Education
Endowment
Foundation

What are the Characteristics of Effective Teacher Professional Development? A Systematic Review and Meta-analysis

October 2021

Sam Sims (UCL)
Harry Fletcher-Wood (Ambition Institute)
Alison O'Mara-Eves (UCL)
Sarah Cottingham (Ambition Institute)
Claire Stansfield (UCL)
Jo Van Herwegen (UCL)
Jake Anders (UCL)

UCL

Ambition
Institute

Please cite this report as follows: Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Van Herwegen, J., Anders, J. (2021). *What are the Characteristics of Teacher Professional Development that Increase Pupil Achievement? A systematic review and meta-analysis.* London: Education Endowment Foundation. The report is available from: https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/teacher-professional-development-characteristics

The EEF Guidance Report ***Effective Professional Development*** is available at: https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/effective-professional-development

Contents

Executive summary

Teachers have an important influence on pupils' academic progress, yet the quality of teaching varies widely (Lee, 2018). Policymakers, school leaders, and teacher educators therefore face the challenge of designing and commissioning professional development (PD) to help all their teachers become as effective as the best teachers.

In the last two decades, a large number of experimental evaluations have tested the impact of different approaches to teacher PD. Evidence has accumulated showing that PD does indeed improve teaching and pupil learning (Fletcher-Wood & Zuccollo, 2019; Kennedy, 2016a; Lynch et al., 2019). However, impact varies widely, which raises the question of what—if anything—differentiates more effective PD from less effective PD. The objective of this review is to identify the characteristics of more effective PD.

Several reviews have attempted to answer this question (Desimone, 2009; Kennedy, 2016a; Timperley et al., 2007; Walter & Briggs, 2012; Wei et al., 2009; Yoon et al., 2007). However, to date, these reviews have either been inconclusive or have had important methodological limitations (Sims & Fletcher-Wood, 2021). This report presents the results of an updated systematic review and meta-analysis, employing novel theory and methods to provide new insights on this important question. It includes the results of analyses pre-registered in our published protocol (Sims et al., 2021).

We systematically searched ten scholarly databases and other resources covering the field of education for evaluations of PD programmes published between 2002 and 2020, screening 3,140 abstracts and reviewing 347 papers. We included papers if they reported an experimental evaluation of PD, delivered to qualified teachers of pupils aged 3–18, and reported a standardised test-score outcome. Ultimately, 121 papers were eligible for inclusion, although 17 of these did not report sufficient data to calculate effect sizes. Our final dataset included 104 papers reporting 205 effects.

For each of these 104 papers, we coded the characteristics of the study, indicators of study quality, and the impact of the PD programme. The majority of studies were conducted in the USA (70%), with just under a quarter in the UK (24%), and a small number elsewhere (6%). Half the studies were conducted in primary/elementary schools (50%) with the rest evenly split between early years/pre-kindergarten and middle/secondary schools. Half the studies focused on literacy (50%).

Crucially, we also coded whether each PD programme incorporated a set of 14 'mechanisms'—empirically-evidenced general principles about how people learn and change their practice. We grouped these mechanisms around four purposes of PD: helping teachers gain new *insights* (I), pursuing new *goal*-directed behaviours (G), acquiring new skills or *techniques* (T), and embedding these changes in their *practice* (P). We refer to this as the IGTP model. We pre-registered our overall IGTP framework, as well as our approach to identifying the mechanisms within it (Sims et al., 2021). We only included a mechanism in our framework if we found empirical evidence for its effects in multiple settings outside of teacher PD (for example, in a lab experiment and in a health psychology study). We theorised that teacher PD programmes incorporating more of these domain general mechanisms would themselves be more effective. We also pre-registered the hypothesis that PD is more likely to be effective if it incorporates at least one mechanism addressing each of the I, G, T and P purposes—we refer to this as a 'balanced design'.

We found an average impact of teacher PD on pupil standardised test scores of .05 standard deviations (95% confidence interval .03-.07). This is equivalent to approximately one month of additional pupil progress.[1] However, the impact of PD varies widely across different PD programmes and our primary interest in this research is in explaining this variation. In line with our expectations, our results show that PD programmes that include more mechanisms tend to have a higher impact on pupil test scores. Programmes incorporating zero mechanisms have effect sizes around zero. By contrast, programmes incorporating (almost) all 14 of our mechanisms have effect sizes around .17—equivalent to approximately two months of additional pupil progress. These results hold true within the subset of PD focused on developing formative assessment and inquiry-based teaching; though not among those focused on data-driven instruction.

In line with our hypothesis, balanced PD designs (incorporating one mechanism addressing each of insight (I), goals (G), techniques (T), and practice (P)) have larger average effects. Balanced designs have an average impact of .15 standard deviations; while for non-balanced designs, the equivalent figure is .05. However, there is considerable statistical uncertainty around our meta-analytic estimates for the impact of balanced designs, with a 95% confidence interval of .02-.27. The *p*-value on the comparison between balance and imbalanced design is .22. Hence, more trials of balanced PD designs would be needed to provide a precise test of whether balanced designs are superior or not. Our qualitative comparative analysis showed that balanced designs were more likely to show impact above various thresholds for substantively important impact, relative to imbalanced designs.

These findings should be interpreted in light of the limitations of this research. For example, while the set of original research papers we analysed all provide causal impact estimates at the within-study level, our analysis looks at the *associations* between PD characteristics and impact at the between-study level. Nevertheless, the independent causal evidence that we present for each of our mechanisms suggests that the associations that we observe reflect an underlying causal relationship. An important additional caveat to our results is that we find much lower effect sizes among pre-registered experiments (in which methods are specified prior to data collection) and experiments with higher levels of attrition (missing outcome data). This suggests that effect size estimates from some experiments may be inflated and should be interpreted with caution.

We also reviewed the implementation and process evaluations of a subset of studies to identify factors supporting implementation. We found that, in practice, PD is often implemented with low fidelity and adapted to suit local needs and constraints. Our thematic analysis suggests that PD is more likely to be implemented with fidelity when interventions are aligned with schools' needs and existing practices, and when planned around the limited time available to teachers.

In sum, our results suggest that policymakers, school leaders, and teacher educators should favour PD designs that incorporate more of the mechanisms we set out in our theoretical framework. Appendix 5 sets out this framework in full, providing examples, explanations, and references for each of the 14 mechanisms. Our findings also suggest that, on the balance of probabilities, educators should prefer balanced PD designs over imbalanced designs. However, this should be kept under review as there is considerable statistical uncertainty in our estimates. Appendix 8 contains examples of balanced PD designs.

[1] https://educationendowmentfoundation.org.uk/help/projects/the-eefs-months-progress-measure/

Part 1: Background, conceptual framework and methods

1.1 Background and rationale

On average, teachers spend 10.5 days per year attending courses, workshops, conferences, seminars, observation visits, or other in-service training (Sellen, 2016). The rationale for this substantial investment in professional development (PD) is clear: meta-analyses find that teacher PD programmes tend to improve pupil academic achievement (Fletcher-Wood & Zuccollo, 2019; Lynch et al., 2019). How this PD should be designed is, however, somewhat less clear. While research has identified some programmes or interventions for which there is persuasive evidence of impact on pupil achievement (for example, My Teaching Partner; Allen et al., 2011; 2015), many schools do not have access to such programmes, due to cost or location. School leaders and PD designers instead need to know which characteristics of PD matter to help them design or commission effective PD (Hill et al., 2013).

Several reviews have attempted to identify the characteristics of effective PD (Desimone, 2009; Kennedy, 2016a; Timperley et al., 2007; Walter & Briggs, 2012; Wei et al., 2009; Yoon et al., 2007). Indeed, these reviews have themselves been summarised in two meta-reviews (Cordingley et al., 2015; Dunst et al., 2015). However, these (meta-)reviews have either been inconclusive or have important methodological limitations (Sims & Fletcher-Wood, 2021). In particular, existing reviews have no clear way of distinguishing causally redundant components of interventions from the 'active ingredients' that contribute to improved teaching and pupil learning (Sims & Fletcher-Wood, 2021). A new synthesis of this literature, using improved methods, is therefore required.

Previous reviews have not used a consistent definition of PD. Indeed, several proceed without offering any explicit definition (Lynch et al., 2019; Cordingley et al., 2015; Dunst et al., 2015; Kennedy, 2016a), with one stating only that professional development is 'hard to define by aggregation and generalities' (Opfer & Pedder, 2011, p.379). A recent review adopted a broad multi-part definition, which can be summarised as: facilitated learning opportunities for qualified professionals that aim to enhance the professionals' knowledge and skills in ways that are relevant for application in practice, that is, to serve ultimate beneficiaries (students) (Filges et al., 2019). While this is a useful starting point, the breadth of this definition is problematic for our purposes. For example, it would seem to include programmes that introduce some new educational technology and incorporate a short training session to familiarise teachers with the software (e.g. Campuzano et al., 2009). Similarly, it would appear to include so-called 'out of the box' curriculum packages, which are accompanied by token training to introduce the teacher to the new curriculum materials (e.g. Miller et al., 2007). Intuitively, we believe that both researchers and teachers would recognise these as educational technology and curriculum programmes, rather than PD.

Our approach, therefore, builds on this broad definition, while also seeking to refine it slightly. **We define teacher PD as structured, facilitated activity for teachers intended to increase their teaching ability.** The focus on *teaching* ability is intended to include a broad range of skills including instruction, classroom management, assessment, and lesson planning. At the same time, it is intended to exclude educational technology programmes with a token training element (e.g. Campuzano et al., 2009). The focus on teaching *ability*, rather than merely knowledge, is intended to distinguish PD from new curriculum programmes with a token training element (e.g. Miller et al., 2007). Furthermore, this will help distinguish PD from activity focused on simply providing teachers with general updates about school business. We acknowledge that our definition will still require a degree of inference on the part of the reviewers, but we submit that this definition is tighter and more transparent than those used in previous reviews.

1.2 Initial conceptual framework

The design of PD can be thought about at three different levels: programmes, forms, and mechanisms (Sims & Fletcher-Wood, 2019). PD *programmes* are specific sets of activities and materials that have their own identity and tend to be located in, or associated with, specific people or institutions. In well-established programmes, the activities are sometimes codified in a programme manual and the materials can sometimes be acquired off-the-shelf as part of a resource pack. An example of a PD programme is Dialogic Teaching, which is a highly codified programme developed by Robin Alexander at Cambridge University. Programmes are generally suitable for evaluation using (quasi-)experimental methods, which can provide evidence on whether a PD programme is effective. However, as alluded to in the previous section, schools may not have access to such effective programmes.

PD *forms* are a type or category of PD, specified at a higher level of abstraction than a programme. Forms are defined by a set of characteristics: typical, identifying features. Conditional on having these characteristics, forms can accommodate variation in the specific materials and activities involved and are not uniquely associated with specific people or institutions. One example of a PD form is lesson study (Fernandez & Yoshida, 2012). Evidence that a form of PD is effective might come from a meta-analysis that looks at whether programmes that exemplify that form are effective on average. Schools might be more able to access PD of a certain form because (unlike a PD programme) there may be multiple providers, or the school can develop in-house provision. However, the latter requires that schools understand the active ingredients of that form, so that they can implement it effectively.

PD *mechanisms* are the 'entities and activities organized in such a way that they are responsible for the phenomenon' (Illari & Williamson, 2012, p.14). The 'phenomenon' of interest here is improved teaching and learning. The 'entities and activities' are the design of the PD programmes. It follows from this definition that a genuine mechanism could not be removed from some PD without changing the impact of the PD. In that sense, mechanisms are the basic building blocks of PD. One example of a PD mechanism might be the rehearsal of a new technique in a realistic classroom setting (Hobbiss et al., 2021). To establish that something is a genuine PD mechanism requires 'evidence of mechanism', which we argue comes 'from basic [causal, empirical] research describing fundamental characteristics of human motivation or learning, which hold across diverse contexts' (Sims & Fletcher Wood, 2021, p.54). The latter part of the definition is intended to stress that evidence of mechanism requires that that the mechanisms apply beyond teacher PD. We insist on this criteria for evidence of mechanism because it is the only way that we can be confident that a given entity and activity in a PD programme is causally responsible for the improved teaching and learning that results from PD. If we have evidence that a given entity/activity is causally active in changing knowledge, motivation, or practice across multiple settings, then we are much better warranted in inferring that same entity is also causally active in teacher PD.

It follows from the above that PD forms and programmes can be defined based on the interlocking set of mechanisms of which they are comprised. For example, the Content Focused Coaching programme (Matsumura et al., 2010) is composed of several mechanisms, including: providing an observable example of a technique, providing communication from a credible source in favour of that technique, and prompting rehearsal of that specific technique (Sims & Fletcher-Wood, 2019). *The aim of this systematic review is to identify effective forms of PD, and their constituent mechanisms.*

Having conceptualised PD in this way, it remains to be theorised how PD is linked with pupil achievement. Several logic models have been put forward in the academic literature (Boylan & Demack, 2018; Desimone, 2009; Kraft et al., 2019), which we have synthesised into the logic model in Figure 1 below. Thus, we conceive of PD as having proximal effects on teachers' insight, goals (for their teaching), techniques/skills, and habits of practice. These proximal effects affect teaching quality, which then affect pupil achievement. In line with Desimone's (2009) model, we also acknowledge that the school environment is likely to interact with each of these steps. The review of implementation (see Part 5 of this report) and the accompanying review of practice (Higton et al., 2021) are intended to probe this part of the logic model. It should be noted, however, that this review does not aim to test each of the linkages set out in the logic model in Figure 1. Rather, the logic model is intended to make transparent some of our background assumptions about the broader system in which PD is embedded, in order to inform our review of theory and act as a useful point of reference for developing our coding frame.

Our primary focus in this review is on the two boxes at the left of Figure 1 ('PD forms' and 'Changes within the teacher') and their relationship with the far-right box ('Pupil performance (achievement)'). In particular, we hypothesise that effective PD has to incorporate a set of mechanisms that are able to achieve four purposes (Goodrich, 2021). These are:

1. To instil new evidence-based **insights**, such as: working memory is composed of separate visual-spatial and phonological systems, each of which has limited capacity (Baddeley & Hitch, 1974).
2. To motivate **goal-directed behaviour** around new ways of teaching that make use of this insight. For example, a teacher resolves to limit the cognitive load their teaching places on either the visual-spatial or the phonological system within working memory.
3. To develop different **techniques** that teachers use to put these insights to work, such as *do not read aloud text that is also being presented visually, in order to avoid overloading the phonological loop*.
4. To help embed this new **practice**—for example, by repeatedly practising remaining silent for a short period, while students read a clearly highlighted quote from a slide.

Table 1 provides a concise summary of our hypotheses about how PD can fail if any of these four purposes are not addressed. If PD brings about the necessary changes to insight (I), goal-directed behaviour (G), and technique (T), but does not embed this in practice (P), then teachers will tend to revert to established ways of working (Hobbiss et al., 2021). This is depicted in row 2 of the table. If PD brings about the necessary changes to I and G, but not to T and P, then the insights might never be translated into practice in the classroom to begin with. This problem has long been referred to as the 'knowing-doing gap' in the teacher education literature (Knight et al., 2013) and is depicted in row 3 of the table. Row 4 shows an extreme case, in which PD brings about the necessary changes to I, but not G, T or P, in which case teachers leave the PD without intending to change their practice. Finally, row 5 depicts the case in which PD brings about the necessary changes to G, T and P, but not I. In this case, the PD has failed to provide an understanding of why (and when) a particular practice is effective. This can lead to misapplication of a technique in a way that renders it ineffective (Mokyr, 2002), sometimes referred to as a 'lethal mutation' in the education literature (Brown & Campione, 1996, p.259). By contrast, when PD succeeds in addressing I, G, T and P, we hypothesise that it is more likely to be effective.

Different mechanisms will be suitable to address each purpose. We previously gave the example (mechanism) of practising a new technique in a classroom environment. This might help embed practice (P) but would be less likely to instil insight (I). In Part 2 of this report, we set out how we identified mechanisms addressing each of the IGTP purposes.

Table 1: Theorising how PD might be ineffective, using the IGTP model

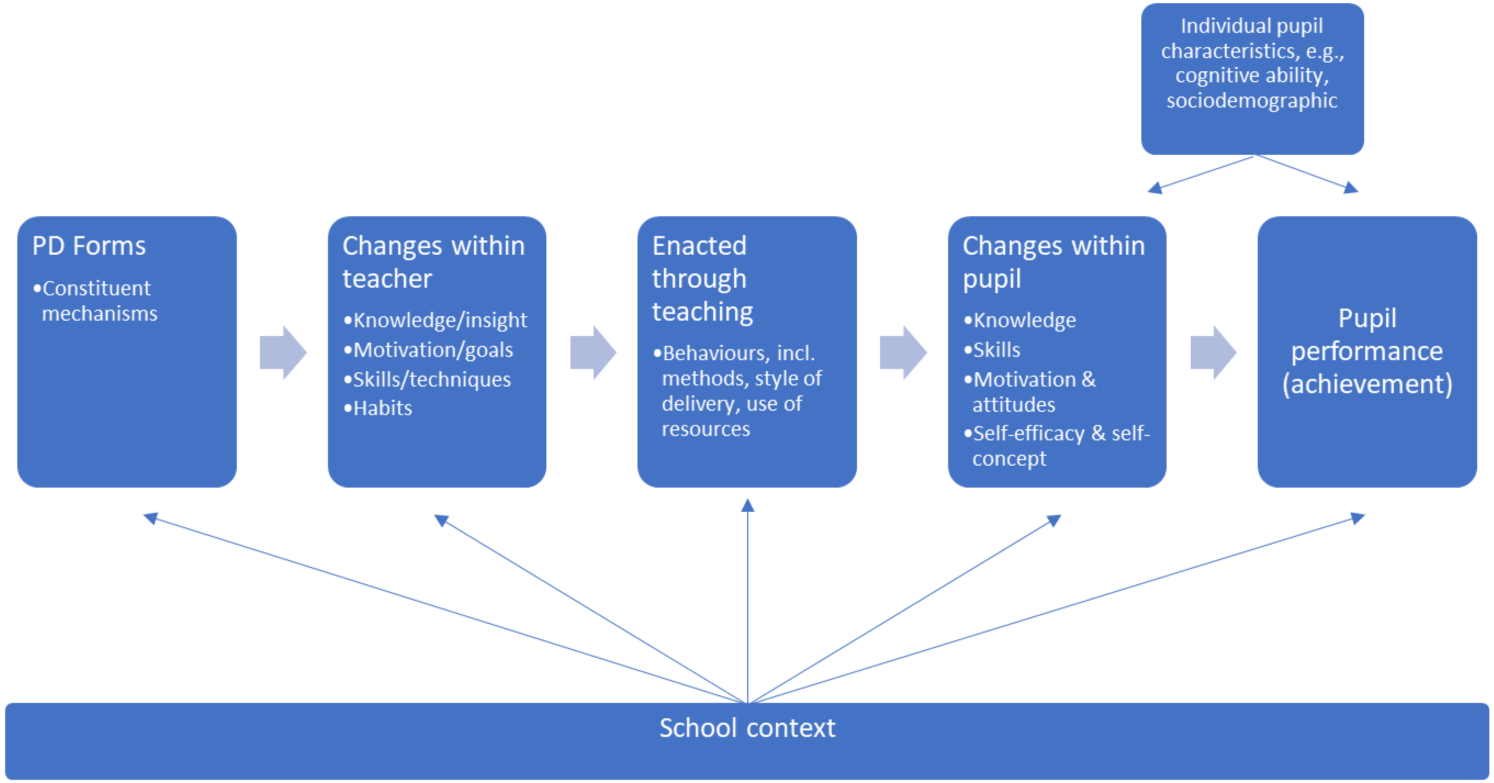
(Instil) **Insight**	**(Motivate)** **Goals**	**(Develop)** **Techniques**	**(Embed)** **Practice**	**Consequences**
✓	✓	✓		Revert to old habits
✓	✓			Knowing-doing gap
✓				No implementation
	✓	✓	✓	Misapplication
✓	✓	✓	✓	May work

Table 2: Glossary of terms

Balanced design	A PD *programme* with a balanced design includes at least one *mechanism* addressing each of the four *purposes* of PD.
Embed practice	To help a teacher apply a technique/act on an insight/pursue a goal consistently in the classroom.
Form	A form is a type or category of PD, specified at a higher level of abstraction than a *programme*. Forms can be defined as clusters of *mechanisms*. Conditional on having these mechanisms, forms can accommodate variation in the specific materials and activities involved. One example of a PD form is lesson study.
IGTP	The acronym summarising the four *purposes* of PD: *instil insight, motivate goal-directed behaviour, teach techniques, embed practice.*
Imbalanced design	In a PD *programme* with an imbalanced design, there is no mechanism addressing one or more of the four *purposes.*
Instil insight	To help a teacher gain a new, evidence-based understanding of teaching, their students, or themselves. An example of an insight would be recognising the impact of limited working capacity on students. This is one of four *purposes* of PD.
Mechanism	In general, a mechanism refers to entities and activities organised in such a way that they are responsible for the phenomenon of interest. In PD, a mechanism is a component of the PD that could not be removed or altered without changing the impact of the PD on teaching and learning. One example of a PD mechanism is the rehearsal of a new technique in a realistic classroom setting in order to *embed it in practice*.
Motivate goal-directed behaviour	To encourage a teacher to pursue an action, in pursuit of a specific, conscious aim. An example would be a teacher resolving to limit the burden on students' working memory in their lessons. This is one of four *purposes* of PD.

Programme	A PD programme is a specific set of activities and materials intended to improve teaching ability. In well-established programmes, the activities are sometimes codified in a programme manual and the materials can sometimes be acquired off-the-shelf as part of a resource pack. An example of a programme is Dialogic Teaching.
Purpose	We hypothesise that PD has four purposes, without which it is less likely to be successful: *instil insight*, *motivate the pursuit of goal-directed behaviour*, *develop new techniques* and *embed practice*.
Teach techniques	To help a teacher master a new teaching practice. An example would be a teacher not reading text which is presented visually. This is one of four *purposes* of PD.

Figure 1: Logic model showing pathway from professional development interventions to pupil achievement



1.3 Research questions

The objective of this review is to synthesise the existing literature to identify characteristics of effective teacher PD (defined as PD which helps teachers increase pupil achievement).

To achieve this, we address four research questions:

1. What are the characteristics of the studies and interventions in the experimental impact evaluation literature on teacher PD?
2. Overall, are teacher PD interventions effective at increasing pupil achievement compared to business as usual?
    a. Does this vary based on study characteristics (features of the evaluation not specific to the intervention itself)?
    b. Does this vary by study-level pupil disadvantage or teacher experience levels?[2]
3. Which forms of PD are associated with the greatest impact?
    a. Which forms (clusters of mechanisms) do we observe in the literature?
    b. Are forms more likely to be effective when they incorporate mechanisms addressing all four purposes: instil insights (I); motivate goal-directed behaviour (G); develop techniques (T); and embed practice (P)?
    c. Which forms (clusters of mechanisms) are associated with the largest effects on teacher practice and pupil achievement?
4. What supports successful implementation of PD programmes?[3]

The results of this analysis will directly inform the development of recommendations for a subsequent EEF guidance report on the characteristics of effective PD. This guidance report will be developed separately by the EEF, in consultation with a panel of practitioners and academics, drawing on the findings of this report.

1.4 Overview of the methods

The methodological approach was designed and planned in relation to the research questions; therefore, in this summary, we present an overview of the methods organised by research question. The detailed methods are presented in Appendix 1.

Research Question 1: What are the characteristics of the studies and interventions in the experimental impact evaluation literature on teacher PD?

This review employed a systematic search of ten scholarly databases and other resources covering the education field, with the aim of capturing all the relevant literature. For inclusion in the systematic review, we screened all studies to meet all the following inclusion criteria:

1. **Publication year:** We restricted our search to studies published during or after 2002. Specifying a start date was necessary in order to ensure that we completed the project by the deadline and within the available budget. We chose this specific start date because it is the year in which the Institute of Education Sciences in the US was established, which marked the beginning of a new era in terms of the funding and conduct of rigorous experimental evaluations of PD in education (Hedges & Schauer, 2018). The end date for the search was 2

[2] The phrase 'teacher experience levels' was added to this question after the protocol was published.
[3] This was not stated as a stand-alone research question in the protocol. However, it was included in the protocol (page 26).

November 2020 and the searches were mainly undertaken between 12 and 16 November 2020.

