/bmjmed-2023-000604>

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv*. <https://doi.org/10.48550/arXiv.2205.01833>

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., et al. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomized or non-randomized studies of healthcare interventions, or both. *BMJ*, 358. <https://doi.org/10.1136/bmj.j4008>

Sutton, A., Clowes, M., & Booth, A. (2019). What is a systematic review? In A. Sutton, D. A. Barrett, A. Booth, & M. Clowes (Eds.), *An introduction to systematic reviews* (2nd ed., pp. 1–16). SAGE Publications.

Thomas, J., Brunton, J., & Graziosi, S. (2023). EPPI-Reviewer software for systematic reviews (version 4) [Computer software]. Social Science Research Unit, UCL Institute of Education, University College London. <https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2919>

Tran, L., Tam, D. N. H., Elshafay, A., Dang, T., Hirayama, K., & Huy, N. T. (2021). Quality assessment tools used in systematic reviews of in vitro studies: A systematic review. *BMC Medical Research Methodology*, 21(1), 101. <https://doi.org/10.1186/s12874-021-01295-w>

Urdaneta-Ponte, M. C., Mendez-Zorrilla, A., & Oleagordia-Ruiz, I. (2021). Recommendation systems for education: Systematic review. *Electronics*, 10(14), 1611. <https://doi.org/10.3390/electronics10141611>

Appendices

Appendix 1. Search string for meta review

Category	Terms
Teacher professional development	"professional development" OR "teacher training" OR "initial teacher education" OR "teacher preparation" OR "continuing professional development" OR "pre-service teacher" OR "preservice teacher" OR "in-service teacher" OR "student teacher" OR "teacher professional learning" OR "teacher professional education" OR "INSET" OR "in-service education and training"
AND	
Education sector	"higher education" OR college* OR universit* OR undergrad* OR graduat* OR postgrad* OR "K-12" OR school* OR kindergarten* OR "primary school*" OR "middle school*" OR "high school*" OR "elementary school*" OR "secondary school*" OR "nursery school*" OR "early years" OR "early childhood education"
AND	
Evidence synthesis	"systematic review" OR "scoping review" OR "narrative review" OR "meta-analysis" OR "evidence synthesis" OR "meta-review" OR "evidence map" OR "rapid review" OR "umbrella review" OR "qualitative synthesis" OR "configurative review" OR "aggregative review" OR "thematic synthesis" OR "framework synthesis" OR "mapping review" OR "meta-synthesis" OR "qualitative evidence synthesis" OR "critical review" OR "integrative review" OR "integrative synthesis" OR "narrative summary" OR "state of the art review" OR "rapid evidence assessment" OR "qualitative research synthesis" OR "qualitative metasummary" OR "meta-ethnography" OR "meta-narrative review" OR "mixed methods synthesis" OR "scoping study" OR "systematic map"

Appendix 2. Impact on teachers - individual reviews

Impact on teachers	Description / calculation
Positive	The review provides clear and consistent evidence of beneficial effects on teachers. This includes improvements in teachers' knowledge, skills, attitudes, and well-being, which contribute to enhanced teaching performance.
Mixed	The review suggests a potential influence on teacher outcomes, but the evidence is variable or inconsistent. Some findings point to positive effects, while others are negative, inconclusive or limited in scope. Overall, the impact cannot be confidently confirmed and may depend on specific contexts or implementation factors.
Negative	The review identifies consistently detrimental effects of the approach, such as diminished teaching effectiveness, increased stress, or negative impacts on professional relationships.
Not reported	The review does not report teacher outcomes and the approach's impact cannot be determined.

Appendix 3. Impact on teachers – strand level

Impact on teachers	Description / calculation
Promising	Research consistently highlights a clear and beneficial impact of the approach on teaching practice and teacher outcomes. The evidence indicates a meaningful and desirable change, showing a strong and repeatable pattern across various studies.
Mixed	Research presents inconsistent or varied findings regarding the impact of the approach on teaching practice and teacher outcomes. Some areas or contexts may show positive changes, while others indicate no significant influence, or even conflicting trends. The impact of the approach is present but not uniform, often depending on specific conditions.
Negative	Research consistently points to a detrimental impact of the approach on teaching practice and/or teacher outcomes. The evidence indicates an undesirable influence, leading to less effective or counterproductive changes.
Not reported	There is insufficient or inconclusive evidence to draw definitive conclusions about the impact of the approach due to a lack of meta-analyses addressing the specific outcome or existing meta-analyses having too little data for clear findings to be established.