2. **Language:** We included studies written in English and excluded studies in languages other than English. This was necessary due to resource constraints.
3. **Reporting:** We included studies reported in journal papers, working papers or institutional report formats, as well as doctoral theses that could be obtained via current UCL subscriptions. We excluded conference papers or extended abstracts on the grounds that they do not contain enough information to assess quality or to extract sufficiently detailed information about intervention components.
4. **Geography:** We only included programmes implemented in OECD countries.[4] This was necessary to ensure some level of comparability between the contexts in which the programmes were conducted.
5. **Intervention:** We included studies that evaluated teacher PD programmes, as defined above. We excluded programmes that incorporated a change in the pupil:teacher ratio (for example, training teachers in the delivery of small-group or one-to-one tuition). Our focus is on the relationship between certain characteristics of PD and the impact of that PD on pupil achievement. Small-group tuition is known to be highly effective in and of itself (Nickow, Oreopoulos, & Quan, 2020), so including evaluations of PD that also incorporated small-group tuition would confound the relationship between the PD characteristics and the impact of the PD. In line with our definition of PD, we excluded programmes that aimed only to briefly familiarise teachers with educational technology or curriculum materials.
6. **Intervention population:** We included studies of qualified teachers working in formal settings (for example, early years settings, schools, colleges) who teach pupils between the ages of 3 and 18. We excluded studies in higher education (HE) settings.
7. **Design:** We included randomised controlled trials (RCTs) and excluded all other evaluation designs. The justification for this is that RCTs are the only design that provide unbiased causal impact estimates—essential for our synthesis approach.
8. **Outcomes:** We included studies that measured pupil achievement using standardised tests in any school curriculum subject. We excluded researcher-designed (as opposed to standardised) tests because they have been shown to display systematically larger effect sizes (Cheung & Slavin, 2016). We excluded studies that only measured achievement using observational protocols (e.g. Assessment of Scientific Argumentation in the Classroom; Sampson et al., 2012) or that used holistic teacher judgements, as opposed to quantitative aggregation of marks from multiple test items (e.g. the Early Years Foundation Stage; DfE, 2020). This is justified because of various studies showing systematic bias in teacher assessments (e.g. Black & New, 2020). We included high-stakes test scores (set by any branch of government) and low-stakes test scores (not set by any branch of government).
9. **Comparison group:** We included studies that provided no PD, business as usual PD, or waitlist intervention for the control group.

For a detailed account of the screening process and coding frame, see Appendix 1. A detailed account of the software, search terms and process used to conduct this search can be found in Appendix 2. A full PRISMA flow diagram can be found in Appendix 3 and a detailed table of study characteristics can be found in Appendix 4.

[4] https://www.oecd.org/about/document/list-oecd-member-countries.htm

Using the information extracted from the studies, we presented descriptions of the PD interventions using text and tables to give an overview of the evidence base.

Research Question 2: Overall, are teacher PD interventions effective at increasing pupil achievement compared to business as usual?

To answer this question, we calculated effect sizes for included studies in the EPPI-Reviewer software. For our student achievement outcomes, the data were continuous test scores and our effect sizes were calculated using Cohen's *d*. These were typically calculated from means and standard deviations (or SEs or CIs). In the absence of those statistics, we used (in decreasing order of preference) t- or F-statistics, or p-values. Any study results that could be converted to an effect size were used (see full list of formulae in Lipsey & Wilson, 2001). For the subset of studies for which it was possible, we also calculated Hedges' *g*. We report some of our main results using both Cohen's *d* and Hedges' *g*.

Where multiple test score outcomes were reported and a primary outcome measure was specified, we used the primary outcome measure to calculate the effect size. If multiple test score outcomes were reported and no primary outcome was specified, we calculated effect sizes for all test score outcomes.

To calculate meta-analytic average effect sizes, while accounting for some studies having multiple outcome measures, we used robust variance estimation random effect meta-analysis (Hedges, Tipton & Johnson, 2010; Tanner-Smith & Tipton, 2013) using Stata. We then reported pre-registered sub-group analyses and sensitivity tests to check for variation in effect sizes based on study characteristics, including indicators of study quality, and teacher and pupil characteristics (see Appendix 6). We tested for publication bias using trim-and-fill (Duval & Tweedie, 2000), *p*-curve (Simonsohn et al., 2014a; Simonsohn et al., 2014b), and weight function methods (Vevea & Hedges, 1995), and graphically depicted a contour-enhanced funnel plot (Peters, Sutton, Jones, Abrams & Rushton, 2008)—see Appendix 7 for full details.

Research Question 3: Which forms of PD are associated with the greatest impact?

As discussed in Section 1.1, previous reviews which have attempted to identify the characteristics of effective teacher PD have lacked a way of distinguishing the causally active from the causally redundant components of teacher PD. To address this, we drew on our definition of 'mechanisms' and 'evidence of mechanism' to develop a coding framework of mechanisms for which we have independent evidence of mechanism (see Part 2 below for more detail). This allowed us to code up all of the 121 PD programmes in our sample based on whether or not they contain a list of probable causally active components. First, we developed a detailed mechanisms coding frame including definitions, examples, non-examples and if-then rules based on extensive practice coding. Then two authors (SS and HFW) double coded 46 papers using this coding frame and achieved 82% agreement at the mechanism level. The two coders met to discuss discrepancies until consensus was reached. The coding frame was then revised to further eliminate ambiguity and to support consistent coding (see Appendix 5 for the final version). The subsequent 58 papers were coded for mechanisms by a single author (HFW).

At this stage, we also coded several other features of the papers, including:

- focus on early-career or newly-qualified teachers;
- intended number of hours participating;
- overall pupil disadvantage;
- attrition; and

- pre-registration/pre-specification.

We also coded the content of the professional development programme in two ways, which we hoped would allow us to condition out the relationship between content and PD impact. First, we coded the aspect(s) of teaching on which the study focused. To do so, we used Kennedy's (2016b) definition of persistent problems in teaching:

- Portraying the curriculum: 'to portray curriculum content in a way that makes it comprehensible to naïve minds, and to decide how that portrait will be constructed from some kind of live activity' (Kennedy, 2016b, p.10).
- Managing behaviour: 'to contain student behavior not only as a matter of public safety but also to ensure that students are not distracting each other, or distracting the teacher, from the lesson' (Kennedy, 2016b, p.12).
- Enlisting participation: fostering 'active engagement' or, if that is impossible, fostering 'cooperation' from students (Kennedy, 2016b, p.11).
- Exposing student thinking: 'teachers must continually find ways to expose their students' thinking' (Kennedy, 2016b, p.12).

Second, we coded the PD programmes against four broad areas of focus:

- Science of learning/cognitive science: PD focused on the use of findings from cognitive science relating to how memory works and how humans learn (for example, Cromley, 2016).
- Inquiry/discovery/problem-based teaching: PD focused on pedagogy that encourages students to construct knowledge for themselves via solving problems and completing authentic tasks, working with autonomy (for example, Meyers et al., 2016).
- Formative assessment/responsive teaching: PD focused on how to elicit evidence of pupil understanding and then use this evidence to adapt the next steps in instruction (for example, Arens et al., 2012).
- Data driven instruction: using cyclical class-wide testing to systematically collect data on pupil progress and then refocusing or differentiating instruction based on the findings (for example, Cavalluzzo et al., 2014). Data-driven instruction is differentiated from formative assessment by the systematic use of testing to collect quantitative data and the longer cycles between collection of evidence and subsequent adaptation of teaching.

To investigate the relationships between mechanisms and PD impact, we ran meta-regressions of the intervention effect sizes on the number of mechanisms per intervention. We then plotted this relationship using 'bubble plots'. To enable us to produce bubble plots, we used the primary outcome from each study (if specified) or else a randomly chosen non-primary outcome. We also produced the equivalent bubble plots for pre-specified subgroups of PD interventions focusing on specific pedagogical content and for indicators of study quality. We did not enter each mechanism separately into meta-regressions because we did not have sufficient sample size to do so and different mechanisms addressing the same purpose (such as 'instil insight') are likely to be substitutes, rather than having an additive linear relationship on the outcome.

As discussed in Section 1.2, forms of PD can be defined based on combinations of mechanisms. To define these forms, we relied on the existing academic literature (see Section 2.2 below). To investigate the relationship between different combinations of mechanism (forms) and PD impact, we first calculated the meta-analytic average effect size for different forms. We did this using robust variance estimation random effect meta-analysis using the 'robumeta' command in the Stata software.

To investigate whether PD programmes incorporating mechanisms addressing the four IGTP purposes of PD are more effective, we used two approaches. The first was—again—robust variance estimation random effect meta-analysis, with the meta-analytic average effect sizes and 95% confidence intervals plotted in charts. The second approach employed some of the tools of crisp set qualitative comparative analysis (QCA; Thomas, O'Mara-Eves, & Brunton, 2014). This involved creating truth table, in which each of the unique combination of I, G, T and P were included as rows of the table. We then calculated the consistency with which the study-specific effect sizes in each row exceeded a given threshold. This was implemented using the 'fuzzy' command in the Stata software.

Research Question 4: What supports successful implementation of PD programmes?

In the protocol (Sims et al., 2021), we specified that we would randomly select 20 studies with Implementation and Process Evaluations (IPEs) from our final sample of studies. We randomly ordered all studies with implementation data. However, we found that the vast majority of such studies offered limited fidelity data. All those studies featuring detailed IPEs were those funded by the EEF. Since 22 EEF studies remained in our sample, we chose to focus this element of the review on them. We believed this was particularly appropriate, since some elements of the supports and barriers to implementation are likely to reflect English educational policies and practices (Lendrum & Humphrey, 2012). Within this sample, we found that one study (Torgerson et al., 2014) contained only fidelity data, so this was excluded, leaving a final sample of 21 studies.

Since different evaluators measured different aspects of implementation in different ways, we concluded that attempting to compare incommensurable quantitative measures of fidelity was unlikely to be fruitful. We therefore conducted a qualitative content analysis of the barriers and supports for professional development from the IPE evaluation sections of the included studies (Graneheim & Lundman, 2004). We began with a loose theoretical framework around the key questions of interest: fidelity, adaptation, and barriers to implementation, derived from the literature (for example, Durlak & DuPre, 2008). We coded for:

- factors supporting teacher responsiveness, fidelity, and (or) positive adaptations; and
- factors hindering teacher responsiveness and fidelity, or encouraging negative adaptations.

We also coded features of the intervention, the support system, and the school context (Domitrovich et al., 2008). Having coded the texts using these broad themes, we iteratively refined our coding to identify sub-themes: for example, we examined and categorised the reasons given for adaptation, responding to the themes emerging from the content (Cohen, Manion & Morrison, 2018, p.677). We excluded from our coding any references to opinions held by isolated individuals, unless they were described as characteristic of the group. For example, having explained that multiple schools had adopted cooperative learning, one evaluation then stated that 'one teacher felt that this particular element was burdensome' (Biggart et al., 2015, p.23). The evaluator was not suggesting that this was a broadly representative opinion, so it was not coded.

Part 2: Full conceptual framework

This review seeks to go beyond previous studies of PD by providing evidence about the forms and mechanisms of effective PD. In this section, we discuss the theoretical and developmental review we conducted in order to identify potentially promising forms and mechanisms. First, we discuss how we identified a set of mechanisms. Then, we define forms of PD using these mechanisms.

2.1 Identifying mechanisms of professional development

Mechanisms are 'entities and activities organized in such a way that they are responsible for the phenomenon' (Illari & Williamson, 2012, p.14).[5] For something to be considered a mechanism, we require it to be 'sufficiently reliable across individuals and context' that it constitutes an 'empirical generalization' (Willingham, 2017, p.168). As explained above, this was necessary to support our inferences about a potential mechanism also being causally active within a PD programme. In line with the Illari & Williamson (2012) definition, we also required that there be some account of how each potential mechanism produced the outcome of interest. These criteria informed our efforts to identify plausible mechanisms in PD. We stress that this process did not aim to identify all possible or plausible mechanisms supporting teacher PD. Our goal was to identify those mechanisms for which we could find (a) empirical causal evidence across domains, and (b) at least one example in the PD studies we reviewed.

For example, one mechanism is practical social support. Practical social support provides the individual with access to a peer who has knowledge and experience of the desired behaviour or practice (Dennis, 2003). The guidance a peer offers makes behaviour change more likely. A systematic review in health promotion found that peer support helps change knowledge, attitudes and beliefs, and encourages practice change (Ramchand et al., 2017). Similar findings emerge from reviews of breastfeeding (Jolly et al., 2012), diabetes care (Dale et al., 2012), and prisoner outcomes (Bagnall et al., 2015).

To develop a list of candidate mechanisms, we built upon existing research identifying and organising behaviour/practice change mechanisms. In particular, Susan Michie and her colleagues (2013) conducted a series of exercises to review existing theoretical accounts of behaviour change and consult with experts to identify and organise mechanisms. Their findings are encapsulated in Version 1 of the Behaviour Change Technique Taxonomy (hereafter, 'the Taxonomy'). This organises 93 mechanisms into 16 groups, such as 'Social support', 'Self belief', and 'Goals and planning'. Mechanisms in the 'Goals and planning' group include 'Action planning', 'Making commitments', and 'Agreeing behavioural contracts'. The Taxonomy provided the starting point for our study, but we conducted two further tasks to ensure that candidate mechanisms did indeed generalise across domains, and were sufficiently well specified for our study. These tasks, described below, were conducted in parallel, but are described sequentially.

We sought empirical evidence that mechanisms in the Taxonomy operate in multiple domains. This step was necessary because Michie et al.'s (2013) review organised plausible mechanisms, but did not test the empirical evidence for each one. We searched for systematic reviews of evidence for the

[5] Note on terminology: Following the definition of a mechanism proposed by Illari and Williamson (2012), a mechanism is the entities and activities organised in such a way that they are responsible for a phenomenon. This term is related to, but distinct from, the idea of a 'mechanism of action', which is sometimes used in related literature (Carey et al., 2019). Some of our mechanisms are equivalent to what Michie et al. (2013) call 'behaviour change techniques'. We have not used the term 'behaviour change techniques' here because our final list of mechanisms includes some components (such as 'manage cognitive load') which are not behaviour change techniques.

effect of the mechanism in one or more domains. Where necessary, our search employed logically connected terms. For example, we searched for 'implementation intentions', while seeking reviews of the mechanism 'Action planning'. We searched for 'peer support' when seeking evidence for the mechanism 'Practical social support' because most reviews of social support reflect theorising around the protective effects of social support on individuals (for example, Cunningham & Barbee, 2000). Searches for 'peer support' found multiple reviews relevant to the mechanism described in the Taxonomy (for example, Ramchand et al., 2017). We found strong support for some mechanisms. For 'goal setting', for example, we identified a systematic review which showed that setting a goal increased the chances of an individual taking action across domains including health, education, and sport (Epton et al., 2017). However, other plausible mechanisms have not accrued commensurate evidence. For example, the same systematic review (Epton et al., 2017) found no studies meeting their inclusion criteria for the suggested mechanism, 'highlighting a discrepancy between the goal and current performance standard'. We retained mechanisms in our coding framework where we found supportive empirical evidence from reviews and we dropped mechanisms where we were unable to find such evidence.

We also tested the relevance and applicability of the mechanisms in the Taxonomy to a randomly-selected sample of studies eligible for inclusion in our review. This was necessary because, while the Taxonomy sets out to offer a domain-general set of mechanisms, the team developing it primarily comprised experts in psychology, health psychology and behavioural medicine (Michie et al., 2013, p.89). Some adaptations and developments were therefore needed in order to apply these mechanisms to research on teacher PD. We iteratively coded 20 studies (in groups of five studies), discussing our coding and refining our understanding of the mechanisms as we did so. This process led us to adapt the Taxonomy in four ways: removing mechanisms, refining definitions, simplifying, and adding mechanisms.

1) Removing mechanisms

Some mechanisms were clearly irrelevant to PD programmes. For example, we know of no PD programmes that have offered 'Biofeedback' (using an external device such as a heart rate monitor to provide feedback). Other mechanisms seemed relevant, but did not occur in our sample of 20 studies. For example, PD providers might plausibly encourage a teacher to identify their strengths, but none of the interventions in our sample included this mechanism. Our subsequent coding of the full set of interventions in our study bore this out.

2) Refining definitions

In some cases, we needed to refine the definition of the mechanism to ensure it was sufficiently specific for our purposes. For example, Michie et al. (2013) state that, when an intervention describes a person receiving training in a skill, three mechanisms should be coded: 'Instruction on how to perform a behavior', 'Behavioral practice/rehearsal', and 'Demonstration of the behavior'. Since all of the interventions in our sample included training of some sort, we only coded rehearsal and demonstration if studies explicitly stated that these latter two had also occurred.

3) Simplifying

In some cases, the Taxonomy's specificity proved too exacting to describe PD. For example, the Taxonomy separates feedback on behaviour (for example, about how many questions a teacher asks) from feedback on the outcome of that behaviour (for example, about students' response to those questions). However, the interventions in our sample universally stated that a teacher received feedback from a coach, without differentiating between the two types of feedback. A coach's feedback might include mention of both a teacher's behaviour ('I noticed you using a different questioning technique') and the outcome ('which seemed to elicit more sophisticated student responses'). Since we could not differentiate between these two mechanisms in the interventions in our sample, we merged them into a single mechanism called 'Feedback'. Our subsequent coding of the full set of interventions in our study also support this decision.

4) Adding mechanisms

The mechanisms in the Taxonomy included several that related to motivating goal-directed behaviour (G) (such as 'Goal setting'), developing new techniques (T) (such as 'Modelling'), and embedding practice (P) (such as 'Context-specific repetition'). However, the Taxonomy did not include mechanisms that might promote changing mental models, and the learning of declarative knowledge—insight (I). Hence, we searched for plausible mechanisms drawn from cognitive science. We identified four such mechanisms: focused presentation, simplified presentation, promoting encoding, and retrieval. We did not find any examples of simplified presentation or promoting encoding in our initial sample, so these were dropped from our full coding framework. This left us with two mechanisms. The first was managing cognitive load (that is, the cognitive load of the teacher taking part in the PD). In a review of the importance of cognitive load, Sweller et al. (2019) set out how this can promote the comprehension and learning of new ideas, through removing redundant information, employing the modality effect (dual coding), and providing completion problems. The second mechanism was revisiting past material. A review of retrieval practice in education identified 23 experiments conducted in classroom settings and found consistent evidence that retrieval practice promoted learning (Moreira et al., 2019). A review of laboratory experiments, again in education, provided evidence that retrieval practice also promotes future learning (that is, that reviewing past learning makes it easier to learn new ideas—Pastotter et al., 2014).

As a result of this process, we identified 14 mechanisms (see Table 3). This list does not represent all possible or plausible mechanisms supporting teacher PD. Rather, it contains all the mechanisms for which (a) we found empirical causal evidence across multiple domains, and (b) we found examples in the teacher PD studies we reviewed. Future research in basic social science may permit researchers to add to (or indeed subtract from) this list.

Organising the mechanisms using the IGTP framework

We organised these mechanisms using the IGTP framework. This allows us to clarify how the mechanisms contribute to effective PD. It also allows us to test the hypothesis that PD programmes with at least one mechanism addressing each of the four purposes of PD (balanced designs) are more likely to be effective. A challenge in allocating each mechanism to a single category is that the four purposes of PD often play out in a broadly sequential way, such that insight (I) precedes motivating goal-directed behaviour (G), which precedes the teaching of techniques (T), which precedes embedding practice (P). While we acknowledge that things sometimes do not play out in this order (practice might precede new insights, for example), this suggests that mechanisms will often have some 'downstream' effects. We therefore allocated mechanisms to the four IGTP purposes by placing the mechanisms in the highest (as listed in Table 3) of the four PD purposes to which they theoretically contribute. (For more information on how each mechanism is intended to work in theory, see the full coding frame in Appendix 5.)

For example, the 'manage cognitive load' and 'revisit prior learning' mechanisms in theory contribute to the purpose of instilling insight and not to the other three purposes of PD. This makes it straightforward to allocate them to the insight purpose (I). 'Goal setting', 'credible source', and ‘praise/reinforce’ clearly do not contribute to instilling insight. However, they do contribute to motivating goal-directed behaviour (G), as well as perhaps to embedding practice (P). We therefore place these three mechanisms in the highest (motivate goals) of these two categories. To reiterate, this reflects our judgement that these three mechanisms in theory primarily contribute to motivating goal-directed behaviour (G) but may also make a secondary contribution downstream to embedding practice. We recognise that some of these theoretical claims will be arguable. In our empirical analysis in Section 4.3, we therefore conduct a number of sensitivity checks to investigate whether moving mechanisms into different purposes affects our overall results.

Table 3: Integrating the mechanisms in the IGTP model

Purpose	Mechanism
Instil insight (I)	0.1 Manage cognitive load
	0.2 Revisit prior learning
Motivate goals (G)	1.1 Goal setting
	9.1 Credible source
	10.4 Praise/reinforce
Teach techniques (T)	4.1 Instruction
	3.2 Practical social support
	6.1 Modelling
	2.2, 2.7 Feedback
	8.1 Rehearsal
Embed practice (P)	7.1 Prompts/cues
	1.4 Action planning
	2.3, 2.4 Self-monitoring
	8.3 Context-specific repetition

Note. Numbers (e.g. 2.3) refer to the codes used in Michie et al. (2013). Mechanisms 0.1 and 0.2 are additions to the Michie taxonomy for this project. Some mechanism labels have been adapted from Michie et al. (2013)’s—for example, we have adopted ‘Modelling’ for ‘Demonstration’ and ‘Context-specific repetition’ for ‘Habit formation’.

To support our subsequent coding, we produced a ‘Mechanisms Coding Framework’. Table 4 provides an example of the type of information we included for each of the mechanisms. The full table, for all 14 mechanisms, can be found in Appendix 5.

Table 4: Extract from the coding frame for the mechanism ‘Action planning’

Mechanism: ‘Action planning’ (Purpose G)	
Information	**Example (for the mechanism)**
Coding guidance (from Michie et al., 2013)	Prompt detailed planning of performance of the behaviour (must include at least one of context, frequency, duration and intensity). Context may be environmental (physical or social) or internal (physical, emotional or cognitive) (includes ‘Implementation Intentions’). Note: evidence of action planning does not necessarily imply goal setting, only code latter if sufficient evidence.
Additional coding guidance for this study (derived from the developmental review)	Includes lesson planning, if planning is to apply techniques in lessons
Example (from one of the studies we reviewed)	‘Staff teams in each of the eight research schools worked to develop their plans to implement teaching initiatives focused on

	enhancing student literacy outcomes via a focus on two of the four aspects of oral language competence targeted by ICPALER.’ (Snow, 2014, p.500)
Non-example (i.e. a superficially-similar example which did not meet our criteria for the mechanism, from the studies we reviewed)	‘Teachers were given time to meet in grade-level groups and as school teams to discuss how to modify the materials to meet their specific students’ needs’ (Olson et al., 2017, p.7). *No specific context/frequency for action.*
Summary of supporting evidence	A review of 94 experimental studies found strong evidence that setting implementation intentions about how to act and when improved performance for a range of populations in health, goal pursuit, academic and laboratory tasks. The review included both correlational and experimental research, but found that implementation intentions were similarly effective in both cases (Gollwitzer & Sheeran, 2006).
Statement of theory explaining the operation of the mechanism	‘Implementation intentions should enhance people’s ability to initiate, maintain, disengage from, and undertake further goal striving and thereby increase the likelihood that strong goal intentions are realized successfully. In other words, this form of planning is expected to bridge the intention–behavior gap.’ (Gollwitzer & Sheeran, 2006, p.82)

2.2 Identifying forms of professional development

As we set out in our initial conceptual framework, one of the benefits of identifying a set of mechanisms is that it allows us to define ‘forms’ of PD in terms of these underlying mechanisms. In the protocol, we set out our intention to use a truth table in order to identify groups of interventions that included the same set of mechanisms (Sims et al., 2021). These commonly-occurring clusters of mechanisms would then be used to define our different forms. However, as a result of ultimately including more mechanisms than we originally anticipated, the truth table ended up having almost as many rows (93 unique combinations of mechanisms) as we had PD programmes in our sample (n = 104). This made it impossible to use the truth table to identify commonly-occurring clusters of mechanisms, because almost all of the rows of our truth table contained one or zero PD programmes.

We therefore relied instead on existing theory to define forms of PD based on configurations of mechanisms. To do this, we went back to the literature and identified terms commonly used to describe different designs of PD. This was challenging because we found many instances of authors using different terms used to describe similar PD designs (for example, professional learning communities, teacher learning communities) and similar terms used to describe quite different PD designs (such as coaching). Indeed, this terminological confusion is one of our motivations for defining forms more precisely in terms of their underlying mechanisms. The research team met as a group to discuss the different terms and how they were used in the literature, with the aim of identifying forms that were (a) clearly defined enough that we could differentiate them in terms of mechanisms used, and (b) sufficiently conceptually distinct from each other. This resulted in a list of three forms: lesson study, instructional coaching, and teacher learning communities. In the following three paragraphs we define each of these three forms in terms of their constituent mechanisms. Future theorising and research might allow for other forms to be identified and defined in terms of their underpinning mechanisms.

Lesson study can be defined as 'observation of live classroom lessons by a group of teachers who collect data on teaching and learning and collaboratively analyze it' (Lewis, Perry & Murata, 2006, p.3). Within this broad definition, however, there is some disagreement about the key features of lesson study. For example, some authors emphasise the role of expert advisors, which appears to incorporate aspects of our 'instruction' mechanism (Fernandez, 2002), while other authors do not emphasise this mechanism (Lewis, Perry, & Murata, 2006). Other authors emphasise that lesson study may or may not involve the same teacher repeating the same lesson for a second time, which is similar to our 'context specific repetition' mechanism (Fernandez & Yoshida, 2012; Mutara, 2011). Despite these ambiguities, there does seem to be consensus on the central importance of the following three mechanisms for lesson study (Fernandez, 2002; Fernandez & Yoshida, 2012; Lee, 2008; Lewis, Perry, & Murata, 2006; Murata, 2011):

- Action planning: when teachers jointly develop the lesson plan for a specific lesson.
- Practical social support: when teachers advise each other on how to develop the lesson.
- Feedback: from the observer-teachers who follow the study lesson.

At a high level, **instructional coaching** can be defined as 'an observation and feedback cycle in an ongoing instructional … situation" (Joyce & Showers, 1981, p.170). Within this broad definition, there is again some variation across different models of instructional coaching. For example, some models of instructional coaching stress the importance of coachees rehearsing new techniques outside the classroom (Bambrick-Santoyo, 2018), whereas others rely on repetition in realistic classroom settings (Gregory et al., 2017). Likewise, some models of instructional coaching emphasise that coaches will need to be more or less directive (corresponding to our instruction mechanism), depending on the expertise and outlook of the coachee (Knight, 2007). Nevertheless, across the literature there seems to be something of a consensus that instructional coaching includes the following core mechanisms (Bambrick-Santoyo, 2018; Gallucci et al., 2010; Gregory et al., 2017; Joyce & Showers, 1981; Knight, 2007; Kraft, Blazar & Hogan, 2018):

- Goal setting: coaches and/or coachees identify specific, focused areas of improvement for coachees.
- Feedback: coaches provide feedback following observations of coachee practice.
- Instruction OR modelling: coaches point coachees toward specific improvements in practice.
- Rehearsal OR context specific repetition: instructional coaching always involves deliberate practice of new techniques, outside or inside the classroom.

Teacher learning communities can be defined as 'teacher learning in a community setting in which teachers come together over time for the purpose of reconsidering their existing beliefs and practice, gaining new professional knowledge and skills and constructing reform agenda that enhances student learning and professional practice' (Chow, 2016, p.288). Again, there is plenty of variation within this broad definition. For example, some authors emphasise the importance of action planning (Wiliam & Leahy, 2012), while other authors do not (Chow, 2016). Similarly, McLaughlin & Talbert (2006) state that teacher learning communities sometimes include analysis of samples of pupils' work or achievement data, but sometimes do not (McLaughlin & Talbert, 2006). Indeed, the only mechanism on which there appears to be a consensus is that teacher learning communities must include practical social support (Chow, 2016; McLaughlin & Talbert, 2006; Priestley et al., 2011; Wiliam & Leahy, 2012; Van Es, 2012).

Defining any PD that incorporates the single 'practical social support' mechanism as a teacher learning community is unlikely to provide much analytical traction. This is particularly true since practical social support turned out to be a very common mechanism in the PD programmes in our study. In order to retain the teacher learning community idea, we therefore consciously adopted a 'strong' definition in which practical social support is combined with two other mechanisms: action planning and goal setting. This is similar to the version of teacher learning communities outlined in Wiliam & Leahy (2012) and tested in Speckesser et al. (2018). To distinguish this from standard teacher learning communities, we explicitly refer to it as a strong teacher learning community.

Table 5 provides a summary of the different mechanisms from our framework incorporated in each of our forms. Indeed, the table illustrates many of the main components of the conceptual framework that we have built up so far. The IGTP purposes are captured in the columns, the three 'forms' of PD outlined in this section are captured in the rows, and eight of our 14 mechanisms are present in the cells. It is important to note that Table 5 sets out the mechanisms that we require a PD programme to have before we consider it to be an example of a certain form of PD. This does not preclude, for example, a lesson study programme having other mechanisms in addition to those outlined in the top row of the table.

Before we move on to our empirical results in Part 3, it is worth making a few further clarifications. In particular, we believe this conceptual framework allows us to talk about PD in a way that is precise, useful, and parsimonious. It is *precise* in the sense that we have reduced a number of well-known PD forms down to a common set of well-defined mechanisms. The framework is also *useful* in the specific sense that, because we have independent empirical evidence that each of these mechanisms holds in multiple domains, PD designers can be more confident that we have identified the active ingredients of each of these forms of PD. This can help avoid lethal mutations (Brown & Campione, 1996, p.259). Finally, the framework is *parsimonious* in that it (intentionally) leaves out much information about a PD programme. For example, looking at Table 5 would clearly be insufficient to guide implementation of any of these forms of PD in schools. Our point here is that this framework is intended to be used for a certain purpose: to characterise the necessary conditions for what PD has to do to improve teaching and learning, and how different PD designs go about attempting to achieve this. There are other purposes for which it would be less useful.

Table 5: Defining 'forms' of PD in terms of their mechanisms

	I mechanisms	**G mechanisms**	**T mechanisms**	**P mechanisms**
Lesson study			Practical social support, Feedback	Action planning
Teacher learning communities		Goal setting	Practical social support	Action planning
Instructional coaching		Goal setting	Instruction, Modelling, Feedback, Rehearsal	Context specific repetition

Note. I = insights, G = goal-directed behaviours, T = techniques, and P = practice.

Part 3: Describing the experimental impact evaluation literature on teacher professional development

3.1 What are the characteristics of the studies and interventions in the experimental impact evaluation literature on teacher professional development?

Research Question 1

Our systematic search and screening process identified 104 studies that met these inclusion criteria and reported sufficient data to calculate an effect size and be included in a statistical meta-analysis. For a detailed account of how these studies were identified, see the PRISMA flow diagram in Appendix 3. Table 6 shows the characteristics of the 104 studies included in this review. Around three quarters are from the US, around half were targeted at primary age pupils, around half were targeted at literacy skills, and around a quarter used high-stakes test score outcomes. Just over two thirds of the PD programmes also included a curriculum component, around four fifths included a focus on portraying curriculum, and 7–15% focused on each of inquiry teaching, formative assessment, and data-driven instruction.

Table 6: Characteristics of the studies included in the statistical meta-analysis

	Characteristics	Count	Proportion
Study characteristics	Location		
	USA	73	70.2%
	UK	25	24.0%
	Other	6	5.8%
	Age group		
	Early years/Pre-kindergarten	29	27.9%
	Primary/Elementary	52	50.0%
	Middle/Secondary/High	28	26.9%
	Subject targeted		
	Literacy/first language	52	50.0%
	Maths	30	28.9%
	Science	12	11.5%
	Other subjects	6	5.8%
	Cross-curricular	17	16.4%
	Test type		
	High-stakes standardised	29	27.9%
	Low-stakes standardised	75	72.1%
	Targeted at early-career teachers	3	2.9%
PD characteristics	PD plus …		
	… curricular reform	41	39.4%
	… educational technology	7	6.7%
	Kennedy’s persistent challenges		
	Portraying curriculum	81	77.9%
	Containing behaviour	16	15.4%
	Enlisting participation	29	27.9%
	Exposing student thinking	17	16.4%
	Broad area of focus		
	Science of learning	1	1.0%
	Inquiry/discovery/problem	16	15.4%
	Formative assessment	14	13.5%
	Data-driven instruction	7	6.73%
	Total:	104	100%

Note. Percentages may not sum to 100 within categories due to rounding or sub-categories not being exhaustive. Counts may not sum to 104 due to categories not being mutually exclusive or exhaustive.

Summary of answer to Research Question 1:	
What are the characteristics of the studies and interventions in the experimental impact evaluation literature on teacher PD?	Most studies are in the USA (70.2%) or UK (24.0%). They are distributed across different phases of education. Half (50.0%) focus on literacy, 28.9% on maths, and 11.5% on science. There is a cross-curricular focus in 16.4% of studies. Only a minority use high-stakes standardised tests (27.9%). Very few (2.9%) are focused specifically on early-career teachers. A large proportion (39.4%) include an element of curricular reform, whereas only 6.7% include an

	element of education technology. Similarly, the largest proportion focus on portraying the curriculum (77.9%).

3.2 Overall, are teacher professional development interventions effective at increasing pupil achievement compared to business as usual?

Research Questions 2, 2a, 2b

Table 7 shows the meta-analytic average impact of teacher PD on standardised test scores. Each estimate is the result of a robust variance estimation random effects meta-analysis implemented using the robumeta command (Hedges, Tipton & Johnson, 2010; Tanner-Smith & Tipton, 2014) in the Stata software. Each row of the table corresponds to a different phase of education: early years (or pre-kindergarten); primary (or elementary); and secondary (or middle/high). The results for different phases of education are shown separately because effect sizes tend to be larger in earlier phases, in part because the standard deviation of test scores is lower among younger children. The first column includes all standardised test scores and the second restricts the sample to experiments reporting impact on high-stakes standardised test scores. These are shown separately because interventions generally show lower effect sizes on high-stakes standardised tests, in part because these tests tend to be less aligned to the content targeted in the PD (Cheung & Slavin, 2016). The third and final column restricts the sample to experiments testing PD interventions that *do not* also incorporate curricular reform. This is intended to isolate the effect of the PD.

Across the three phases of education, impacts on all standardised test scores (column 1) are between .03 and .09 (p < .01). These estimates are each based on at least 28 primary studies and at least 32 separate effect sizes. Impacts on high-stakes test score outcomes (column 2) are similar in magnitude, ranging from .02 to .07. Among the high-stakes test scores, the early years estimate is based on a single study and neither the early years nor the primary estimate is statistically significant (p > .05). Impacts of PD programmes that do not involve curriculum change (column 3) range from .03-.07. Again, the early years estimate is not statistically significant (p > .05). Taken together, these results suggest that evaluated PD programmes tend to have a small positive impact, on average. Although there is some variation in the point estimates across the cells in the table (phases and subsamples), the estimates are generally quite similar and arguably quite small.

Table 7: Overall effect of PD on pupil achievement

	All standardised test scores	High-stakes standardised test scores	PD without curriculum change
Early years k[n]	0.089** (0.043–0.135) 92[29]	0.052 (-0.023–0.123) 1[1]	0.053 (-0.012–0.12) 41[15]
Primary k[n]	0.032** (0.013–0.05) 90[52]	0.018 (-0.055–0.092) 22[16]	0.032* (0.008–0.056) 63[36]
Secondary k[n]	0.046** (0.013–0.078) 32[28]	0.073** (0.023–0.123) 19[15]	0.065* (0.013–0.117) 18[15]

Note. Early years < 4 years of age. Primary is 4–11 years of age. Secondary includes all pupils 11+ years of age, as well as middle school pupils. High-stakes test scores are derived from tests required by law and may be high stakes for either the child, the teacher or the school. Numbers in round parentheses are 95% confidence intervals. k is number of effect sizes and n is number of separate experimental studies. Effect sizes are not strictly comparable across rows or columns. ***p* < 0.01. **p* < 0.05. Produced using random effects meta-analysis, incorporating all standardised test score outcomes using robust variance estimation.

Table 7b in Appendix 6 provides a range of sensitivity tests based on indicators of quality of the underlying studies. Studies with 'acceptable' levels of attrition[6] have much lower average impact (.01) than those with 'unacceptable' levels of attrition (.08). Likewise, studies with more than 50 units randomised to either treatment or control have half the effect size (.04) as those with fewer than 50 units (.10). However, all estimates remain positive and statistically significant ($p < .05$). Of more concern is that studies that were pre-registered have an effect size of .01, much lower than those that are not pre-registered (.07, $p = 0.0001$). Indeed, the meta-analytic average effect size for studies that are pre-registered is not significantly different from zero ($p > .05$). We return to this point in the discussion in Section 6. Table 7c in Appendix 6 looks at variation in effect sizes across different groups of participants. Average impact is slightly smaller among PD programmes focusing on early-career teachers compared with more experienced teachers, and higher among higher deprivation pupils compared with lower deprivation pupils ($p = 0.001$). Appendix 7 provides tests of publication bias which are consistent with either small positive or zero publication bias due to unpublished studies. Taken together, the publication bias tests support the conclusion that there is a positive true effect of teacher PD on test scores across the 104 studies in our sample.

The meta-analytic averages in Table 7 may hide variation in the effects of PD based on the content or substantive focus of the PD. Table 8 sheds light on this by reporting separate estimates within three broad categories of PD content for which we found a sufficient number of studies to support such an analysis. There is little substantive variation in effect sizes (.04-.07) across the three content areas and the 95% confidence intervals all overlap considerably.

Table 8: Overall effect of PD on pupil achievement, by PD content

	Formative assessment	Inquiry-based	Data-driven
By content area:	0.038 (-0.005–0.082)	0.073** (0.021–0.125)	0.042 (-0.051–0.135)
k[n]	24[14]	22[16]	9[7]

Note. Formative assessment = PD focused on formative assessment. Inquiry-based = PD focused on inquiry-based teaching, problem-based learning or discovery learning. Data-driven = PD focused on data-driven instruction. Numbers in round parentheses are 95% confidence intervals. k[n]: k is number of effect sizes and n is number of experimental studies. ***p* < 0.01 different from zero. **p* < 0.05 different from zero. Produced using random effects meta-analysis, incorporating all standardised test score outcomes using robust variance estimation.

Summary of answer to Research Question 2	
Overall, are teacher PD interventions effective at increasing pupil achievement compared to business as usual? a. Does this vary based on study characteristics	Overall, across different phases of education, we find consistent, small, positive effects on student learning, ranging from .03 to 0.9 ($p <$ .01).

[6] Based on What Works Clearinghouse standards.

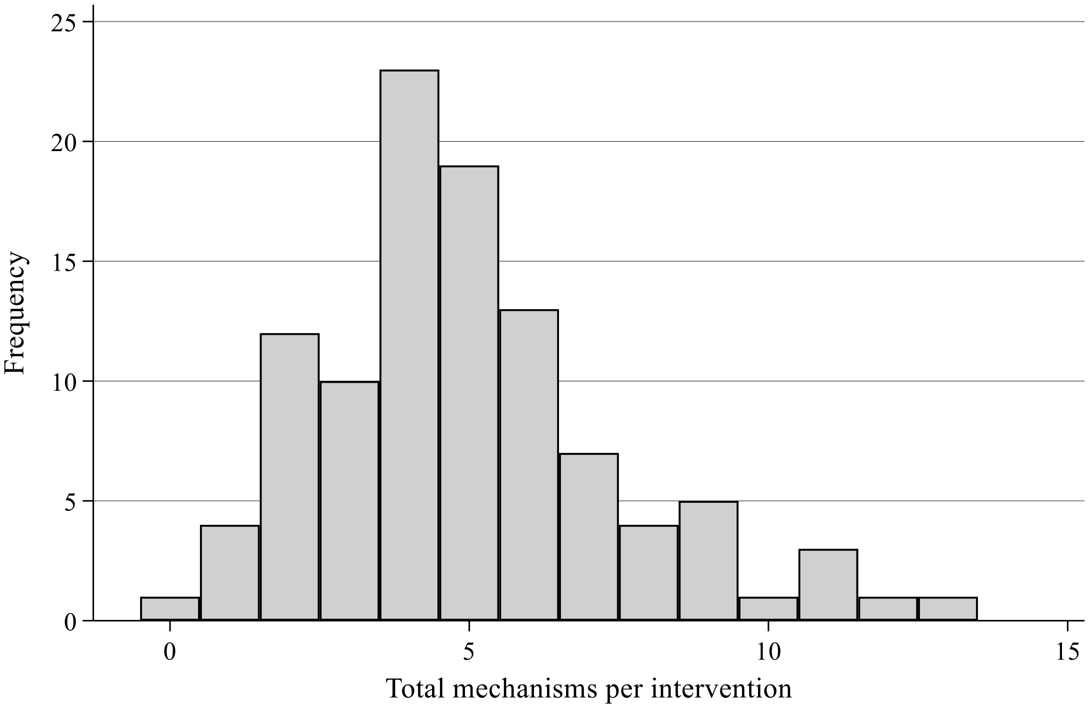
(features of the evaluation not specific to the intervention itself)?	We find variation based on some study characteristics. Impact is lower for studies with 'acceptable' levels of attrition and for larger studies (those with more than 50 units randomised). The effect of pre-registered studies (.01) is much lower than the effect for those not pre-registered (.1).
b. Does this vary by study-level pupil disadvantage or teacher experience levels?[7]	Average impact is slightly smaller among PD programmes focusing on early-career teachers compared with more experienced teachers, and higher among higher deprivation pupils compared with lower deprivation pupils.

Part 4: What are the characteristics of effective teacher professional development?

Figure 2 shows the distribution of the number of mechanisms per PD intervention. The distribution ranges from 0 to 13, has a modal value of 4, and a long right tail between 10 and 13 mechanisms. Figure 3 shows the frequency with which each of the 14 mechanisms occurs across our sample of 104 PD programmes. All 14 mechanisms occur at least once in our sample and the majority of mechanisms occur between 10 and 35 times each. The least common is 'environmental cues', which occurs three times. The most common mechanisms are all in the techniques (T) category, with 'instruction' (96 occurrences) and 'practical social support' (79 occurrences) being the most frequent.

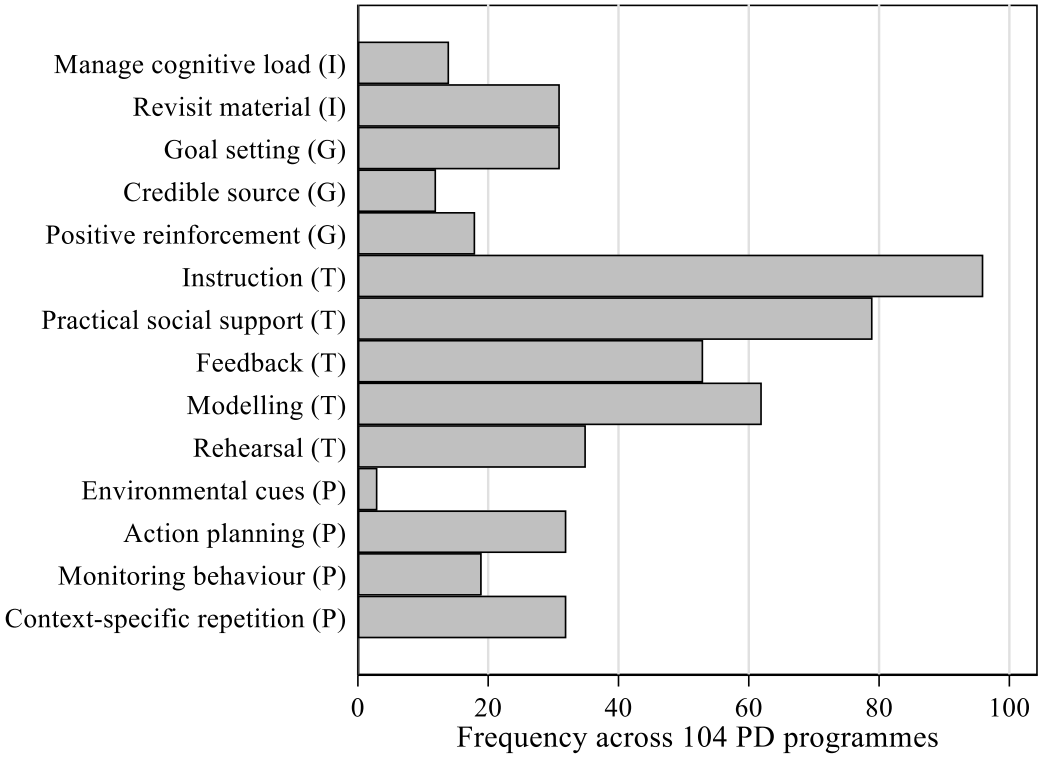
[7] The phrase 'teacher experience levels' was added to this question after the protocol was published.

Figure 2: Number of mechanisms per PD programme



Note. n = 104 studies.

Figure 3: Frequency with which each mechanism occurs across the 104 PD programmes

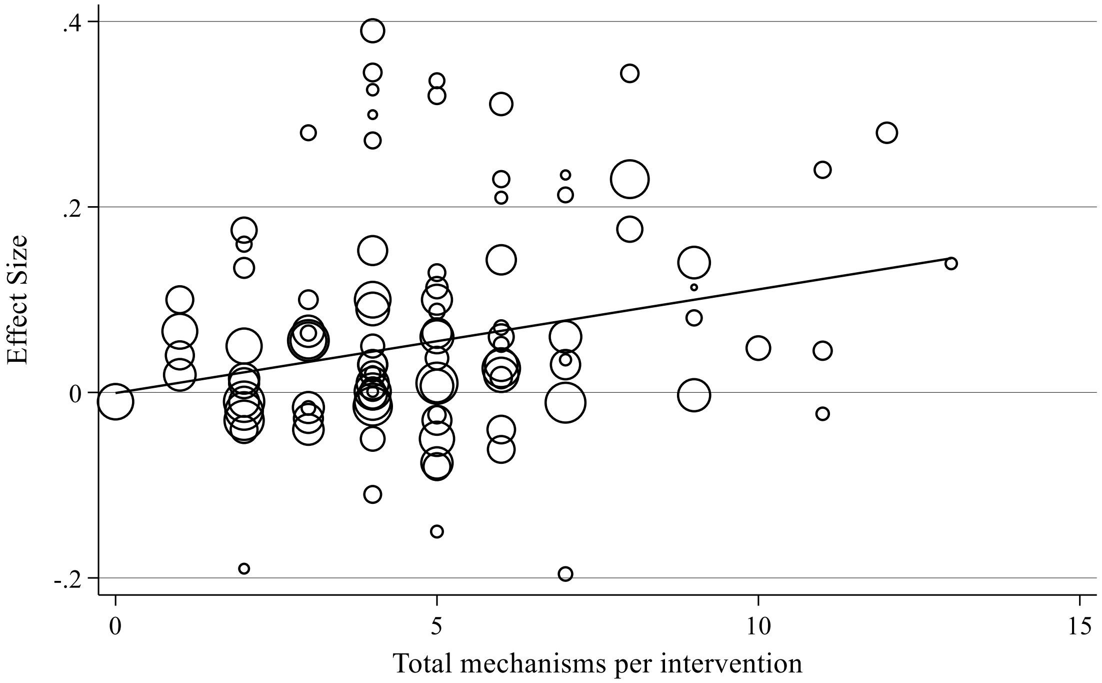


Note. n = 104 studies.

4.1 Do interventions with more mechanisms have higher impact?

In Sections 2 and 3, we hypothesised that each of the 14 mechanisms identified in Section 2.1 makes a contribution to the effectiveness of teacher PD. In order to investigate this empirically, we ran meta-regressions of the intervention effect sizes on the number of mechanisms per intervention. We use the primary outcome from each study (if specified) or else a randomly chosen outcome. Figure 4 plots the number of mechanisms (horizontal axis) against the effect size of the intervention (vertical axis) for all 104 PD programmes in our sample. Each circle in the figure represents a study, with larger circles representing studies with more precise estimates, which are given more weight in the analysis. The meta-regression line of best fit is upward sloping, with a gradient of .01 (*p* = .02). PD interventions incorporating zero mechanisms have an expected effect size close to zero and PD mechanisms incorporating 14 mechanisms have an expected effect size close to .17. Reproducing the chart using Hedges' *g* effect sizes and standard errors provides nearly identical results, with a gradient of .01 and *p* = .02. The relationship shown in Figure 4 is an association, rather than a counterfactual causal estimate. However, we provide independent causal evidence that each of these mechanisms is a domain general cause of learning or practice change (see Section 2.1 and Appendix 5), which provides reason to believe that this association reflects an underlying causal relationship.