Appendix 4. Impact on pupils - individual reviews

Impact on pupils	Description
Positive	The review presents clear evidence of a beneficial effect on pupils. This includes gains in their academic performance, improved attitudes toward learning, enhanced skills, or better emotional and social well-being.
Mixed	The review suggests a potential influence on pupil outcomes, but the evidence is variable or inconsistent. Some findings point to positive effects, while others are negative, inconclusive or limited in scope. Overall, the impact cannot be confidently confirmed and may depend on specific contexts or implementation factors.
Negative	The review identifies consistently detrimental or unintended consequences for pupils. These may include reduced engagement, lower academic performance, or negative effects on emotional or social development.
Not reported	The review does not report pupil outcomes and the approach's impact cannot be determined.

Appendix 5. Impact on pupils – strand level

Impact on pupils	Description
Promising	The approach shows a consistent and positive influence on pupil outcomes across a range of reviews. The evidence points to meaningful and desirable changes, such as improved academic achievement, motivation, or well-being, with a strong and repeatable pattern.
Mixed	The approach shows varied or inconsistent outcomes. Some reviews report positive effects, while others show negative, minimal or conflicting results. This suggests that the impact may depend on specific factors such as context, implementation, or student demographics.
Negative	The approach is associated with consistently negative outcomes for pupils. Evidence indicates undesirable effects, such as reduced learning gains or negative emotional or behavioural impacts.
Not reported	There is insufficient evidence to determine the overall impact on pupils. This may result from a lack of relevant meta-analyses or limited data that prevents clear conclusions.

Appendix 6. Reporting Quality Assessment (QuEST; Bond et al., 2024)

Reporting quality (QuEST)	Description / calculation (out of 10 QuEST criteria)	Notes
Excellent	9-10	The scores indicate a very robust and methodologically sound review.
High	7.5-8.5	The scores signify a strong and well-conducted review with minor limitations.
Medium	5-7	The scores suggest a review with some methodological limitations but still providing valuable insights.
Low	3-4.5	The scores suggest a review with significant methodological weaknesses.
Critically low	0-2.5	The scores point to a review with severe and critical methodological flaws.

Note: The version of QuEST applied in this review comprised 10 items, scored as follows: Yes = 1, Partly = 0.5, and No = 0, in line with previous studies (e.g., Buntins et al., 2023; Tran et al., 2021). Where certain synthesis types (e.g., scoping or mapping reviews; Sutton et al., 2019) did not require a specific quality criterion, items were marked 'Not Applicable' and assigned a score of 1. To ensure consistency, five initial papers were jointly assessed by all review team members. Subsequently, 30 reviews were screened by two reviewers, achieving 90% agreement. The remaining reviews were independently assessed by two reviewers, with a third reviewer appraising five randomly selected items from each quality category ('critically low', 'low', 'high', and 'excellent'). Only one discrepancy in scoring was identified, which was resolved through discussion.

Criterion	Score	Interpretation
RQs	Yes	RQs, aims or objectives are explicitly and clearly defined.
	Partly	Some mention of objectives or aims are alluded to.
	No	No RQs, aims or objectives are identifiable.
Inclusion/exclusion	Yes	The criteria used are explicitly defined in the paper.
	Partly	The criteria are implicit.
	No	The criteria are not defined and cannot be readily inferred.
Publication years	Yes	The publication years are clearly stated

	Partly	The publication years state from a year OR until a year.
	No	The publication years are not defined at all.
Search coverage	Yes	4 or more digital libraries searched and included additional search strategies (e.g., snowballing) OR identified and referenced all pertinent journals.
	Partly	3 or 4 digital libraries searched with no extra search strategies OR searched a defined but restricted set of journals and conference proceedings.
	No	2 digital libraries searched or an extremely restricted set of journals.
Search string	Yes	The search string was reported in full and is replicable.
	Partly	Examples of keywords only were given.
	No	The search terms were not provided in any form.
Inter-rater reliability	Yes	An inter-rater reliability value is reported (e.g., Cohen's kappa).
	Partly	Mention is made of how disagreements were reconciled.
	No	No inter-rater reliability is mentioned.
Data extraction	Yes	The full coding scheme was provided.
	Partly	Examples are provided, but not the full list.
	No	The coding scheme was not provided at all.
Quality assessment	Yes	Explicitly defined quality criteria extracted for each study.
	Partly	The research question involved quality issues that are addressed by the study.
	No	No explicit quality assessment of individual papers has been attempted.
Study description	Yes	Information is presented about each paper
	Partly	Only summary information is presented about individual papers.
	No	The results for individual studies are not specified
Review limitations	Yes	Yes, there is a specific identifiable limitations section.
	Partly	There is some mention of limitations
	No	There is no limitations section or reflection on limitations