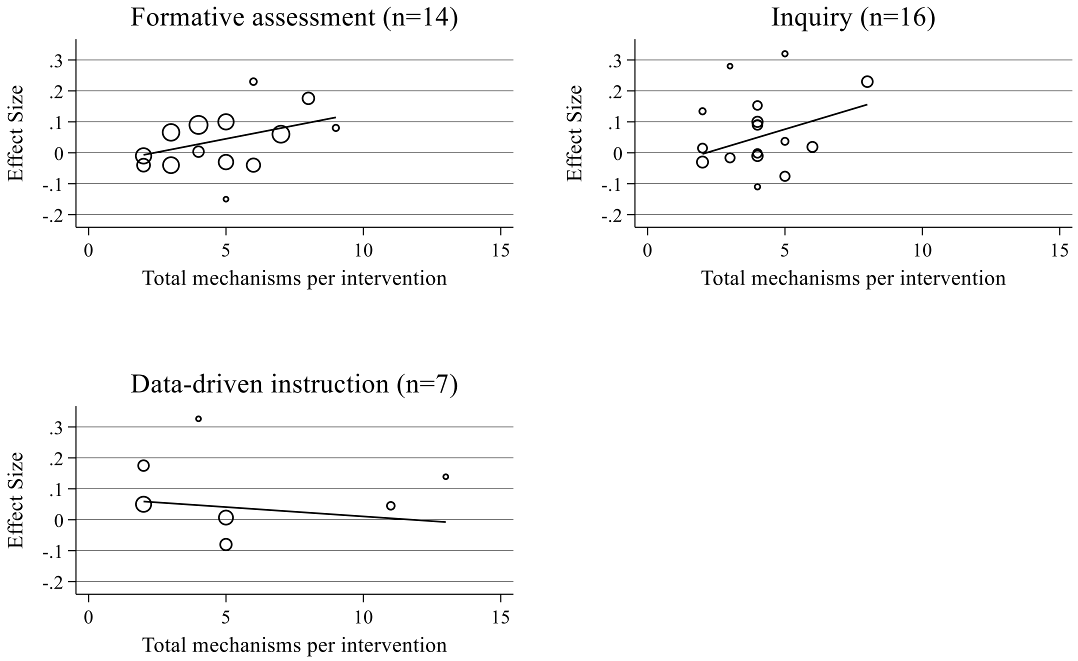
Figure 4: Association between the number of mechanisms included in a PD programme and the impact of the programme on pupil test scores



Note. n = 104 studies. Line of best fit derived from a meta-regression using random effect weights and either the primary outcome as specified in the study, or else one randomly selected outcome per study. Effect sizes >.5 or <-.2 are used in the underlying meta-regression but are not shown in the chart in order to aid visual clarity.

The PD interventions (circles) in Figure 4 vary widely in their distance from the line of best fit, with the number of mechanisms explaining only 16% of the variation in effect sizes. Figure 5 sheds light on this by reporting the same analysis separately for each of our three broad categories of PD content. The slopes of the meta-regression lines of best fit are very similar for formative assessment and inquiry-based teaching. The interventions (circles) are also more closely clustered around the line of best fit, with the number of mechanisms able to explain 36% of the variation in effect sizes. By contrast, the line has a slight negative slope for data-driven instruction. Our data do not allow us to investigate the reasons for this. However, we can speculate that data-driven instruction involves making many decisions outside of live lessons (for example, reviewing pupil test data, and assigning pupils to certain intervention groups or activities) and some of our mechanisms (such as feedback on practice) might therefore be less relevant. Alternatively, this could just be explained by the small number of studies (n = 7) in which the PD focused on data-driven instruction. Further research would be needed to understand this. Figure 5 looks almost identical when produced using Hedges' *g* effect sizes and standard errors. In Appendix 6, Figure 13 shows another version of Figure 5 in which the results are reported separately by Kennedy's persistent challenges of teaching. The results look very similar for PD addressing the persistent challenges of portraying curriculum, enlisting participation, and exposing student thinking. However, the gradient is much shallower for the fourth persistent challenge of containing behaviour.

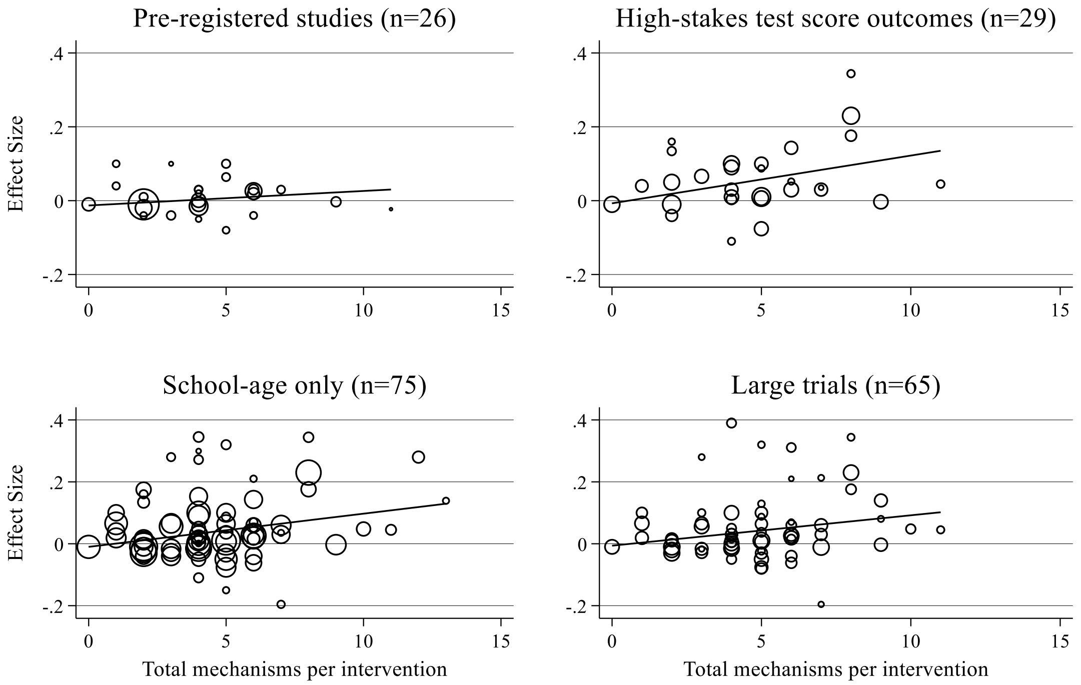
Figure 5: Association between the number of mechanisms included in a PD programme and the impact of the programme on pupil test scores, by area of focus for the PD



Note. n = number of studies. Line of best fit derived from a meta-regression using random effect weights and either the primary outcome as specified in the study, or else one randomly selected outcome per study. Effect sizes >.5 or <-.2 are used in the underlying meta-regression but not shown in the chart in order to aid visual clarity.

Figure 6 further investigates this unexplained heterogeneity based on various sample restrictions. Broadly speaking, we see a similar meta-regression line of best fit among studies reporting high-stakes test score outcomes, among studies with school age children, and in larger studies that randomised 50+ units to treatment and control. This is reassuring, since these three types of studies have all been found to have smaller effect sizes in past research (Cheung & Slavin, 2016). However, in line with our findings in Table 7b (Appendix 6), the meta-regression line of best fit is almost flat among studies that are pre-registered. As can be seen from Figure 5, this largely reflects the narrow range of variation in effect sizes across pre-registered studies. We will return to this important point in our discussion, in Section 6.

Figure 6: Association between the number of mechanisms included in a PD intervention and the impact of the programme on pupil test scores, among different subsamples of studies



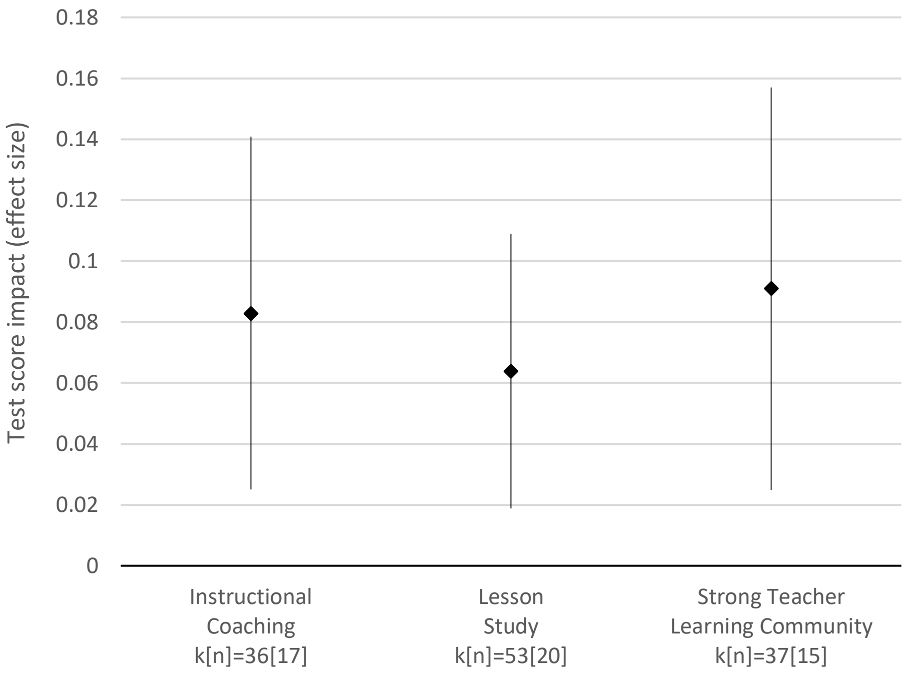
Note. n = number of studies. 'Large trials' involve more than 50 units randomised to treatment or control. Line of best fit derived from a meta-regression using random effect weights and either the primary outcome if specified in the study, or else one randomly selected outcome per study. Effect sizes > .5 or < .2 are used in the underlying meta-regression but not shown in the chart in order to aid visual clarity.

4.2 Impact of different forms of professional development

In Section 2.2, we outlined three different forms of PD, defined as combinations of mechanisms. Having coded each PD programme in our sample for the presence or absence of our 14 mechanisms, we are now also able to identify which interventions constitute examples of each of these three forms. For example, we define an intervention as being 'lesson study' if it includes all of the following three mechanisms: action planning, practical social support, and feedback. In total, 29 of the 104 PD programmes in our sample belong to one or other of these three forms of PD. The other 75 do not belong to any of the three forms we have defined. This is a consequence of the approach we took to defining forms based on terms commonly used in the academic literature. Intuitively, the remaining 75 PD programmes can be thought of as comprising a set of mechanisms for which there is no corresponding commonly-used descriptor in the academic literature.

In Figure 7, we investigate the relative effectiveness of these three forms of PD. The meta-analytic average effect sizes are all very similar, at around .06-.09. This is similar to our sample-wide average effect size of .05 (see Appendix 6). The confidence intervals for each of the three forms are wide and overlap with each other to a large extent—implying no statistically significant difference at conventional levels. However, all three forms of PD have 95% confidence intervals that exclude zero. Taken together, this suggests that PD programmes of any of these forms can be effective, but none are clearly more effective than the others.

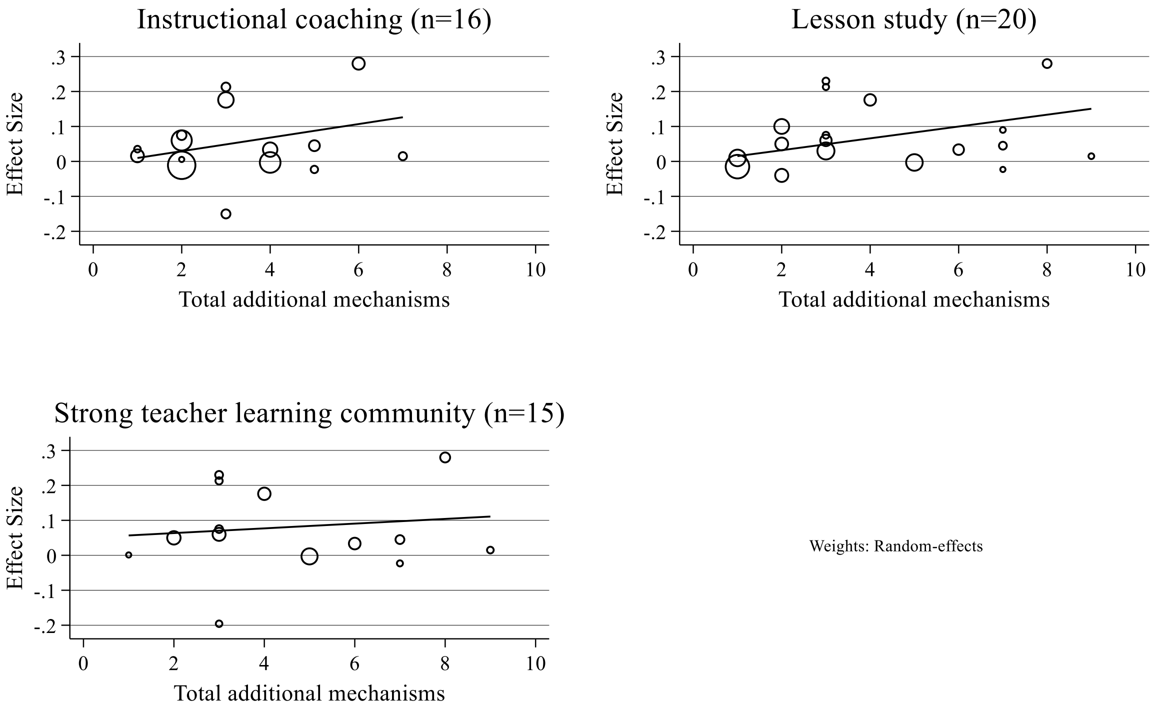
Figure 7: Meta-analytic average impact of different forms of teacher PD



Note. k = number of effect sizes. n = number of separate experimental studies.

Our approach to defining an intervention as belonging to a certain form leaves room for substantial heterogeneity within a given form. For example, 'lesson study' programmes often incorporate other mechanisms in addition to action planning, practical social support, or feedback. To investigate this further, Figure 8 plots the number of 'additional' mechanisms for each intervention against effect size, within each of the three forms. For example, the top right panel looks at the relationship between the number of mechanisms (besides action planning, social support, and feedback) and effect sizes, for all PD programmes which belong to the form 'lesson study'. In all three panels, we observe an upward sloping relationship between the number of mechanisms incorporated in a PD programme and effect size. Indeed, comparing Figure 7 and Figure 8, we find much more variation when comparing PD programmes of the same form with different numbers of mechanisms than we do when comparing PD programmes that belong to different forms. In all three panels in Figure 8, we also see that interventions with zero *additional* mechanisms have an expected effect size just above zero. This suggests that much of the average effect for interventions that belong to any of the given forms may be due to mechanisms that are not themselves integral to that form.

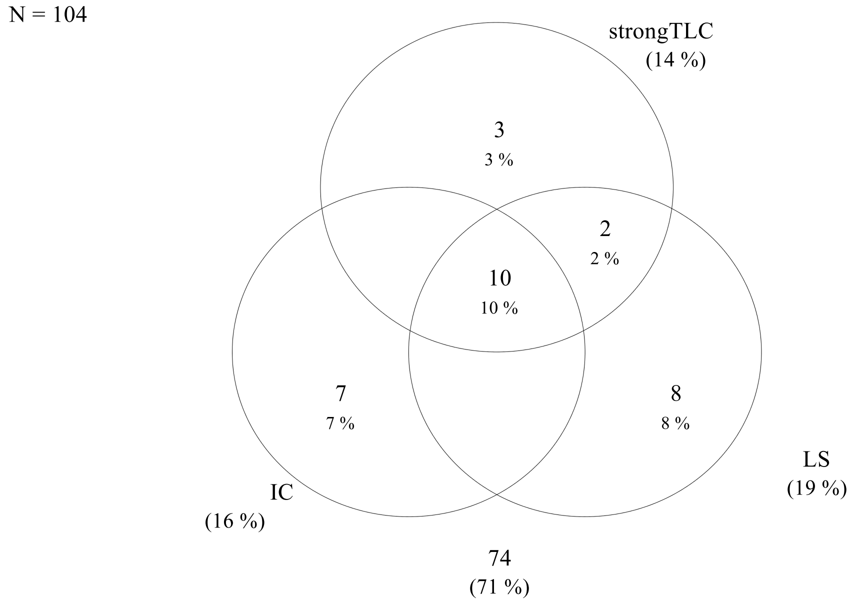
Figure 8: Association between the number of additional mechanisms included in a PD intervention and the impact of the intervention, by PD design



Note. n = number of studies. Line of best fit derived from a meta-regression using random effect weights and either the primary outcome if specified in the study, or else one randomly selected outcome per study. Effect sizes >.5 or < -.2 are used in the underlying meta-regression but not shown in the chart for reasons of visual clarity.

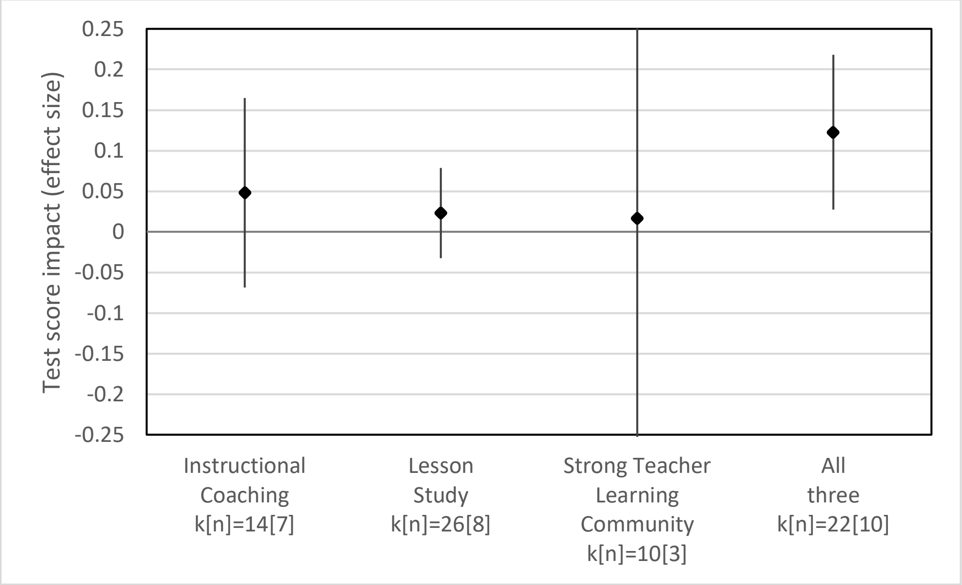
One potential challenge is a consequence of our method of defining forms of PD: it is logically possible that a PD programme can belong to more than one form. In extremis, a PD programme that incorporated all 14 of our mechanisms would belong to all three forms of PD. Indeed, Figure 9 shows that 12 PD programmes belong to more than one form simultaneously. This means that the similarity of the meta-analytic average effect sizes in Figure 7 might be explained by some of the same PD programmes appearing in all three estimates. To overcome this challenge and investigate further, Figure 10 plots the average effect sizes of studies in the individual regions of the Venn diagram separately. That is, it shows as 'lesson study' only those studies which do qualify as lesson study but do not also qualify as instructional coaching or teacher learning communities, and so on. There are three important points to notice. First, the point estimates for the three forms remain very similar to each other and the confidence intervals still overlap to a large extent. Second, the point estimates for the three forms are all now much lower than in Figure 7 (.01-.02) and the 95% confidence intervals include zero. Third, and relatedly, the meta-analytic average effect size for the PD programmes that simultaneously belong to all three forms is markedly higher (.11) and statistically significant ($p < .05$). Taken together, this evidence provides further justification for the argument that mechanisms explain much more variation in effectiveness than forms. It also suggests that average effect sizes from meta-analyses of specific forms (such as instructional coaching) might misattribute the effect size to the instructional coaching, rather than the additional mechanisms that are present alongside the instructional coaching.

Figure 9: Venn diagram showing the frequency of different forms, and combinations thereof



Note. n = 104 studies. Numbers (and percentages) refer to the frequency (and proportion) of PD programmes that fall in each segment of the Venn diagram. Empty segments imply zero PD programmes fall in that segment. The value of 74 (71%) outside the segments indicates the number (and proportion) of PD programmes in the review that do not represent either of these three forms. StrongTLC = strong teacher learning community, LS = lesson study, IC = instructional coaching.

Figure 10: Meta-analytic average impact of different forms of PD (or combination thereof)



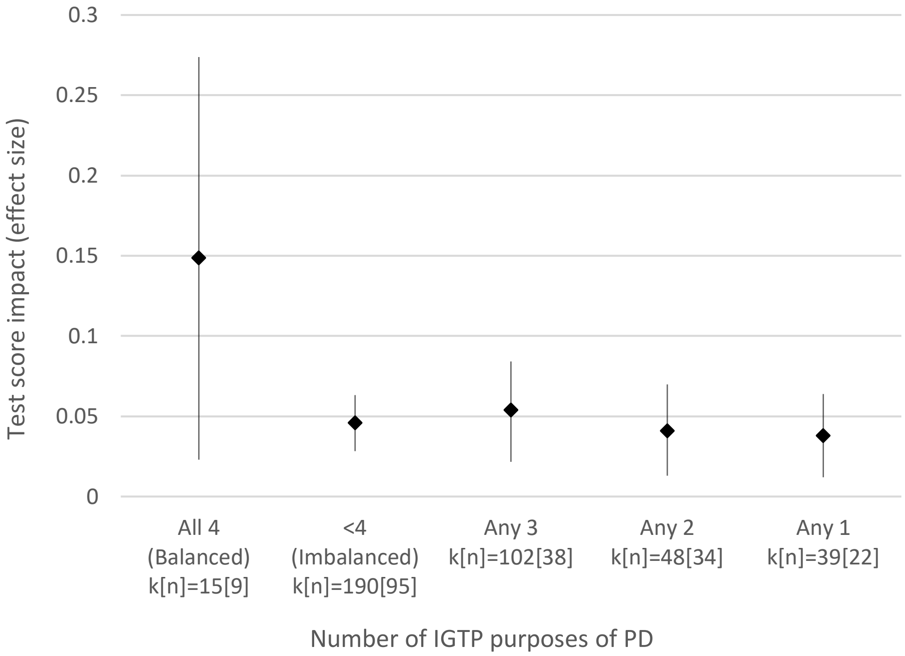
Note. k = number of effect sizes. n = number of separate experimental studies. Produced using random effects meta-analysis, incorporating all standardised test score outcomes using robust variance estimation. Vertical lines represent 95% confidence intervals. The confidence interval for the strong teacher learning community design extend beyond the boundary of the chart. In each category, only studies which fit that form and not the other two forms are shown (as depicted in Figure 9), hence the lower number of included studies than in Figure 8.

4.3 Mechanisms addressing I, G, T and P

In our initial conceptual framework (see protocol and Section 1.2), we theorised that PD programmes are more likely to be effective if they include mechanisms addressing all four purposes: instil insight (I), motivate goals (G), teach technique (T), and embed practice (P). We refer to programmes which address all four purposes as 'balanced' designs. The word 'balanced' implies a design in which none of the four purposes have been neglected. It is not intended to imply that there should be an equal number of mechanisms addressing each of the four purposes. Programmes that address three or fewer of the IGTP purposes are referred to as 'unbalanced' designs.

Figure 11 shows that balanced PD programmes have a point estimate (.15 SD) three times the size of imbalanced PD programmes (.05 SD). There is also no clear upward gradient in the point estimates between imbalanced designs addressing any one, any two, or any three of the four IGTP purposes of PD. This suggests that balanced designs are not associated with higher effect sizes simply because they contain more (at least four) mechanisms. Having said that, the 95% confidence interval for balanced PD programmes is wide, it overlaps with the confidence interval for all imbalanced designs, and the difference is not statistically significant at conventional levels (*p* =.22). Finally, it is also noteworthy that the 95% confidence intervals for all of the imbalanced categories also exclude zero. Reproducing Figure 11 using Hedges' *g* effect sizes results in an almost identical graph (not included in this report).

Figure 11: Meta-analytic average impact depending on whether PD programmes incorporate mechanisms addressing the IGTP purposes of PD

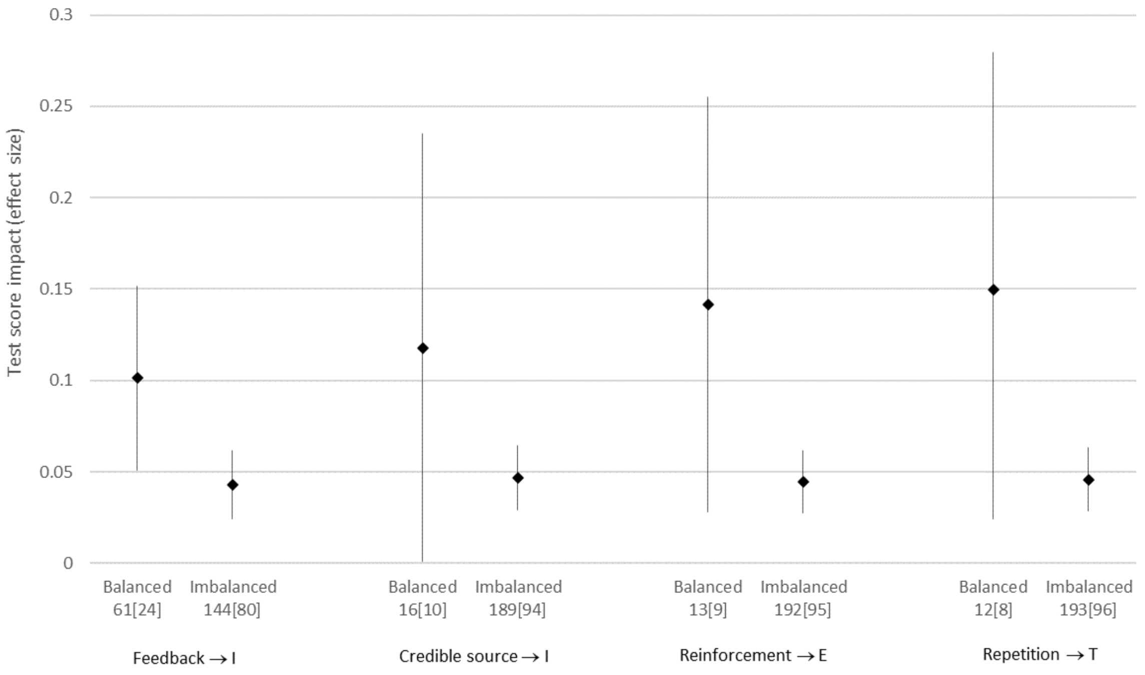


Note. k = number of effect sizes. n = number of separate experimental studies. Produced using random effects meta-analysis, incorporating all standardised test score outcomes using robust variance estimation. Vertical lines represent 95% confidence intervals.

One concern with Figure 11 is that the balanced designs might have higher average effect sizes due to methodological artefacts, rather than because the PD is more effective. For example, we might be worried that evaluations of balanced designs tend to have used more low-stakes (as opposed to high-stakes) test score outcomes, have higher levels of attrition (missing outcome data), or were less likely to be pre-registered. We can explore these three indicators across all 104 studies in our data. First, balanced designs are actually less likely to be evaluated using low-stakes exams (55.6%) than imbalanced designs (73.7%). Second, evaluations of balanced designs are more likely to show 'unacceptable' (based on What Works Clearinghouse standards) levels of attrition (89.9%) versus imbalanced designs (63.21%). Third, evaluations of balanced and imbalanced designs are approximately equally likely to pre-register their analysis (22.2% and 25.2% respectively). In sum, while there is some concern that evaluations of balanced designs are more likely to show high attrition, taking the three indicators together, there is no clear evidence that evaluations of balanced designs are likely to be inflated relative to evaluations of imbalanced designs.

As discussed in Section 2.1, the results in Figure 11 may depend in part on assumptions we have made about how to allocate mechanisms to the four IGTP purposes of PD. In order to test this, we conducted sensitivity analysis in which we reallocated four mechanisms for which the initial allocation appeared debatable. First, we reallocated the feedback mechanism to the insight (I) purpose. Second, we reallocated the credible source mechanisms to the insight (I) purpose. Third, we reallocated the praise/reinforce mechanism to the embed practice (P) purpose. Fourth, we reallocated the context-specific repetition mechanism to the embed practice (P) purpose. Figure 12 shows the results. The results are qualitatively similar to those in Figure 11, which suggests our findings are not overly sensitive to assumptions we have made about allocating mechanisms to the four IGTP purposes of PD.

Figure 12: Sensitivity analysis of the meta-analytic average impact depending on whether PD programmes incorporate mechanisms addressing the IGTP purposes of PD



Note. k = number of effect sizes. n = number of separate experimental studies. Produced using random effects meta-analysis, incorporating all standardised test score outcomes using robust variance estimation. Vertical lines represent 95% confidence intervals.

We also investigated the relationship between effect sizes and whether interventions incorporated mechanisms addressing the various IGTP categories using qualitative comparative analysis (QCA). This involves invoking assumptions in order to dichotomise effect sizes into substantively important and unimportant. As set out in our protocol (Sims et al., 2021), we use *d* = .10 as our main cut-off, because it is approximately equivalent to the additional learning that would be necessary for an average student to overtake one other pupil in a classroom of average size.[8] However, we also report sensitivity tests for cut-off values of *d* = .08 and .12. This allows us to calculate the proportion of interventions addressing different combinations of the IGTP purposes that have effect sizes larger than these three thresholds. This proportion is known as the 'consistency'.

The added value of conducting a QCA analysis is that it provides information about PD programmes with all possible combinations of I, G, T and P, even in a sample of only 104 studies. By contrast, using meta-regression to analyse all possible combinations of I, G, T and P would require a model with many interaction terms, which would require a sample with more than 104 studies. The QCA is also answering a subtly different question to the meta-analysis, in that it provides information about necessary conditions for impact, rather than correlates of impact. Relatedly, the Quine-McCluskey algorithm used in QCA helps identify any of I, G, T and P that are causally redundant (not necessary) for effective PD. An important limitation of QCA is that it does not account for random error in the RCT impact estimates.

[8] In statistical terms, an effect size of 0.10 means that the average pupil in the intervention group would score higher than 54% of pupils in the control group on the achievement outcome measure.

Table 9 shows the results. Each row represents a different combination of IGTP purposes of PD, with small letters indicating that a given purpose is not addressed by any mechanism and large letters indicating a purpose that is addressed by at least one mechanism. For example, igTP (row 3) including all PD programmes that include mechanisms addressing T and P but no mechanisms addressing i or g. Using our preferred cut-off ($d \geq .10$), balanced designs have a consistency of 67%. To put it another way, in the final row of our table, there are nine studies with a balanced design and six of them have effect sizes higher than .10. Balanced designs have a higher consistency than any of the other (imbalanced) designs. The pattern of results are similar using the other two cut-off values (.08 and .12), with balanced designs always having the highest levels of consistency. Having said that, the consistency for balanced designs is always lower than the conventional QCA consistency criteria of 80%, which means the Quine-McCluskey algorithm cannot be used to test for causal redundancy.

Table 9: Qualitative comparative analysis—consistency measures

	Cut-off = 0.08	Cut-off = 0.10	Cut-off = 0.12
Igtp [n=1]	0%	0%	0%
igTp [n=22]	31.8%	31.8%	13.6%
igTP [n=11]	27.3%	27.3%	18.2%
iGTp [n=11]	27.3%	18.2%	18.2%
iGTP [n=22]	36.4%	31.8%	31.8%
IgTp [n=11]	54.5%	45.5%	45.5%
IgTP [n=10]	50%	50%	50%
IGtp [n=1]	0%	0%	0%
IGTp [n=6]	16.7%	16.7%	16.7%
IGTP [n=9]	66.7%	66.7%	55.6%

Note. n = number of separate experimental studies in each row. N = 104 studies in the table in total. Each study uses either the primary outcome if specified, or else one randomly selected outcome per study.

Research Question 3 Which forms of PD are associated with the greatest impact?	We identified three forms of PD in Section 2.2: lesson study, instructional coaching, and teacher learning communities.
a. Which forms (clusters of mechanisms) do we observe in the literature?	Programmes which address all four purposes are on average more effective: the point estimate for programmes addressing all four purposes (.15) is three times that for programmes addressing three or fewer (.05). However, the 95% confidence interval for programmes addressing all four purposes is wide and the difference in impact between balanced and imbalanced designs is not significant at conventional levels ($p = .22$).
b. Are forms more likely to be effective when they incorporate mechanisms addressing all four purposes: instil insights (I), motivate goal-directed behaviour (G), develop techniques (T), and embed practice (P)?	In general, we find that PD which incorporates more mechanisms is associated with greater impact on test scores. Programmes incorporating zero mechanisms have an expected effect size close to zero. PD

	mechanisms incorporating 14 mechanisms have an effect size close to .17.
c. Which forms (clusters of mechanisms) are associated with the largest effects on teacher practice and pupil achievement?	Across the three forms that we identify (instructional coaching, lesson study, and strong teacher learning communities), the average effect sizes are similar. The confidence intervals for each of the three forms also overlap to a large extent. This suggests that none of these three forms is clearly more effective than others.

Part 5: Review of implementation

Alongside our review of PD impact, we examined the implementation of PD programmes. A teacher development programme cannot succeed without an appropriate theory of student learning and an appropriate theory of teacher change (Wayne et al., 2008). But, no matter how good the intent and design, the implementation of the programme makes a substantial difference to the outcomes it achieves (Durlak & DuPre, 2008; O'Donnell, 2008). The study of implementation allows evaluators to identify critical intervention components, identify the support needed for the programme and for scale up, and provide formative feedback (Lendrum & Humphrey, 2012). While our understanding of active ingredients increases, we also need to look at the complex factors within an intervention which determine whether or not they have the desired effect (Karanika-Murray & Biron, 2015). Examining implementation also permits evaluators to explain how and why a trial has succeeded, if a positive outcome is found (Humphrey et al., 2016). This section uses the Implementation and Process Evaluations from a selection of the included studies to examine the extent to which PD programmes are implemented with fidelity, as well as the key supports and barriers.

5.1 Background

The study of implementation focuses on the gap between intentions and policies on one hand, and what happens on the ground on the other. Studying implementation provides conclusions which are often both obvious and frequently overlooked (Pressman & Wildavsky, 1984). If an impact evaluation shows that an intervention works, an implementation and process evaluation shows how and why it works (Humphrey et al., 2016). This permits the identification of critical components and the scale-up of successful interventions, provides formative guidance to PD providers, and avoids false attributions of success (Lendrum & Humphrey, 2012). Our work on mechanisms helps to show how and why PD trials work, but a review of implementation and process evaluations provides further insight. Eight dimensions are usually suggested for implementation and process evaluations: fidelity, dosage, quality, reach, responsiveness (of recipient), programme differentiation, control group activity, and adaptation (Durlak & DuPre, 2008). Much school implementation literature focuses on prevention programmes, designed to discourage teenagers from drugs, smoking, drinking, and gang involvement (for example, Dariotis et al., 2008; Domitrovich et al., 2008). We are unaware of any published work synthesising the findings of implementation and process evaluations from PD programmes. This section of the report should therefore be seen as offering an initial overview of these evaluations.

5.2 Purpose

We focused on three questions:

1. To what extent were interventions implemented as planned?
2. What factors supported or obstructed effective implementation?
3. What was the nature and effect of programme adaptations?

Of the eight dimensions of implementation (Humphrey et al., 2016), our first research question incorporated fidelity, dosage, quality, reach, and responsiveness, and our third studied adaptations. We did not seek to address programme differentiation and monitoring of control groups in this study.

5.3 Data

Of the 121 studies in our final sample,[9] 71 (59.7%) had some form of process evaluation. A review of schoolwide programmes published between 1993 and 2012 found 46% contained quantitative data on fidelity (Bruhn et al., 2015). Since the earliest study in our sample was published in 2007, and the most recent in 2020, this suggests that implementation evaluation has become more prevalent in recent years.

In the protocol (Sims et al., 2021), we specified that we would randomly select 20 studies with Implementation and Process Evaluations (IPEs) from our final sample of studies. We randomly ordered all studies with implementation data. However, we found that the vast majority of such studies included only limited fidelity data. All those studies featuring detailed IPEs were those funded by the EEF. Since 22 EEF studies remained in our sample, we chose to focus our study on them. We believe this is particularly appropriate, since some elements of the supports and barriers to implementation are likely to reflect English educational policies and practices (Lendrum & Humphrey, 2012). Within this sample, we found that one study (Torgerson et al., 2014) contained only fidelity data, so this was excluded, leaving a final sample of 21 studies.

Synthesis across IPEs is challenging because aspects of what is measured are specific to the programme (Lendrum & Humphrey, 2012). There is substantial variety in the approach to data collection and analysis in the studies in our sample. This is because 'There is no single, universally agreed way to conduct an IPE' (Humphrey et al., 2016, p.3). In particular, older EEF IPEs show substantial variation in their approach (Humphrey et al., 2016). This can be illustrated simply by examining the lengths of IPEs in this sample, which range from five pages (Hanley et al., 2016, IEE, 2016) to 27 (Humphrey et al., 2018). Likewise, the approaches to data collection vary. Some evaluators relied primarily on structured surveys and administrative data (for example, Kitmitto et al., 2018), and some relied on case studies (McNally et al., 2014). Others used observations, interviews, or combinations of these approaches. Similarly, the conceptualisation of fidelity varies substantially depending on the implementation design. Moreover, some evaluators also used data which became available—for example, from emails sent to the PD provider (Humphrey et al., 2018) or from impromptu focus groups (Sutherland et al., 2019).

Beyond the variety of the data collected, an additional challenge is that the respondents are often self-selecting. The implications were described well by one evaluation report, which noted:

> 'The findings reported in this section draw largely on survey, focus group, and interview findings, which represent respondents' self-reports and perceptions and may not be representative of the sample as a whole. We also have to acknowledge that the surveys are not random samples, hence selection bias may influence responses. For example, those most motivated about Digital Feedback may have been those most willing to respond to the survey. Furthermore, the surveys did not track respondents across time-points and therefore should not be interpreted as longitudinal.' (Sutherland et al., 2019, pp.38–9)

The implications of these restrictions on data collection are not always clear. It is possible that this overestimates fidelity, since successful schools and teachers are more likely to respond or to participate in a case study (Tracey et al., 2019, p.58). Alternatively, another evaluation suggested that

[9] Note that this is larger than our analytical sample, since it includes studies which met our inclusion criteria, but from which we were unable to extract an effect size.

survey responses were ‘quite polarised’: participants were motivated to respond by either very good or very bad experiences (Biggart et al., 2015, p.23).

Since different evaluators measured different aspects of implementation in different ways, we concluded that attempting to compare incommensurable quantitative measures of fidelity was unlikely to be fruitful. We therefore focused on the occurrence of an issue within a case (that is, a study) as our primary unit of analysis. Our assumption in doing so was that all crucial issues—such as barriers to implementation—were likely to be mentioned at least once in a case study, irrespective of the method of research used. We excluded from our coding any references to opinions held by isolated individuals, unless they were described as characteristic of the group. For example, having explained that multiple schools had adopted cooperative learning, one evaluation then stated that ‘one teacher felt that this particular element was burdensome’ (Biggart et al., 2015, p.23). The evaluator was not suggesting that this was a broadly representative opinion, so it was not coded.

5.4 To what extent were interventions implemented as planned?

Fidelity

We coded the studies for evidence that the intervention was implemented as planned, or it was not implemented as planned. The latter incorporates participants not following the intervention as planned, and participants adapting it.

Table 10: Fidelity

	Number of studies
Faithful to plans	19 (95%)
Deviation from plans	18 (86%)

Note. Categories are not exclusive, so may sum to more than 21. This is because an evaluation may refer to elements of both faithful implementation and deviation from plans.

It appears there were substantial variations to plans. References to faithful implementation were rarely unmitigated. For example:

- ‘Pupil observations were carried out in most, although not all, of the schools visited.’ (McNally et al., 2014, p.30)
- ‘Of the 39 schools that attended and completed the training programme, 21 submitted an action plan that was judged to be compliant.’ (Culliney et al., 2019, p.34)
- ‘Good fidelity was defined as the nursery being rated by the NDNA as being very or partially engaged in all of the core components of the intervention … A third (n=18) of the 54 settings randomly allocated to the intervention group achieved a good fidelity rating.’ (Robinson-Smith et al., 2018, p.57)

Even apparently strong statements of fidelity suggest substantial adaptation. For example, one evaluation reported that most changes ‘were minor, with the majority of respondents involved in delivery reporting that they had changed the content of the lesson very little’ (Foliano et al., 2019, p.32). A review of implementation evaluations noted as an established finding that: ‘The degree of achieved implementation is almost never 100%’ (Durlak, 2010, p.350). We must repeat that these measures are far from comparable, and coded references varied from ‘anecdotal evidence from the school visits interviews that a good level of implementation was maintained’ (IEE, 2016, p.28) to detailed quantitative measures. Nonetheless, it is clear that even faithfully-implemented programmes see substantial adaptation and deviation from plans.

Types of deviation from programme design

We examined deviations from plans more closely, dividing them into four categories:

1) Changes to the intended strategies/techniques experienced by students.
2) Changes to the intended teacher learning activities.
3) Not implementing the intervention—for example, not using the intended strategies in the classroom.
4) Not participating in the intervention—for example, dropping out of a trial, not attending training.

Table 11: Deviations from programme design

Type of deviation	**Number of studies**
Changes to student experience	14 (67%)
Changes to teacher learning	4 (19%)
Not implementing	7 (33%)
Not participating	4 (19%)

Note. Categories are not exclusive, so may sum to more than 21. This is because an evaluation may refer, for example, to changes to student experience and changes to teacher learning.

The most common changes were made to student experience, with changes occurring in 14 (67%) of the studies. These included changes to the activities used in lessons (Foliano et al., 2019), such as omitting some activities (Tracey et al., 2019), and changing which students participated (Hanley et al., 2016). Most often, teachers adapted the intervention by limiting how much they did, and focusing on the elements they perceived as most important. Four evaluations reported changes to teacher learning. These are less common, in part because they were only possible where an intervention planned school-led teacher learning as part of the intervention (for example, Speckesser et al., 2018). Where all training was designed and led by the intervention team, substantial changes were rare. Seven evaluations (33%) collected explicit evidence of teachers simply not implementing changes. For example, a trial of digital feedback saw *decreasing* fidelity over time (Sutherland et al., 2019).

The causes of these adaptations varied substantially. In some cases, schools questioned the relevance of the intervention to their work (McNally et al., 2014). In some, schools simply did not comply with measures of fidelity established by intervention teams (Culliney et al., 2019). Finally, some teachers did not participate and missed the training (Sutherland et al., 2019). Some schools withdrew from trials (for example, 24% of schools in the Good Behaviour Game intervention—Humphrey et al., 2018). We examine barriers to successful implementation and the nature and effect of adaptations in more detail below. This finding echoes the wider literature. For example, a study comparing the implementation of youth prevention programmes in schools, communities, and families found that the greatest degree of variation from plans lay in schools (Dariotis et al., 2008).

5.5 What was the nature and effect of programme adaptations?

Reason for adaptation

Both fidelity and adaptation can contribute to the success of a programme (Lendrum & Humphrey, 2012). However, to understand the consequences of an adaptation, we can examine its nature and effect. Adaptations can be divided by the grounds on which they are made: whether they are made for logistical or philosophical reasons (Moore et al., 2013). When an adaptation is made for logistical reasons, a practical barrier has emerged to the teacher or student acting as intended, such as a lack of resources or time. For example, teachers made logistical adaptations where they omitted planned elements of a lesson due to lack of time (Hanley et al., 2016, p.23), or adapted lessons to keep pupils engaged (Styles et al., 2014, p.25). When an adaptation is made for philosophical reasons, this reflected more deep-seated challenges to the underlying logic of an intervention, based on teachers' beliefs about learning, and their experience. For example, teachers (and leaders) made philosophical adaptations where they shifted an intervention intended to be completed in cross-curricular groups to departmental ones, because teachers wanted to work together to apply new strategies in subject-specific ways (Speckesser et al., 2018). Distinguishing between these two forms of adaptation can be tricky. In some cases, the reasons for adaptation were unclear (these cases are excluded from the table and discussion below). Moreover, a teacher might describe a practical barrier in philosophical terms, or vice versa. Nonetheless, distinguishing between logistical and philosophical adaptations seems valuable in helping to distinguish between barriers which providers can address through changing practical arrangements for a programme, and more substantial barriers to teacher engagement.

Table 12: Types of adaptation

Type of adaptation	Number of studies
Logistical	11 (52%)
Philosophical	8 (38%)

Note. Categories are not exclusive, so may sum to more than 21. This is because an evaluation may refer to both logistical and philosophical adaptations.

More studies saw interventions adapted for logistical reasons than for philosophical reasons. This echoes a finding from American schools that two thirds of adaptations were made for logistical reasons, and one third for philosophical reasons (Moore et al., 2013). This is understandable, given the substantial constraints teachers are under: the same study found the most common reason teachers gave for making adaptations was time. Nonetheless, it implies that organisations may wish to prepare for such adaptations, and discuss how best to ensure they succeed. We found three studies which had discouraged adaptation in some way, and ten that had encouraged it in some way. In some cases, adaptation was encouraged or sanctioned by project teams, provided it was in line with the 'spirit' and 'essence' of the intervention (Speckesser et al., 2018). It is also possible that teachers were less explicit about philosophical reasons for their changes than logistical ones, as they sought to respond politely or constructively in interviews or focus groups. For example, as part of the Good Behaviour Game (Humphrey et al., 2018), teachers are intended not to support pupils while playing the game (to encourage peer support and autonomy). Teachers reported not following this for logistical reasons (because some students struggle without support), but this may also be a philosophical scepticism about the intended approach. We suggest that all interventions need to consider which adaptations they would be happy to encourage/discourage, and how best to introduce them.

Effect of adaptation

We planned to examine the effects of adaptations. However, we found it difficult to draw robust conclusions about the actual effect on student learning of these adaptations. The precise impact of a specific adaptation was rarely clear, to teachers or evaluators. We identified some adaptations which appeared to build positively upon the intended intervention, such as:

> '… small tweaks to the content to enhance its relevance to pupils, or to make the activity more practical or engaging for pupils to complete. This included relating the content to pupils' upcoming SATS exams and turning whole class exercises into paired or group exercises where

> more children had the opportunity to contribute their thoughts or ideas to the task at hand.' (Foliano et al., 2019, p.33)

Similarly, mentors were described as having:

> 'adapted the timings and content of the intervention to create more flexibility. For example, one mentor reported 'adapting' and 'filtering' parts of the programme to make engagement more manageable for teachers.' (Jay et al., 2017, p.33)

Such adaptations appeared to be widespread. For example, in survey evidence collected in the Grammar for Writing effectiveness trial, 35% of teacher respondents stated that they had made adaptations to meet the ability levels of pupils (Tracey et al., 2019, p.56). Similarly, some teachers applied the tools from the Digital Feedback intervention (videos for students) to explain complicated concepts in advance (Sutherland et al., 2019, p.44).

We also identified some adaptations which may have limited the effect of the intended intervention. For example, teachers met less frequently (e.g. Jay et al., 2017; Speckesser et al., 2018) or did not conduct specific elements of the programme, such as peer observations (Speckesser et al., 2018). Similarly, teachers often limited the extent to which students were exposed to interventions—for example, by combining planned lessons (Tracey et al., 2019) or omitting planned elements of the lesson (Gorard et al., 2015).

Our uncertainty about the effects of these adaptations means our conclusions must be tentative. For example, teachers may have missed out through meeting less frequently than intended, but more frequent meetings may have been unnecessary or impossible. In the broader literature, there is a growing consensus that principled adaptation can increase the impact of a programme (Durlak & DuPre, 2008; Lendum & Humphrey, 2012). Additionally, teachers in England want flexibility in implementing interventions (Lendrum and Humphrey, 2012). However, better-implemented programmes (which usually means more faithfully-implemented programmes) tend to have a greater impact (Durlak & DuPre, 2008). Moreover, it is not always the case that adaptations *are* principled or planned. Providers may wish to encourage tweaks to tailor the programme to teachers' and students' needs, and to discourage the omission of crucial elements of the programme. Most adaptations are made with good intentions—but that does not mean they always have desirable consequences (Fagan & Mihalic, 2003). The challenge is to find the right balance between adaptation and fidelity (Durlak & DuPre, 2008). Programme guidance rarely specifies what can be adapted (Dane & Schneider, 1998), so offering some guidance around the core elements of the programme and underlying mechanisms may make it easier for teachers to make positive adaptations.

5.6 What factors supported or obstructed effective implementation?

Various models have been advanced to model the elements contributing to effective implementation (for example, Durlak & DuPre, 2008). We adopted the division used by Moore et al. (2008), between:

- intervention design—the mechanisms by which the intervention sought to influence teacher practice and student learning;
- intervention support system—the training, resources and support offered by the developer;
- school context—for example, time, leadership support, competing priorities; and
- teachers—teachers' reactions to the intervention.

These distinctions are artificial: a teacher's experience of a training event is a response to the intervention design and how it is introduced (the support system). However, they help us to isolate supports and barriers: a strong intervention design may falter due to a weak support system if, for example, resources are delivered late or incomplete.

Table 13: Elements acting as barriers and supports

Elements	Number of studies mentioning them as a barrier or support
Intervention design	20 (95%)
Support system	19 (90%)
School context	19 (90%)
Teachers	11 (52%)

Note. Categories are not exclusive, so may sum to more than 21. This is because an evaluation may refer to multiple barriers and supports.

Perhaps unsurprisingly, all these elements play a role in the success of the intervention. Teachers are cited least often, but this is partly due to the nature of the data. Most emerges from teachers' responses (to surveys, focus groups, interviews), so they are describing their experience. Where teachers were described as a barrier or support to an intervention, this was often in terms of describing their response to other elements of the intervention. For example, teachers cited personal commitments preventing them from attending training, as well as ambivalence about what was offered (Sutherland et al., 2019), doubts about the value of the intervention (Humphrey et al., 2018), and concerns about their absence from the class (Murphy et al., 2017). Breaking these elements down further offers greater insight.

Intervention design

Three substantial themes emerged: pitch, value, and feasibility (see Table 14).

Table 14: Intervention design—themes of barriers and supports

Barrier/support theme	Number of studies mentioning this theme as affecting implementation
Pitch	15 (71%)
Value	18 (86%)
Feasibility	12 (57%)

Note. Categories are not exclusive, so may sum to more than 21. This is because an evaluation may refer to more than one theme.

The most important element of an intervention that helped cause success was demonstrating value to teachers. This was raised in 18 of the 21 studies. Broadly, this meant providing teachers with tools or opportunities which helped them improve, and which they would otherwise lack. In particular, teachers valued time to meet, discuss teaching, and plan with peers (Murphy et al., 2017; Speckesser et al., 2018), and the chance to gain fresh insight into students' needs (McNally et al., 2014). They also valued tools which would help them meet students' needs or developed students' capabilities (Gorard et al., 2015; Humphrey et al., 2018). In some cases a study was immediately appealing. For example, teachers saw the efficacy trial of Changing Mindsets as 'appealing and attractive and readily adopted what they saw as its key features' (Rienzo et al., 2015, p.31). Similarly, 90% of survey respondents saw the Reciprocal Reading programme as necessary (O'Hare et al., 2019).

In other cases, however, teachers recognised the value of the programme only with experience. For example, in the effectiveness trial of Changing Mindsets, it was reported that 'some members of staff were initially resistant to these ideas [but], as they learnt more about this approach and came to see its potential benefit for pupils, they became receptive to changing their working practices' (Foliano et al., 2019, p.36). Teachers also cited students valuing the intervention as reasons for its success, in terms of student enjoyment of the activities (for example, Biggart et al., 2015; Gorard et al., 2015). Where teachers questioned the value of a programme, this could reflect having tried it, but concluded it was unhelpful, such as the digital feedback intervention, which was seen as distracting students (Sutherland et al., 2019). Teachers also struggled with interventions where there was a lack of apparent clarity and direction in the intervention (for example, McNally et al., 2014; Sutherland et al., 2019). Facilitating teacher buy-in is often cited as an important factor in the success of PD (for example, Cordingley et al., 2015). These findings suggest that demonstrating value by helping teachers improve, or enjoy their work more, is a crucial element of this. It is also noteworthy that there are multiple ways in which a programme might achieve this, including an immediate appeal, tools to help teachers meet students' needs, and an experience of success.

Unsurprisingly, the success of an intervention was also influenced by how feasible it proved to be—a point raised in 13 studies. One element of this was alignment. Where an intervention was seen as aligned to the school's existing priorities and needs, it was welcomed. For example, Embedding Formative Assessment was seen as 'about embedding good practice and enhancing what teachers were already doing' (Speckesser et al., 2018, p.26). Conversely, where an intervention did not meet the school's current priorities, implementing it was seen as impractical or unnecessary (for example, Humphrey et al., 2016; McNally et al., 2014). We discuss alignment further as a reflection of school context below. The time demands that the intervention places on teachers was another important issue. Again, this is discussed further below, since the issue is one of finding time among the many priorities teachers face. Finally, flexibility was often described as making an intervention workable or feasible. Teachers welcomed approaches which did not prescribe a particular teaching behaviour, but instead offered a method to improve teaching and learning (for example, Murphy et al., 2017; Speckesser et al., 2018). They also welcomed the ability to adapt programmes to meet students' needs (for example, Styles et al., 2014), and disliked programmes they deemed 'too prescriptive' (Biggart et al., 2015, p.24). Overall, the feasibility of an intervention in a school seemed to be a function of the interaction between that intervention and the school's current situation.

A specific element of the feasibility of the programme which emerged as particularly important was the pitch of the programme: how realistic expectations were. For many teachers, the issue was whether the intervention was accessible and worthwhile for students. This might include the roles students were expected to take on (O'Hare et al., 2019) and their demands for sustained attention (Foliano et al., 2019). It also depended on whether activities were suitable for all students, particularly those with lower prior attainment (Kitmitto et al., 2018). In some cases, teachers were concerned that intervention designs obstructed key elements of teaching and learning, like student/teacher interaction. For example, the Good Behaviour Game asked teachers not to interact with students to help them during the 'game' (lesson time). This was seen as 'impeding the extent to which the teacher could aid their academic progression' (Humphrey et al., 2018, p.43). More common were concerns that interventions demanded additional planning time, which was seldom provided. This might mean preparing resources (Sloan et al., 2018) or spending additional time planning, particularly when an intervention was new to teachers (Humphrey et al., 2018). The PD activity expected of teachers could also be a barrier. For example, the demands of catching up with the class's needs after a training session and participating in the Integrating English programme were considered by many respondents to be 'too much for a practising teacher' (Culliney et al., 2019, p.36). Interventions were more welcome, and effective, where intervention designers were able to limit these concerns—for example, by providing easily usable resources and making limited demands on teachers (e.g. Kitmitto et al., 2018). We return to the question of time pressures on teachers below.

Support system

Resources were the most frequently referenced element of the support system, mentioned in 18 evaluations. Resources took three crucial forms: guidance documents, teaching resources (such as lesson plans or resources for students), and technological resources (such as iPads or computer programmes). The unifying theme across these categories was clear: teachers welcomed resources which met their needs simply and effectively but abandoned resources which did not do so. The precise nature of the resource depended on the intervention. Teachers welcomed the laminated Dialogic Teaching Repertoires sheet of nine key talk moves and it was 'regularly singled out as being the most useful and consistently used' resource (Jay et al., 2017, p.32). For a more directed intervention, Changing Mindsets, they appreciated the comprehensive manual: 'all the materials and resources were there that they needed to deliver the sessions with very little preparation required' (Foliano et al., 2019, p.33). However, criticisms were more common than praise. In some cases, resources proved hard to use: for example, 'the lengthy process of stamping individual books [in the Good Behaviour Game] was deemed to be so time consuming that it inadvertently defeated the purpose of the game' (Humphrey et al., 2016, p.52). This problem was particularly common for technological resources: most references to technological resources were negative. This included difficulty accessing videos (Foliano et al., 2019), cumbersome and time-consuming feedback apps (Sutherland et al., 2019), and a virtual learning environment for teachers which was 'not user-friendly' (Culliney et al., 2019, p.36). Resources can be valuable in guiding teachers' actions, and are sometimes essential for a programme. If teachers are to use them—and to implement the intervention effectively—they must be as easy to use as possible.

Table 15: Support system—themes of barriers and supports

Barrier/support theme	**Number of studies mentioning this theme as affecting implementation**
Guidance and communication	11 (52%)
Resources	18 (86%)
Teacher educators	11 (52%)
Training events	16 (76%)

Note. Categories are not exclusive, so may sum to more than 21. This is because an evaluation may refer to multiple themes.

The next most common element of the support system mentioned was training events, which were mentioned as supports or barriers in 16 evaluations. Generally, reactions to training events were positive: several evaluations described initial events as enthusing. Embedding Formative Assessment's launch was described as contributing to 'a high-buy-in and enthusiasm among participants' (Speckesser et al., 2018, p.24). Similarly, RISE's launch left participants 'very positive about the skills they acquired, the learning from peers, and the expertise of the trainers' (Wiggins et al., 2019, p.41). Beyond inspiration, teachers particularly valued practical support: 'the opportunity to share experiences and good practice, and gain further ideas for activities and use of resources' (Sloan et al., 2018, p.29). Sometimes, positive initial responses to training belied subsequent confusion which training had not addressed (Sutherland et al., 2019). Where teachers were critical of training events, this was usually because the training failed to solve the practical problems they faced. For example, teachers disliked training that 'focused too heavily on the theoretical principles' (Humphrey et al., 2018, p.58). Similarly, they stated that they would have liked 'greater sharing of ideas of how to implement [Embedding Formative Assessment] and overcome specific barriers faced in schools' (Speckesser et al, 2018, p.24). Follow-up training received more mixed reactions: in some cases it was seen as vital to build on teachers' initial experience and clarify key points (for example, O'Hare et al., 2019, p.47). In other cases, teachers were unsure whether they had learned more, or whether the time had been well spent (Humphrey et al., 2018). These findings are interesting in the light of the review of evidence reported above. They primarily focus on teachers' immediate reactions to the training, and there is little mention of the mechanisms which make training effective. In doing so, they underscore the importance of offering teachers practical guidance which addresses their problems, and the value of an enthusing initial training experience.

The support provided by teacher educators was referred to in 11 evaluations. (Here we are referring to teacher educators working for the developer; the role of school leaders is discussed below in the 'School context' section.) Most evaluations described the support offered by teacher educators in positive terms. For example, TEEP training sessions were 'delivered skilfully' (IEE, 2016, p.27), while observers 'returned from [Integrating English] training days with praise for the professionalism, relevance, and supportiveness of the trainers and the programme' (Culliney et al., 2019, p.34). Teachers were often able to access ad hoc support where needed, and this was usually satisfactory. For example, Embedding Formative Assessment participants 'emphasised that there had been open lines of communication, and that they had regularly kept in contact via emails and phone conversations, and the schools knew where to go if there had been any problems' (Speckesser et al., 2018, p.26). In some interventions, external coaches were able to build strong relationships with teachers, challenging and supporting them to improve (Humphrey et al., 2018). Where concerns were expressed, they often arose because teacher educators were unable to reconcile the intervention design with the school's needs. For example, where schools had concerns about turnover of project staff (Murphy et al., 2017) or the applicability of the intervention to their context (Humphrey et al., 2018), their satisfaction hinged on teacher educators' success in resolving these concerns.

The importance of clear guidance and effective communication was mentioned in 11 studies. Teachers wanted to know exactly what they needed to do. In particular, they wanted models, such as 'examples of how schools had previously modelled the programme' (Speckesser et al., 2018, p.25). Programmes which did not offer this clear guidance struggled to gain traction: participants came to feel 'there was a lack of clarity and direction' in the Hampshire Hundreds project (McNally et al., 2014, p.33). Similarly, teachers left Digital Feedback training unsure about 'implementation details', such as 'how to implement the intervention when a year group has stream classes or mixed age groups' (Sutherland et al., 2019, p.40). Sometimes guidance was needed around nuances in programme intervention, such as adaptations to meet students' needs at different ages (O'Hare et al., 2019). Poor communication could leave participants unclear about what was intended or required. For example, some participants reached training sessions unclear about 'what the training would be about (its main goal), or what they would be committing to prior to joining the session' (Sutherland et al., 2019, p.39). A lack of initial clarity could have substantial consequences for the intervention: for example, where schools had signed up for a digital intervention but did not have sufficient iPads for students (Sutherland et al., 2019). It could also lead to additional and unplanned work, for example, when teachers 'had not realised that they would have to write up pupil case studies after each round of Lesson Study' (Murphy et al., 2017, p.37). Clear guidance and communication contributed to maintaining teacher willingness to participate, and ensuring they knew what to do, and why.

A cross-cutting theme emerged across these forms of support, mentioned in ten evaluations: the value of developers' flexibility and prioritisation of convenience for teachers. Programme flexibility supported teacher buy-in, since teachers felt able to 'take charge of their own personal development rather than being required to deliver a "one-size-fits-all" scheme' (Speckesser et al., 2018, p.26). The Literacy Octopus intervention (which compared multiple interventions) also highlighted the importance of accessibility, achieved by providing ‘accessible information, different modes of engagement (e.g. printed materials, face-to-face training, webinars) and flexibility (e.g. time and location)’ (Lord et al., 2017, p.54). Teachers valued efforts to adapt training in response to feedback (Murphy et al., 2017), and discussions about how to adapt the programme to their circumstances (Biggart et al., 2015). Flexibility is double-edged: while some interventions reported flexibility as a strength (for example, O’Hare et al., 2019; Speckesser et al., 2018), others criticised excessive flexibility as a symptom of a lack of clarity about the intervention which left teachers unenthused and confused (McNally et al., 2014). Developers need to provide sufficient clear guidance about the purpose, goals and principles of the intervention, while maintaining the flexibility needed to ensure teachers can fit the intervention into their working patterns.

School context

Table 16: School context—themes of barriers and supports

Barrier/support theme	Number of studies mentioning this theme as affecting implementation
Alignment	9 (43%)
Leadership support	16 (76%)
Resources	4 (19%)
Time	16 (76%)
Turnover and absence	10 (48%)

Note. Categories are not exclusive, so may sum to more than 21. This is because an evaluation may refer to multiple themes.

Leadership support was mentioned in 16 studies, and emerged as a crucial barrier or support for success. Leadership support is referenced frequently in PD evaluations (for example, Glazerman et al., 2010; Jacob et al., 2017) and PD reviews (such as Cordingley et al., 2015). However, systematically reviewing these evaluations allows us to identify how leadership support affects programme participation. Most importantly, leadership support allowed teachers to prioritise the intervention, and provided the resources needed to do so. For example, in several studies, leaders provided non-contact time for teacher meetings (Culliney et al., 2019; Jay et al., 2017; Murphy et al., 2017). If leaders did not provide this support, it was often difficult to make the intervention work (Hanley et al., 2016; Jay et al., 2017; Lord et al., 2017; Sloan et al., 2018). Direct leadership participation in PD was also described as a contributing factor to success, since it could help them to understand the programme (Sutherland et al., 2019) and show their support for the intervention (Speckesser et al., 2018). The decision to adopt a programme was often driven by leaders (Humphrey et al., 2016; Murphy et al., 2017). Conversely, if leaders came to see an intervention as unworkable at any point, teachers and schools tended to disengage (Humphrey et al., 2016; Murphy et al., 2017). In particular, leaders could be pulled away from an intervention by other concerns, including academisation and teacher turnover (Humphrey et al., 2016). Where the leader responsible for the intervention moved roles or schools, this often meant the programme lost momentum, or was abandoned entirely (Culliney et al., 2019; Gorard et al., 2015). Securing *and maintaining* leadership support, with clear communication of what the intervention requires, appears to be essential in ensuring intervention success.

Alignment between an intervention and the school’s current practice, or aspirations, contributed to successful adoption of the change. Conversely, schools struggled where interventions clashed with their priorities and practices. For example, schools were ‘more receptive’ to the Good Behaviour Game where it matched their ‘existing ethos and practices’ (Humphrey et al., 2018, p.62). Some schools which withdrew from the intervention noted that it was in ‘conflict with the school’s behaviour management policy', or would not be extended to the next school year (Humphrey et al., 2018, p.43). There is an inevitable tension here: while greater alignment supports implementation, complete alignment might imply that a school had little to learn from the intervention. Where possible, developers may wish to assess school's current practices to identify schools which are sufficiently aligned to make the intervention viable, but which still have room to learn.

Teachers needed sufficient time to participate in interventions. Time was so important that it was described as 'the most crucial condition' for success (Jay et al., 2017, p.31). Sometimes, teachers needed additional planning and preparation time (Humphrey et al., 2018). Moreover, programme participation often required both cover and opportunities to arrange meetings between teachers and peer observations (Jay et al., 2017; Speckesser et al., 2018). The need to cover teachers' absence to attend training creates additional pressure (Tracey et al., 2019). In nurseries, teacher withdrawal was particular problematic because it affected staffing ratios (Robinson-Smith, 2018). Where interventions are not aligned to the national curriculum—such as Philosophy for Children—the requirement to find teaching time for the intervention could also be challenging (Gorard et al., 2015). A similar challenge arose when lesson resources had been designed and timed in ways which did not fit teaching timetables: one evaluation reported that 'most [teachers] were left feeling dissatisfied' with their teaching in consequence (Biggart et al., 2015, p.24). Inevitably, teachers were under pressure to maintain some existing routines even while implementing the intervention—for example, in assessing students regularly—which interventions might not have anticipated or left space for (for example, Styles et al., 2014).

It is worth considering these findings regarding time in the light of the current consensus. Reviewers tend to suggest PD should be sustained (Cordingley et al., 2015; Desimone, 2009; Dunst et al., 2015). There is a clear logic to this, with decades of criticism of 'drive by' workshops. However, three meta-analyses have failed to find a link between longer duration and greater impact (Basma & Savage, 2018; Kraft et al., 2018; Lynch et al., 2018). What our findings highlight is the intense time pressure placed on teachers. Every attendance at a training day, or observation of a peer, comes at a cost in cover and disruption. The accumulation of these costs can put teachers under sufficient pressure that they may not implement an intervention, or they may withdraw from it. This is exacerbated by the challenge of staff turnover and absence, which was raised as a barrier in almost half of the evaluations. In other words, a bigger, longer intervention is not always better. Developers must balance the desire to promote lasting and meaningful learning with the imperative to *minimise* the pressure they are placing on curriculum and teacher time.

Finally, resources were recorded as an important barrier in only four studies. For example, in both Digital Feedback and TEEP, schools referred to not having sufficient IT resources, such as an iPad for every student (Sutherland et al., 2019). However, the effect of a resource issue should not be underestimated: where it is central to the intervention design, it can be very problematic to the intervention.

5.7 Conclusions from the review of implementation

Three strong themes emerge from this analysis.

1. High fidelity is unlikely

Few interventions saw high fidelity. Many interventions recorded fairly low fidelity, based on limited data from schools which had completed the programme. Even programmes which were implemented relatively faithfully saw substantial adaptation and deviation from planned programmes. Developers should expect this kind of variation, and plan both the type of adaptations which can be permitted and encouraged, and ways to communicate this scope clearly to teachers. Adaptation is to be expected, with mixed effects: developers can play an important role in ensuring that adaptation is effective and beneficial. Equally, this review highlights what developers can do to encourage fidelity, notably offering clear guidance and making intervention resources easy to use.

2. The support system, intervention design, and school context influence implementation

Interventions are implemented effectively when the support system, intervention design, and school context work together to achieve success. Where an intervention design is problematic, enthusing teachers is challenging, even with a strong support system. The support system—particularly the guidance, training, and teacher educator support—plays an important role in translating good ideas into practical, usable guidance. The support of leaders, alignment with school priorities, and sufficient time combine to make an intervention work in school. While we cannot test the counterfactual with the data available, it seems very likely that it is the combination of these supports which make interventions succeed.

3. Interventions are more likely to be implemented when they fit schools' and teachers' needs

One theme emerges in every section of this analysis: the importance of fit between the intervention, the school's priorities, and the reality of the classroom. In part, this is an issue of alignment. Interventions work when they meet a perceived need in the school, and can build upon school policies and practices. More importantly, however, evaluation reports concur that interventions work when they make participating straightforward and convenient for teachers. This means clear communication and resources which are simple to use. Perhaps most importantly, it means recognising the severe constraints on teachers' time, and planning accordingly.

Limitations

We should acknowledge obvious limitations to this evaluation. We examined a small sample of PD interventions conducted in England in the last decade. These findings may not generalise to other education systems. The small sample and variation of measures used also confined us to a simple statistical approach of identifying references by study. A more comparable sample—particularly one which compared similar measures across studies—might offer fresh insights. The review of practice, published alongside this review, sheds further light on these issues, as will the EEF's forthcoming review of implementation. Moreover, as we mentioned at the beginning of this part of the document, studying implementation provides conclusions which are often both obvious and frequently overlooked (Pressman & Wildavsky, 1984).

Alongside these issues, we should also highlight some strengths of our sample. All but three of the evaluations included were prespecified (the exceptions were Gorard et al., 2015; McNally et al., 2014; Rienzo et al., 2015). All studies were conducted in English schools in the last decade (the oldest studies in the sample were published in 2014). This suggests they are likely to offer a representative impression of the challenges and supports to implementation of PD programmes in English schools. Finally, all included studies were efficacy or effectiveness trials, so they are likely to provide realistic data about the challenges facing programme developers.

Consequently, we believe the points raised here are worth consideration for PD developers and school leaders.

Research Question 4	
What supports successful implementation of PD programmes?	Few interventions were implemented with high fidelity. Even successfully-implemented

	programmes saw substantial adaptation and deviation from planned programmes. Interventions tend to be implemented successfully when the support system, intervention design, and school context are aligned.

Part 6: Summary and Discussion

6.1 Summary of findings on each research question

Research Question 1 What are the characteristics of the studies and interventions in the experimental impact evaluation literature on teacher PD?	Most studies are in the USA (70.2%) or UK (24.0%). They are distributed across different phases of education. Half (50.0%) focus on literacy, 28.9% on literacy, and 11.9% on science. There is a cross-curricular focus in 16.9% of studies. Only a minority use high-stakes standardised tests (27.9%). Very few (2.9%) are focused specifically on early-career teachers. A large proportion (39.4%) include an element of curricular reform, whereas only 6.7% include an element of education technology. Similarly, the largest proportion focus on portraying the curriculum (77.9%).
Research Question 2 Overall, are teacher PD interventions effective at increasing pupil achievement compared to business as usual? a. Does this vary based on study characteristics (features of the evaluation not specific to the intervention itself)? b. Does this vary by study-level pupil disadvantage or teacher experience levels?	Overall, across different phases of education, we find consistent, small, positive effects on student learning, ranging from .03 to 0.9 ($p < .01$). We find variation based on some study characteristics. Impact is lower for studies with 'acceptable' levels of attrition and for larger studies (those with more than 50 units randomised). The effect of pre-registered studies (.01) is much lower than the effect for those not pre-registered (.1). Average impact is slightly smaller among PD programmes focusing on early-career teachers compared with more experienced teachers, and higher among higher deprivation pupils compared with lower deprivation pupils.
Research Question 3 Which forms of PD are associated with the greatest impact?	We identified three forms of PD in Section 2.2: lesson study, instructional coaching, and (strong) teacher learning communities.

a. Which forms (clusters of mechanisms) do we observe in the literature?	
b. Are forms more likely to be effective when they incorporate mechanisms addressing all four purposes: instil insights (I), motivate goal-directed behaviour (G), develop techniques (T), and embed practice (P)?	Programmes that address all four purposes are on average more effective: the point estimate for programmes addressing all four purposes (.15) is three times that for programmes addressing three or fewer (.05). However, the 95% confidence interval for programmes addressing all four purposes is wide and the difference is not significant at conventional levels (*p* = 0.22).
c. Which forms (clusters of mechanisms) are associated with the largest effects on teacher practice and pupil achievement?	In general, we find that PD which incorporates more mechanisms is associated with greater impact on test scores. Programmes incorporating zero mechanisms have an expected effect size close to zero. PD mechanisms incorporating 14 mechanisms have an effect size close to .17. Across the three forms that we identify (instructional coaching, lesson study, and strong teacher learning communities), the average effect sizes are similar. The confidence intervals for each of the three forms also overlap to a large extent. This suggests that none of these three forms is clearly more effective than others.
Research Question 4	
What supports successful implementation of PD programmes?	Few interventions were implemented with high fidelity. Even successfully-implemented programmes saw substantial adaptation and deviation from planned programmes. Interventions tend to be implemented successfully when the support system, intervention design, and school context are aligned.

6.2 Discussion

We systematically reviewed the literature in order to investigate the characteristics of effective teacher PD. To do so, we developed and then tested a theoretical framework (the IGTP model) against 104 experimental PD studies. In this section, we summarise and discuss our findings, highlight some important limitations, and draw out implications for teacher educators and researchers.

We found an average impact of teacher PD on pupil standardised test scores of around .05 standard deviations – or approximately one month of additional pupil progress.[10] This is similar to the finding

[10] https://educationendowmentfoundation.org.uk/help/projects/the-eefs-months-progress-measure/

from a recent systematic review and meta-analysis of the effects of science and maths PD, which found an average impact of .06-.08 on standardised test scores (see Table 11 in Lynch et al., 2019). Although we find some indication of variation in effect sizes among studies with low/high sample sizes and low/high attrition rates, the overall impact remains positive and statistically significant. However, when we restrict our sample to studies that were pre-registered, the effect of PD is no longer distinguishable from zero. To our knowledge, this is the first review to present this finding in the context of teacher PD. We discuss possible explanations and implications of this in Section 6.4 below. In general, this finding suggests that we should exercise caution in interpreting the results from experimental evaluations of PD that have not been pre-registered.

The main goal of this review was to investigate what distinguishes more effective PD from less effective PD. To do this, we developed a conceptual framework incorporating 14 mechanisms: domain-general empirical regularities about how people learn new knowledge and skills and change their practice. Across 104 experimental evaluations, we found a positive association between the number of mechanisms that a PD programme incorporated and the impact of that programme on pupils' standardised test scores. More precisely, we found that each additional mechanism was associated with an increase in the effect size of .01. To put it another way, PD incorporating none of the 14 mechanisms tends to have an effect close to 0 (zero months' additional progress), PD incorporating five mechanisms tends to have an impact close to our sample average of .05 (one month of additional progress), and PD incorporating all 14 mechanisms tends to have an impact close to .17 (two months of additional progress). This suggests that the .05 average impact hides wide variations in effectiveness, depending on the design of PD.

Despite only including experimental studies in our review, the relationship we observe between mechanisms and impact *across studies* is not itself the result of random assignment. However, we argue that the way in which our theoretical framework has been developed provides warrant for interpreting this association as reflecting an underlying causal relationship (Sims & Fletcher-Wood, 2021). Crucially, we only included mechanisms in our framework if we had causal evidence that they are effective across multiple domains outside of teacher PD. For example, we know that setting specific goals helps motivate people to change their practice across health, education, sport, and laboratory contexts (Epton, 2017). Knowing that goal setting causally contributes to changing practice across a range of settings increases our confidence that the association we find between goal setting (amongst our other mechanisms) and the impact of teacher PD reflects an underlying causal relationship. This represents an advance over previous meta-regressions conducted in this literature.

The positive relationship that we find between the number of mechanisms and impact is evident among PD programmes evaluated using high-stakes test score outcomes (such as GCSEs), among trials with larger sample sizes (50 or more units randomised to treatment and control) and among school-age children (excluding early years settings in which effect sizes tend to be larger). Again, however, we found that the relationship almost disappears among pre-registered trials.

What explains the way that our results vary across trials that are pre-registered and not pre-registered? The finding that pre-registered studies have effect sizes close to zero could reflect (a) otherwise higher methodological standards (for example, use of high-stakes test scores) in pre-registered studies; (b) differences in the types of interventions tested using pre-registered trials; or (c) increased analytical flexibility in studies that are not pre-registered, allowing researchers to generate small positive impact estimates through trying alternative specifications.

Table 7e in Appendix 6 probes this empirically. It shows that pre-registered trials are indeed nine percentage points more likely to use high-stakes test score outcomes, 41 percentage points more likely to meet What Work Clearinghouse standards for acceptable attrition, and randomise almost twice the number of units to treatment and control. This suggests that some of the gap in effect sizes is likely to be explained by higher methodological standards among pre-registered studies (explanation a). At the same time, pre-registered trials tend to incorporate one less mechanism (four versus five), are one percentage point less likely to evaluate PD programmes with balanced designs (8% versus 9%), and are three percentage points less likely to evaluate PD programmes that also include curriculum reforms or new educational technology (45% versus 42%). This suggests that some of the gap in effect sizes could be explained by pre-registered studies tending to evaluate less intensive PD programmes (explanation b).

Our data cannot probe explanation c. However, there are documented examples of education trials that have found a precise zero when analysed in line with the protocol (Worth et al., 2017) and then returned positive and statistically significant results when subsequently reanalysed (Burgess et al., 2019). Similarly, meta-science research in psychology has found that pre-registered many-labs studies find smaller effects than published meta-analyses of the same hypotheses that include studies that were not pre-registered (Kvarven et al., 2020). On balance, we suspect that all three explanations (a, b and c) are likely to contribute to some extent to the differences in effect sizes between studies that are and are not pre-registered.

An important limitation of our analysis is that we can only code a mechanism as present in a study if the intervention is clearly and thoroughly described in the evaluation report. The extent to which we may have missed certain mechanisms in our coding because they were not described in the research reports is a ‘known unknown’ for our analysis. This is in addition to issues around inter-rater agreement on the mechanisms, where we found initial inter-rater mechanism-level agreement of 82% across 46 studies. Having said this, we take some reassurance about our ability to correctly identify mechanisms incorporated in PD programmes from our finding that PD programmes which we coded as having zero mechanisms tended to have an effect size of around zero. If we had found positive effect sizes among studies coded as having zero interventions, this would have been more of a concern. Furthermore, the R squared figures from our regressions of impact on number of mechanisms were 36% within PD programmes focusing on formative assessment or inquiry-based teaching. This provides some reassurance that we have captured much of what is important in terms of mechanisms, holding the content of the PD broadly constant.

Conditional on these limitations, one advantage of our theoretical framework is that it allows us to precisely characterise PD in terms of mechanisms. For example, lesson study can be thought of as comprising three mechanisms: practical social support, feedback, and action planning. Likewise, instructional coaching can be thought of as comprising goal setting, feedback, modelling/instruction, and rehearsal/context-specific repetition. Characterising PD in this way can help guard against the jingle fallacy, in which people use the same term (such as ‘instructional coaching’) to refer to quite different types of PD, and the jangle fallacy, in which people use two different terms (such as ‘lesson study’ and ’collaborative enquiry’) to refer to fundamentally similar types of PD (Thorndike, 1904; Kelley, 1927). Furthermore, focusing on mechanisms reveals that that average effect sizes from meta-analyses of specific forms (for example, instructional coaching) might misattribute some of the average impact to the instructional coaching, rather than the additional mechanisms that tend to be present alongside the instructional coaching. By tying terms like ‘instructional coaching’ directly to the mechanisms, our theoretical framework highlights what is essential to PD forms, and what is additional, thus helping to avoid such confusion.

Characterising PD programmes in terms of mechanisms allowed us to compare the impact of different ‘forms’ of PD. Interestingly, the point estimates for the impact of instructional coaching, lesson study, and (strong) teacher learning communities were all very similar. While our meta-analytic estimates remain somewhat imprecise, we found no evidence that any of these is superior to any of the others. Furthermore, when we looked at the relationship between impact and number of mechanisms within each form (for example, within the group of lesson study PD programmes), we still observed a positive relationship between mechanisms and impact. Indeed, across all our analyses, our results suggest that more of the variation in the effectiveness of PD is explained by PD mechanisms, rather than PD forms.

The IGTP framework provides one such way of thinking about PD in terms of mechanisms rather than forms. In the protocol, we hypothesised that balanced PD designs—those that include at least one mechanism addressing insight (I), at least one mechanism addressing goals (G), at least one mechanism addressing technique (T), and at least one mechanism addressing embedding practice (P)—would be more effective than PD designs that were not balanced. In our meta-analytic test of this hypothesis, we found that the point estimate for balanced designs (.15—approximately two months' additional progress) was three times higher than the point estimate for unbalanced designs (.05—approximately one month's additional progress). The notable increase in the point estimates between balanced and imbalanced designs, combined with the lack of increase in the point estimates across PD programmes that address one, two, or three of the IGTP purposes of PD, is consistent with our hypothesis. Having said that, the interval estimate for the balanced designs is wide and the contrast with imbalanced designs is not statistically significant at conventional levels (p >.05). In our QCA analysis, we found that PD programmes with balanced designs were consistently the most likely to have effect sizes above a range of thresholds for substantive significance, relative to all possible imbalanced designs. Our conclusion is that, on the balance of probabilities, balanced designs are more likely to be effective. However, there is considerable statistical uncertainty in this conclusion and it should be kept under close review as new studies emerge.

6.3 Implications for teacher educators

We believe that our review has a number of implications for teacher educators. First, teacher PD can improve pupil achievement. However, there is wide variation in effectiveness, depending on the design of the PD, ranging from zero to two months' additional progress. This suggests that it is worth paying careful attention when developing or commissioning teacher PD. The quality of PD makes all the difference in terms of impact on teaching and pupil learning.

Second, thinking about the mechanisms incorporated in a PD programme is a good way to go about evaluating PD. We found good empirical support for the importance of these mechanisms in explaining the effectiveness of PD. Appendix 5 contains detailed information about each of the mechanisms in our framework, including definitions, examples, non-examples, explanations of how they affect learning and/or practise, and the underpinning empirical evidence. This can be used to assess the 'active ingredients' in a PD programme and systematically think through its likely impact.

Third, thinking about PD programmes in terms of mechanisms is likely to be more fruitful than thinking about PD programmes in terms of PD forms. We found no clear evidence that lesson study, instructional coaching, or (strong) teacher learning communities were more or less effective than each other, on average. Instead, all three have similar, small, positive overall effects on average. Furthermore, we found that for any given PD form, the programmes that incorporated more additional mechanisms tended to be more effective.

Fourth, the IGTP framework provides, at the very least, a useful heuristic tool for thinking about whether a PD programme addresses the purposes of teacher PD. Appendix 8 provides three examples of balanced PD designs—those that incorporate at least one mechanism addressing each of insight (I), goals (G), technique (T), and embedding new practice (P). We found that such balanced designs tend to have three times higher effects on average, albeit with high levels of statistical uncertainty. Teacher educators should look out for future evaluations of balanced PD designs to see if further research confirms or disconfirms the superiority of balanced designs.

Fifth, our results reinforce the difficulties of comparing effect sizes across experiments, due to potentially misleading variation in effect sizes across outcome measures, age groups, PD content, and whether or not studies were pre-registered. This suggests that teacher educators should exercise caution when comparing the impact estimates across evaluations of different PD programmes. It would be advisable to focus instead on recent meta-analyses where efforts have been made to account for these confounding influences on impact estimates (for example, Kraft et al., 2018; Lynch et al., 2019).

6.4 Implications for research

Our review has a number of implications for research. First, we hope that our framework is useful in helping to disambiguate different forms of teacher PD. For example, we found 15 randomised controlled trials of instructional coaching interventions with test score outcomes using our mechanism-based definition, whereas Kraft et al. (2018) found 20 studies in an earlier review using a more process-based definition. The two reviews use slightly different inclusion criteria, so this is perhaps not surprising. However, the discrepancy serves to highlight the important ambiguities in how we classify PD interventions. We submit that researchers should consider using mechanism-based definitions where possible, since this places the emphasis on the likely 'active ingredients' of the PD, thus minimising ambiguity where ambiguity matters most. Specifying PD interventions in terms of their constituent mechanisms in evaluation reports would also help support teacher educators with implementation of evidence-based PD programmes.

Second, our analysis reveals striking disparities in the average impact of PD interventions depending on whether the analysis was pre-registered. Indeed, the average effect is no longer distinguishable from zero among pre-registered trials. Our analysis suggests that this likely reflects a mix of (a) otherwise higher methodological standards (such as use of high-stakes test scores) in pre-registered studies, (b) differences in the types of interventions tested using pre-registered trials, and (c) increased analytical flexibility in studies that are not pre-registered, allowing researchers to 'dredge up' small positive impact estimates regardless of the true impact. From the perspective of research, however, only the pre-registration of PD evaluations can rule out explanation (a) when interpreting future experimental findings. Trials that are not pre-registered should be treated with caution, particularly when effect sizes are small. Researchers conducting pre-registered trials should also increase their expectations about the power required to detect likely effects, relative to what is suggested by findings from previous trials that were not pre-registered.

Third, our conceptual framework suggests a number of interesting avenues for future research. The IGTP framework is adaptable in that new mechanisms can be added (or indeed removed) based on developments in basic social science research. Although we searched hard for well-evidenced mechanisms, the framework can likely be strengthened as basic social science research progresses. In particular, we identified only two mechanisms addressing insight (I). Separately, while we found clear differences in the point estimates for balanced and imbalanced designs, the RVE random effects meta-analysis model yielded imprecise estimates for the balanced designs. We believe that the empirical findings presented in this report are strong enough to justify funders commissioning additional evaluations of PD interventions with balanced designs, in order to increase the precision of future meta-analytic tests. Our empirical findings are also consistent with the view that pre-registered field trials using high stakes test scores remain the most rigorous way to assess effect sizes. In the first instance, however, tests of the IGTP framework might be better conducted in higher-powered laboratory experiments using simulated teaching tasks (for example, Cohen et al., 2020).

Fourth, and finally, our review provides further reason to doubt the value of review-of-reviews exercises, in which researchers summarise the results from previous reviews of the PD literature (for example, Cordingley et al., 2015; Dunst et al., 2015). Half of the experiments that we identified in this review have been published since 2015, which reflects a wider increase in the number of rigorous experimental studies being conducted in education research (Connolly, Keenan & Urbanska, 2018; Dawson, Yeomans & Brown, 2018; Hedges & Schauer, 2018). Since the publication process can take a year or more, reviews-of-reviews (which imply three stages of publication) are likely to be missing a sizable proportion of the existing literature. More generally, as the pace of publication of good research accelerates, up-to-date meta-analyses become more important for getting an accurate picture of the existing evidence base.

References (cited in text)

Abe, Y., Thomas, V., Sinicrope, C. & Gee, K. A. (2012). *Effects of the Pacific CHILD Professional Development Program. Final Report* (NCEE 2013–4002). National Center for Education Evaluation and Regional Assistance.

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659-701.

Al-Saud, L. M., Mushtaq, F., Allsop, M. J., Culmer, P. C., Mirghani, I., Yates, E., ... & Manogue, M. (2017). Feedback and motor skill acquisition using a haptic dental simulator. *European Journal of Dental Education*, *21*(4), 240–247.

Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y. & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner—Secondary intervention. *Journal of Research on Educational Effectiveness, 8*(4), 475–489.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y. & Lun, J. (2011, August 19). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045) 1034–1037.

August, D., Branum-Martin, L., Cárdenas-Hagan, E., Francis, D. J., Powell, J., Moore, S. & Haynes, E. F. (2014). Helping ELLs meet the Common Core State Standards for literacy in science: The impact of an instructional intervention focused on academic language. *Journal of Research on Educational Effectiveness*, 7(1), 54–82.

Baddeley, A. & Hitch, G. (1974). Working memory. In Bower, G. (Ed), *The Psychology of Learning and Motivation*, ed. Bower, G. (pp. 47–89). Academic Press.

Bagnall, A. M., South, J., Hulme, C., Woodall, J., Vinall-Collier, K., Raine, G., ... & Wright, N. M. (2015). A systematic review of the effectiveness and cost-effectiveness of peer education and peer support in prisons. *BMC Public Health*, 15(1), 1–30.

Bambrick-Santoyo, P. (2018). *Leverage leadership 2.0: A practical guide to building exceptional schools*. John Wiley & Sons.

Black, N. & de New, S. C. (2020). Short, heavy and underrated? Teacher assessment biases by children's body size. *Oxford Bulletin of Economics and Statistics*, *82*(5), 961–987.

Boylan, M. & Demack, S. (2018). Innovation, evaluation design and typologies of professional learning. *Educational Research*, *60*(3), 336–356.

Brown, A. & Campione, J. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In: Schauble, L. & Glaser, R. (Eds.) *Innovations in learning: New environments for education* (pp. 289–325). Lawrence Erlbaum Associates, Inc.

Bruhn, A. L., Hirsch, S. E. & Lloyd, J. W. (2015). Treatment integrity in school-wide programs: A review of the literature (1993–2012). *The journal of primary prevention*, 36(5), 335–349.

Burgess, S., Rawal, S. & Taylor, E. S. (2019). Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools. https://scholar.harvard.edu/files/erictaylor/files/peer-evaluation-brt.pdf

Burke, L. E., Wang, J. & Sevick, M. A. (2011). Self-monitoring in weight loss: a systematic review of the literature. *Journal of the American Dietetic Association*, 111(1), 92–102.

Calzolari, G. & Nardotto, M. (2017). Effective reminders. *Management Science*, *63*(9), 2915–2932.

Campuzano, L., Dynarski, M., Agodini, R. & Rall, K. (2009). *Effectiveness of Reading and Mathematics Software Products—Findings from two student cohorts* (NCEE 2009–4041). National Center for Education Evaluation and Regional Assistance.

Carey, R. N., Connell, L. E., Johnston, M., Rothman, A. J., De Bruin, M., Kelly, M. P. & Michie, S. (2019). Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. *Annals of Behavioral Medicine*, *53*(8), 693–707.

Compernolle, S., DeSmet, A., Poppe, L., Crombez, G., De Bourdeaudhuij, I., Cardon, G., ... & Van Dyck, D. (2019). Effectiveness of interventions using self-monitoring to reduce sedentary behavior in adults: a systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity*, *16*(1), 1–16.

Cheung, A. C. & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283–292.

Chow, A. W. (2016). Teacher learning communities: the landscape of subject leadership. *International Journal of Educational Management, 30*(2), 287–307.

Clements, D. H., Sarama, J. & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-based Early Maths Assessment. *Educational Psychology, 28*, 457–482.

Cohen, J., Wong, V., Krishnamachari, A. & Berlin, R. (2020). Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, *42*(2), 208–231.

Connolly, P., Keenan, C. & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, *60*(3), 276–291.

Cordingley, P., Higgins, S., Greany, T., Buckler, N., Coles-Jordan, D., Crisp, B., Saunders, L. & Coe, R. (2015). *Developing great teaching: Lessons from the international reviews into effective professional development*. Teacher Development Trust.

Cordovani, L. & Cordovani, D. (2016). A literature review on observational learning for medical motor skills and anesthesia teaching. *Advances in Health Sciences Education*, *21*(5), 1113–1121.

Cromley, J. G., Weisberg, S. M., Dai, T., Newcombe, N. S., Schunn, C. D., Massey, C. & Merlino, F. J. (2016). Improving middle school science learning using diagrammatic reasoning. *Science Education*, *100*(6), 1184–1213.

Cunningham, M. R. & Barbee, A. P. (2000). *Social support.* In C. Hendrick & S. Hendrick (Eds.), *Close relationships: A sourcebook* (p. 273–285). Sage Publications.

Dale, J. R., Williams, S. M. & Bowyer, V. (2012). What is the effect of peer support on diabetes outcomes in adults? A systematic review. *Diabetic Medicine*, 29(11), 1361–1377.

Dane, A. V. & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical psychology review*, 18(1), 23–45.

Dariotis, J. K., Bumbarger, B. K., Duncan, L. G. & Greenberg, M. T. (2008). How do implementation efforts relate to program adherence? Examining the role of organizational, implementer, and program factors. *Journal of Community Psychology*, 36(6), 744–760.

Dawson, A., Yeomans, E. & Brown, E. R. (2018). Methodological challenges in education RCTs: reflections from England's Education Endowment Foundation. *Educational Research*, *60*(3), 292–310.

Delin, C. R. & Baumeister, R. F. (1994). Praise: More than just social reinforcement. *Journal for the theory of social behaviour*, *24*(3), 219–241.

Dennis, C. L. (2003). Peer support within a health care context: a concept analysis. *International journal of nursing studies*, 40(3), 321–332.

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

DfE [Department for Education] (2019). Early years foundation stage assessment and reporting arrangements (ARA). Department for Education.

Dunst, C. J., Bruder, M. B. & Hamby, D. W. (2015). Metasynthesis of in-service professional development research: Features associated with positive educator and student outcomes. *Educational Research and Reviews, 10*(12), 1731–1744.

Durlak, J. A. (2010). The importance of doing well in whatever you do: A commentary on the special section, "Implementation research in early childhood education". *Early Childhood Research Quarterly*, 25(3), 348–357.

Durlak, J. A. & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American journal of community psychology*, 41(3–4), 327–350.

Duval, S. & Tweedie, R. (2000). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463.

Epton, T., Currie, S. and Armitage, C. J., (2017). Unique effects of setting goals on behavior change: Systematic review and meta-analysis. Journal of Consulting and Clinical Psychology, *85*(12), 1182–1198.

Cabell, S. Q., Justice, L. M., Piasta, S. B., Curenton, S. M., Wiggins, A., Turnbull, K. P. & Petscher, Y. (2011). The impact of teacher responsivity education on preschoolers' language and literacy skills. American Journal of Speech-Language Pathology, 20, 315–330.

Ericsson, K. A., & Harwell, K. W. (2019). Deliberate practice and proposed limits on the effects of practice on the acquisition of expert performance: Why the original definition matters and recommendations for future research. *Frontiers in psychology*, 10, 2396.

Ericsson, K. A., Krampe, R. and Tesch-Rohmer, C. (1993). The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review*, 100(3), 363–406.

Fagan, A. A. and Mihalic, S. (2003). Strategies for enhancing the adoption of school-based prevention programs: Lessons learned from the Blueprints for Violence Prevention replications of the Life Skills Training program. *Journal of Community Psychology*, 31(3), 235–253.

Fernandez, C. (2002). Learning from Japanese approaches to professional development: The case of lesson study. *Journal of Teacher Education*, *53*(5), 393–405.

Fernandez, C. & Yoshida, M. (2012). *Lesson study: A Japanese approach to improving mathematics teaching and learning*. Routledge.

Filges, T., Torgerson, C., Gascoine, L. & Dietrichson, J. (2019). Effectiveness of continuing professional development training of welfare professionals on outcomes for children and young people: a systematic review. *Campbell Systematic Reviews*, *15*(4), e1060.

Finkelstein, N., Hanson, T., Huang, C. W., Hirschman, B. & Huang, M. (2010). *Effects of Problem Based Economics on High School Economics Instruction Final Report* (NCEE 2010–4002). National Center for Education Evaluation and Regional Assistance.

Fletcher-Wood, H. & Zuccollo, J. (2020). *The effects of high-quality professional development on teachers and students: A rapid review and meta-analysis*. Wellcome Trust.

Fong, C. J., Patall, E. A., Vasquez, A. C. & Stautberg, S. (2019). A meta-analysis of negative feedback on intrinsic motivation. *Educational Psychology Review*. 31, 121–162.

Fraser, K. L., Ayres, P., & Sweller, J. (2015). Cognitive load theory for the design of medical simulations. *Simulation in Healthcare*, *10*(5), 295-307.

Furnari, S., Crilly, D., Misangyi, V. F., Greckhamer, T., Fiss, P. C. Aguilera, R. (2021). Capturing Causal Complexity: Heuristics for configurational theorizing. *Academy of Management Review*.

Gallucci, C., Van Lare, M. D., Yoon, I. H., & Boatright, B. (2010). Instructional coaching: Building theory about the role and organizational support for professional learning. *American Educational Research Journal, 47*(4), 919–963.

Gardner, B. & Rebar, A. L. (2019). Habit formation and behavior change. In *Oxford Research Encyclopedia of Psychology*.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M. & Jacobus, M. (2010). Impacts of comprehensive teacher induction: Final results from a randomized controlled study. Washington, DC: National Center for Education Evaluation. Retrieved from https://ies.ed.gov/ncee/pubs/20104027/

Gleason, P., Crissey, S., Chojnacki, G., Zukiewicz, M., Silva, T., Costelloe, S. & O'Reilly, F. (2019). Evaluation of Support for Using Student Data to Inform Teachers' Instruction. NCEE 2019–4008. *National Center for Education Evaluation and Regional Assistance*.

Goodrich, J. (2022). *Responsive Coaching.* John Catt Educational.

Gollwitzer, P. M. & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology*, 38, 69–119.

Graneheim, U. H. and Lundman, B., (2004). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse education today, 24(2), pp. 105–112.

Gregory, A., Ruzek, E., Hafen, C. A., Mikami, A. Y., Allen, J. P. & Pianta, R. C. (2017). My Teaching Partner—Secondary: A video-based coaching model. *Theory into Practice, 56*(1), 38–45.

Gollwitzer, P. M. & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. Advances in experimental social psychology, 38, 69–119.

Grierson, L. E., Barry, M., Kapralos, B., Carnahan, H. & Dubrowski, A. (2012). The role of collaborative interactivity in the observational practice of clinical skills. *Medical Education*, *46*(4), 409–416.

Guy, R., Hocking, J., Wand, H., Stott, S., Ali, H. & Kaldor, J. (2012). How effective are short message service reminders at increasing clinic attendance? A meta-analysis and systematic review. *Health services research*, *47*(2), 614–632.

Hambrick, D. Z., Macnamara, B. N. & Oswald, F. L. (2020). Is the Deliberate Practice View Defensible? A Review of Evidence and Discussion of Issues. *Frontiers in Psychology*, 11.

Hanley, P., Bohnke, J., Slavin, B., Elliott, L. & Croudace, T. (2016). *Let's Think Secondary Science: Evaluation report and executive summary*. Education Endowment Foundation.

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *Doing meta-analysis with R: a hands-on guide*. CRC Press.

Harris, D. J., Vine, S. J., Wilson, M. R., McGrath, J. S., LeBel, M. E. & Buckingham, G. (2018). Action observation for sensorimotor learning in surgery. *Journal of British Surgery*, *105*(13), 1713–1720.

Hatala, R., Cook, D. A., Zendejas, B., Hamstra, S. J. & Brydges, R. (2014). Feedback for simulation-based procedural skills training: a meta-analysis and critical narrative synthesis. *Advances in Health Sciences Education*, *19*(2), 251–272.

Hedges, L. V. & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research*, *60*(3), 265–275.

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.

Hill, H. C., Beisiegel, M. & Jacob, R. (2013). Professional development research: Consensus, crossroads, and challenges. *Educational Researcher, 42*(9), 476–487.

Hobbiss, M., Sims, S. & Allen, R. (2021). Habit formation limits growth in teacher effectiveness: A review of converging evidence from neuroscience and social science. *Review of Education*, *9*(1), 3–23.

Hornikx, J. (2005). A review of experimental research on the relative persuasiveness of anecdotal, statistical, causal, and expert evidence. *Studies in Communication Sciences*, 5(1), 205–216.

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R. & Kerr, K. (2016). Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook. London: The Education Endowment Foundation.

Illari, P. M. & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119–135.

Ismagilova, E., Slade, E., Rana, N. P. et al (2020). The effect of characteristics of source credibility on consumer behaviour: a meta-analysis. *Journal of Retailing and Consumer Services*. 53: 101736.

Jacob, R., Hill, H. & Corey, D. (2017). The Impact of a Professional Development Program on Teachers' Mathematical Knowledge for Teaching, Instruction, and Student Achievement. *Journal of Research on Educational Effectiveness*, 10(2), 379–407.

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L. & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for research in mathematics education*, 258–288.

Jolly, K., Ingram, L., Khan, K. S., Deeks, J. J., Freemantle, N. & MacArthur, C. (2012). Systematic review of peer support for breastfeeding continuation: metaregression analysis of the effect of setting, intensity, and timing. *BMJ*, 344.

Joyce, B. R. & Showers, B. (1981). Transfer of training: The contribution of "coaching". *Journal of Education, 163*(2), 163–172.

Karanika-Murray, M. and Biron, C. (2015). Introduction—Why Do Some Interventions Derail? Deconstructing the Elements of Organizational Interventions for Stress and Well-Being. In *Derailed Organizational Interventions for Stress and Well-Being* (1–15). Springer.

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book Company.

Kennedy, M. M. (2016a). How does professional development improve teaching? *Review of Educational Research, 86*(4), 945–980.

Kennedy, M. M. (2016b). Parsing the practice of teaching. *Journal of Teacher Education*, *67*(1), 6–17.

Kirschner, P., Sweller, J. & Clark, R. E. (2006). Why unguided learning does not work: An analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. *Educational Psychologist*, *41*(2), 75–86.

Kitmitto, S., González, R., Mezzanote, J. & Chen, Y. (2018). *Thinking, Doing, Talking Science Evaluation report and executive summary.* Education Endowment Foundation.

Knight, J. (2007). *Instructional coaching: A partnership approach to improving instruction*. Corwin Press.

Knight, B., Turner, D., & Dekkers, J. (2013). The future of the practicum: Addressing the knowing-doing gap. *Teacher education in Australia: Investigations into programming, practicum and partnership*, 63-76.

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

Kluger, A. and DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin, 119(2), 254-284.

Kraft, M. A., & Hill, H. C. (2020). Developing ambitious mathematics instruction through web-based coaching: A randomized field trial. *American Educational Research Journal*, 57(6), 2378–2414.

Kvarven, A., Strømland, E. & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434.

Lee, J. F. (2008). A Hong Kong case of lesson study—Benefits and concerns. *Teaching and Teacher Education, 24*(5), 1115–1124.

Lee, S. W. (2018). Pulling back the curtain: Revealing the cumulative importance of high-performing, highly qualified teachers on students' educational outcome. *Educational Evaluation and Policy Analysis*, *40*(3), 359–381.

Lendrum, A. & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, 38(5), 635–652.

Lewis, C., Perry, R. & Murata, A. (2006). How should research contribute to instructional improvement? The case of lesson study. *Educational Researcher*, *35*(3), 3–14.

Llosa, L., Lee, O., Jiang, F., Haas, A., O'Connor, C., Van Booven, C. D. & Kieffer, M. J. (2016). Impact of a large-scale science intervention focused on English language learners. *American Educational Research Journal*, *53*(2), 395–424.

Lynch, K., Hill, H. C., Gonzalez, K. E. and 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.

Macnamara, B. N., Moreau, D. & Hambrick, D. Z. (2016). The relationship between deliberate practice and performance in sports: A meta-analysis. *Perspectives on Psychological Science*, 11(3), 333–350.

McGaghie, W. C., Issenberg, S. B., Cohen, M. E. R., Barsuk, J. H. & Wayne, D. B. (2011). Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Academic medicine: journal of the Association of American Medical Colleges*, 86(6), 706.

McLaughlin, M. W. & Talbert, J. E. (2006). *Building school-based teacher learning communities: Professional strategies to improve student achievement*. Teachers College Press.

Matsumura, L., Garnier, H. & Spybrook, J. (2012). The effect of content-focused coaching on the quality of classroom text discussions. *Journal of Teacher Education, 63*(3), 214-228.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., ... & Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine*, *46*(1), 81–95.

Miller, G., Jaciw, A., Ma, B. & Wei, X. (2007). *Comparative effectiveness of Scott Foresman Science: A report of randomized experiments in five school districts*. Empirical Education.

Mokyr, J. (2002). *The Gifts of Athena: Historical origins of the knowledge economy*. Princeton University Press.

Moore, J. E., Bumbarger, B. K. & Cooper, B. R. (2013). Examining adaptations of evidence-based programs in natural contexts. *The Journal of Primary Prevention*, 34(3), 147–161.

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V. & Jaeger, A. (2019, February). Retrieval practice in classroom settings: a review of applied research. Frontiers in Education, 4, 5.

Murata, A. (2011). Introduction: Conceptual overview of lesson study. In Hart, L. C., Alston, A. S. & Murata, A. (Eds.) *Lesson study research and practice in mathematics education*, 1–12. Springer.

Neal, D. T. & Wood, W. (2009). Automaticity in situ and in the lab: The nature of habit in daily life. *Oxford handbook of human action*, 442–457.

Nickow, A., Oreopoulos, P. & Quan, V. (2020). *The impressive effects of tutoring on prek-12 learning: a systematic review and meta-analysis of the experimental evidence* (NBER w27476). National Bureau of Economic Research.

O’Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. Review of educational research, 78(1), 33–84.

O'Keefe, D. J. (1998). Justification explicitness and persuasive effect: A meta-analytic review of the effects of varying support articulation in persuasive messages. *Argumentation and advocacy*, 35(2), 61–75.

Olson, C. B., Matuchniak, T., Chung, H. Q., Stumpf, R. & Farkas, G. (2017). Reducing achievement gaps in academic writing for Latinos and English learners in Grades 7–12. *Journal of Educational Psychology*, 109(1), 1.

Opfer, V. D. & Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research*, *81*(3), 376–407.

Pastötter, B. & Bäuml, K. H. T. (2014). Retrieval practice enhances new learning: the forward effect of testing. *Frontiers in Psychology*, *https://doi.org/10.3389/fpsyg.2014.00286*.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology, 61*, 991–996.

Powell, D. R., Diamond, K. E., Burchinal, M. R. & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on head start teachers and children. *Journal of educational psychology*, *102*(2), 299–312.

Pressman, J. L. & Wildavsky, A. (1984). Implementation: How great expectations in Washington are dashed in Oakland: Third Edition. Berkeley: University of California Press.

Priestley, M., Miller, K., Barrett, L. & Wallace, C. (2011). Teacher learning communities and educational change in Scotland: the Highland experience. *British Educational Research Journal, 37*(2), 265–284.

Ramchand, R., Ahluwalia, S. C., Xenakis, L., Apaydin, E., Raaen, L. & Grimm, G. (2017). A systematic review of peer-supported interventions for health promotion and disease prevention. *Preventive Medicine, 101*, 156–170.

Reinard, J. C. (1988). The empirical study of the persuasive effects of evidence the status after fifty years of research. *Human Communication Research*, 15(1), 3–59.

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science*, 38(1), 1–37.

Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review*, *27*(4), 635-643.

Sampson, V., Enderle, P. J. & Walker, J. P. (2012). The development and validation of the assessment of scientific argumentation in the classroom (ASAC) observation protocol: A tool for evaluating how students participate in scientific argumentation. In Khine, M. (Ed.) *Perspectives on scientific argumentation* (pp. 235–264). Springer.

Sellen, P. (2016). *Teacher workload and professional development in England's secondary schools: Insights from TALIS*. Education Policy Institute.

Shojania, K. G., Jennings, A., Mayhew, A., Ramsay, C., Eccles, M. & Grimshaw, J. (2010). Effect of point-of-care computer reminders on physician behaviour: a systematic review. *Cmaj*, *182*(5), E216–E225.

Simonsohn, U., Nelson, L. D. & Simmons, J. P. (2014a). P-curve: a key to the file-drawer. *Journal of experimental psychology: General*, *143*(2), 534.

Simonsohn, U., Nelson, L. D. & Simmons, J. P. (2014b). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681.

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Stansfield, C., Van Herwegen, J., Cottingham, S. & Higton, J. (2021). *What are the characteristics of teacher professional development that increase pupil achievement? Protocol for a systematic review.* Education Endowment Foundation.

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

Stoll, L., Bolam, R., McMahon, A., Wallace, M. & Thomas, S. (2006). Professional learning communities: A review of the literature. *Journal of Educational Change, 7*, 221–258.

Sweller, J., van Merriënboer, J. J. & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292.

Sutton, A. J. (2009). Publication bias. *The Handbook of Research Synthesis and Meta-Analysis*, *2*, 435–452.

Tanner-Smith, E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*(1), 13–30.

Todd, J. & Mullan, B. (2014). The role of self-monitoring and response inhibition in improving sleep behaviours. *International Journal of Behavioral Medicine*, *21*(3), 470–477.

Thomas, J., O'Mara-Eves, A. & Brunton, G. (2014). Using Qualitative Comparative Analysis (QCA) in systematic reviews of complex interventions: a worked example. *Systematic Reviews, 3*(67).

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. Columbia University Press.

Timperley, H., Wilson, A., Barrar, H. & Fung, I. (2007). *Teacher professional learning and development: Best evidence synthesis iteration (BES)*. New Zealand Ministry of Education.

Van Achterberg, T., Huisman-de Waal, G. G., Ketelaar, N. A., Oostendorp, R. A., Jacobs, J. E. & Wollersheim, H. C. (2011). How to promote healthy behaviours in patients? An overview of evidence for behaviour change techniques. *Health Promotion International*, 26(2), 148–162.

van Aert, R. C., Wicherts, J. M. & van Assen, M. A. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, *11*(5), 713–729.

Van Es, E. A. (2012). Examining the development of a teacher learning community: The case of a video club. *Teaching and Teacher Education, 28*(2), 182–192.

Vescio, V., Ross, D. & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education*, *24*(1), 80–91.

Vevea, J. L. & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435.

Walter, C. & Briggs, J. (2012). *What professional development makes the most difference to teachers*. Oxford University Press.

Wayne, A., Yoon, K., Zhu, P., Cronen, S. & Garet, M. (2008). Experimenting With Teacher Professional Development: Motives and Methods. *Educational Researcher*, 37(8), 469–479.

Webel, A. R., Okonsky, J., Trompeta, J. & Holzemer, W. L. (2010). A systematic review of the effectiveness of peer-based interventions on health-related behaviors in adults. *American journal of public health*, 100(2), 247–253.

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

Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree Press.

Wiliam, D. & Leahy, S. (2012). *Sustaining formative assessment with teacher learning communities*. Dylan Wiliam Centre.

Willingham, D. T. (2017). A mental model of the learner: Teaching the basic science of educational psychology to future teachers. *Mind, Brain, and Education*, *11*(4), 166–175.

Wood, W. & Rünger, D. (2016). Psychology of habit. *Annual Review of Psychology*, 67.

Worth, J., Sizmur, J., Walker, M., Bradshaw, S. & Styles, B. (2017). Teacher Observation: Evaluation Report and Executive Summary. *Education Endowment Foundation*.

Yang, C., Potts, R. & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: a review of the forward testing effect. *NPJ Science of Learning*, *3*(1), 1–9.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B. & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. Institute of Education Sciences.

References (included in meta-analysis)

Abe, Y., Thomas, V., Sinicrope, C. & Gee, K. A. (2012). *Effects of the Pacific CHILD Professional Development Program. Final Report* (NCEE 2013–4002). National Center for Education Evaluation and Regional Assistance.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y. & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*(6045), 1034–1037.

Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y. & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner-Secondary intervention. *Journal of Research on Educational Effectiveness*, *8*(4), 475–489.

Ansari, A. & Pianta, R. C. (2018). Effects of an early childhood educator coaching intervention on preschoolers: The role of classroom age composition. *Early Childhood Research Quarterly, 44*, 101–113.

Arens, S. A., Stoker, G., Barker, J., Shebby, S., Wang, X., Cicchinelli, L. F. & Williams, J. M. (2012). *Effects of Curriculum and Teacher Professional Development on the Language Proficiency of Elementary English Language Learner Students in the Central Region Final Report* (NCEE 2012–4013). National Center for Education Evaluation and Regional Assistance.

Argentin, G., Pennisi, A., Vidoni, D., Abbiati, G. & Caputo, A. (2014). Trying to raise (low) math achievement and to promote (rigorous) policy evaluation in Italy: Evidence from a large-scale randomized trial. *Evaluation review*, *38*(2), 99–132.

Arteaga, I., Thornburg, K., Darolia, R. & Hawks, J. (2019). Improving Teacher Practices With Children Under Five: Experimental Evidence From the Mississippi Buildings Blocks. *Evaluation review*, *43*(1-2), 41–76.

August, D., Branum-Martin, L., Cárdenas-Hagan, E., Francis, D. J., Powell, J., Moore, S. & Haynes, E. F. (2014). Helping ELLs meet the Common Core State Standards for literacy in science: The impact of an instructional intervention focused on academic language. *Journal of Research on Educational Effectiveness*, *7*(1), 54–82.

Babinski, L. M., Amendum, S. J., Knotek, S. E., Sánchez, M. & Malone, P. (2018). Improving young English learners’ language and literacy skills through teacher professional development: A randomized controlled trial. *American Educational Research Journal*, *55*(1), 117–143.

Biggart, A. (2015). *Quest: Evaluation Report and Executive Summary*. Education Endowment Foundation.

Boardman, A. G., Klingner, J. K., Buckley, P., Annamma, S. & Lasser, C. J. (2015). The efficacy of Collaborative Strategic Reading in middle school science and social studies classes. *Reading and Writing*, *28*(9), 1257–1283.

Bos, J. M., Sanchez, R. C., Tseng, F., Rayyes, N., Ortiz, L. & Sinicrope, C. (2012). *Evaluation of Quality Teaching for English Learners (QTEL) Professional Development. Final Report* (NCEE 2012–4005). National Center for Education Evaluation and Regional Assistance.

Brendefur, J., Strother, S., Thiede, K., Lane, C. & Surges-Prokop, M. J. (2013). A professional development program to improve math skills among preschool children in Head Start. *Early Childhood Education Journal*, *41*(3), 187–195.

Buysse, V., Castro, D. C. & Peisner-Feinberg, E. (2010). Effects of a professional development program on classroom practices and outcomes for Latino dual language learners. *Early childhood research Quarterly*, *25*(2), 194–206.

Campbell, P. F. & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, *111*(3), 430-454.

Cavalluzzo, L., Geraghty, T. M., Steele, J. L., Holian, L., Jenkins, F., Alexander, J. M., & Yamasaki, K. Y. (2014). *"Using Data" to Inform Decisions: How Teachers Use Data to Inform Practice and Improve Student Performance in Mathematics. Results from a Randomized Experiment of Program Efficacy*. CNA Corporation.

Chuang, C. C., Reinke, W. M. & Herman, K. C. (2020). Effects of a Universal Classroom Management Teacher Training Program on Elementary Children with Aggressive Behaviors. *School Psychology*, *35*(2), 128-136.

Clements, D. H., Sarama, J., Wolfe, C. B. & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*(4), 812–850.

Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C. & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science*, *315*(5811), 464.

Cordray, D., Pion, G., Brandt, C., Molefe, A. & Toby, M. (2012). *The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement. Final Report* (NCEE 2013–4000). National Center for Education Evaluation and Regional Assistance.

Correnti, R., Matsumura, L. C., Walsh, M., Zook-Howell, D., Bickel, D. D. & Yu, B. (2020). Effects of Online Content-Focused Coaching on Discussion Quality and Reading Achievement: Building Theory for How Coaching Develops Teachers' Adaptive Expertise. *Reading Research Quarterly*. https://doi.org/10.1002/rrq.317

Culliney, M., Moore, N., Coldwell, M. & Demack, S. (2019). *Integrating English: Evaluation Report*. Education Endowment Foundation.

DeCesare, D., McClelland, A. & Randel, B. (2017). *Impacts of the Retired Mentors for New Teachers Program* (REL 2017-225). Regional Educational Laboratory Central.

Dix, K., Hollingsworth, H., & Carslake, T. (2018). *Thinking Maths: Learning impact fund evaluation report: evaluation report and executive summary*. Social Ventures Australia.

Dolfin, S., Richman, S., Choi, J., Streke, A., DeSaw, C., Demers, A., & Poznyak, D., & Mathematica (2019). *Evaluation of the Teacher Potential Project*. Mathematica Policy Research.

Engelstad, A. M., Holingue, C., & Landa, R. J. (2020). Early Achievements for Education Settings: An Embedded Teacher-Implemented Social Communication Intervention for Preschoolers With Autism Spectrum Disorder. *Perspectives of the ASHA Special Interest Groups*, *5*(3), 582-601.

Finkelstein, N., Hanson, T., Huang, C. W., Hirschman, B., & Huang, M. (2010). *Effects of Problem Based Economics on High School Economics Instruction Final Report* (NCEE 2010–4002). National Center for Education Evaluation and Regional Assistance.

Foliano, F., Rolfe, H., Buzzeo, J., Runge, J. & Wilkinson, D. (2019). *Changing mindsets: Effectiveness trial.* Education Endowment Foundation.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., Uekawa, K., Falk, A., Bloom, H. S., Doolittle, F., Zhu, P., & 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.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., Garrett, R., Yang, R., & 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.

Gerde, H. K., Duke, N. K., Moses, A. M., Spybrook, J., & Shedd, M. K. (2014). How much for whom? Lessons from an efficacy study of modest professional development for child care providers. *Early Education and Development*, *25*(3), 421-441.

Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S., & Santoro, L. E. (2010). Teacher study group: Impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *American Educational Research Journal*, *47*(3), 694-739.

Goodson, B., Wolf, A., Bell, S., Turner, H., & Finney, P. B. (2010). *The Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB): Kindergarten Final Evaluation Report* (NCEE 2010–4014). National Center for Education Evaluation and Regional Assistance.

Gorard, S., Siddiqui, N. & See, B. H. (2015). *Philosophy for children.* Education Endowment Foundation.

Greenleaf, C. L., Litman, C., Hanson, T. L., Rosen, R., Boscardin, C. K., Herman, J., Schneider, S. A., Madden, S. & Jones, B. (2011). Integrating literacy and science in biology: Teaching and learning impacts of reading apprenticeship professional development. *American Educational Research Journal*, *48*(3), 647-717.

Hanley, P., Bohnke, J., Slavin, B., Elliott, L., & Croudace, T. (2016). *Let's Think Secondary Science: Evaluation report and executive summary*. Education Endowment Foundation.

Haring, C. D. (2016). *The effects of coaching on teacher knowledge, teacher practice and reading achievement of at-risk first grade students*.[Doctoral dissertation, University of Texas at Austin]. UT Electronic Theses and Dissertations. https://repositories.lib.utexas.edu/bitstream/handle/2152/23148/HARING-DISSERTATION-2013.pdf

Hitchcock, J., Dimino, J., Kurki, A., Wilkins, C., & Gersten, R. (2011). *The Impact of Collaborative Strategic Reading on the Reading Comprehension of Grade 5 Students in Linguistically Diverse Schools. Final Report* (NCEE 2011-4001). National Center for Education Evaluation and Regional Assistance.

Humphrey, N., Ra, H., Ashworth, E., Frearson, K., Black, L., & Petersen, K. (2018). *Good Behaviour Game Evaluation Report and Executive Summary*. Education Endowment Foundation.

Institute for Effective Education (2016). *Teacher Effectiveness Enhancement Programme*. Education Endowment Foundation.

Jaciw, A. P., Hegseth, W. M., Lin, L., Toby, M., Newman, D., Ma, B. & Zacamy, J. (2016). Assessing impacts of Math in Focus, a "Singapore Math" program. *Journal of Research on Educational Effectiveness*, *9*(4), 473-502.

Jaciw, A. P., Schellinger, A. M., Lin, L., Zacamy, J., & Toby, M. (2016). *Effectiveness of Internet-Based Reading Apprenticeship Improving Science Education (" "iRAISE"): A Report of a Randomized Experiment in Michigan and Pennsylvania. Research Report*. Empirical Education Inc.

Jacob, B. (2017). When evidence is not enough: Findings from a randomized evaluation of Evidence-Based Literacy Instruction (EBLI). *Labour Economics*, *45*, 5–16.

Jay, T., Willis, B., Thomas, P., Taylor, R., Moore, N., Burnett, C., Merchant, G. & Stevens, A. (2017). *Dialogic teaching: Evaluation report and executive summary*. Education Endowment Foundation.

Jayanthi, M., Gersten, R., Taylor, M. J., Smolkowski, K., & Dimino, J. (2017). *Impact of the Developing Mathematical Ideas Professional Development Program on Grade 4 Students' and Teachers' Understanding of Fractions* (REL 2017-256). Regional Educational Laboratory Southeast.

Jayanthi, M., Dimino, J., Gersten, R., Taylor, M. J., Haymond, K., Smolkowski, K., & Newman-Gonchar, R. (2018). The impact of teacher study groups in vocabulary on teaching practice, teacher knowledge, and student vocabulary knowledge: A large-scale replication study. *Journal of Research on Educational Effectiveness*, *11*(1), 83–108.

Jerrim, J. & Vignoles, A. (2016). The link between East Asian 'mastery' teaching methods and English children's mathematics skills. *Economics of Education Review*, *50*, 29–44.

Johanson, M., Justice, L. M. & Logan, J. (2016). Kindergarten impacts of a preschool language-focused intervention. *Applied developmental science*, *20*(2), 94–107.

Kinzie, M. B., Whittaker, J. V., Williford, A. P., DeCoster, J., McGuire, P., Lee, Y. & Kilday, C. R. (2014). MyTeachingPartner-Math/Science pre-kindergarten curricula and teacher supports: Associations with children's mathematics and science learning. *Early Childhood Research Quarterly*, *29*(4), 586-599.

Kitmitto, S., González, R., Mezzanote, J., & Chen, Y. (2018). *Thinking, Doing, Talking Science Evaluation report and executive summary.* Education Endowment Foundation.

Kraft, M. A. & Hill, H. C. (2020). Developing ambitious mathematics instruction through web-based coaching: A randomized field trial. *American Educational Research Journal*, *57*(6), 2378-2414.

Kushman, J., Hanita, M. & Raphael, J. (2011). *An Experimental Study of the Project CRISS Reading Program on Grade 9 Reading Achievement in Rural High Schools Final Report* (NCEE 2011–4007). National Center for Education Evaluation and Regional Assistance.

Landry, S. H., Zucker, T. A., Taylor, H. B., Swank, P. R., Williams, J. M., Assel, M., Crawford, A., Huang, W., Clancy-Menchetti, J., Lonigan, C. J., Phillips, B. M., Eisenberg, N., Spinrad, T. L., de Villiers, P., Barnes, M., Starkey, P. & Klein, A. (2014). Enhancing early child care quality and learning for toddlers at risk: the responsive early childhood program. *Developmental psychology*, *50*(2), 526–541.

Lewis Presser, A., Clements, M., Ginsburg, H. & Ertle, B. (2015). Big math for little kids: The effectiveness of a preschool and kindergarten mathematics curriculum. *Early education and development*, *26*(3), 399–426.

Llosa, L., Lee, O., Jiang, F., Haas, A., O'Connor, C., Van Booven, C. D. & Kieffer, M. J. (2016). Impact of a large-scale science intervention focused on English language learners. *American Educational Research Journal*, *53*(2), 395–424.

Lonigan, C. J., Farver, J. M., Phillips, B. M. & Clancy-Menchetti, J. (2011). Promoting the development of preschool children's emergent literacy skills: A randomized evaluation of a literacy-focused curriculum and two professional development models. *Reading and writing*, *24*(3), 305-337.

Lord, P., Rabiasz, A., & Styles, B. (2017). *'Literacy Octopus' Dissemination Trial: Evaluation Report and Executive Summary.* Education Endowment Foundation.

Martin, T., Brasiel, S. J., Turner, H., & Wise, J. C. (2012). *Effects of the Connected Mathematics Project 2 (CMP2) on the Mathematics Achievement of Grade 6 Students in the Mid-Atlantic Region. Final Report* (NCEE 2012–4017). National Center for Education Evaluation and Regional Assistance.

Matsumura, L. C., Garnier, H. E. & Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model. *Learning and Instruction*, *25*, 35-48.

Mattera, S., Jacob, R., & Morris, P. (2018). *Strengthening children's math skills with enhanced instruction: The impacts of Making Pre-K Count and High 5s on kindergarten outcomes.* MDRC.

McMaster, K. L., Lembke, E. S., Shin, J., Poch, A. L., Smith, R. A., Jung, P. G., Allen, A. A., & Wagner, K. (2019). *Supporting teachers' use of data-based instruction to improve students' early writing skills*. https://files.eric.ed.gov/fulltext/ED595445.pdf

McNally, S. (2014). *Hampshire Hundreds: Evaluation Report and Executive Summary*. Education Endowment Foundation.

Meyers, C. V., Molefe, A., Brandt, W. C., Zhu, B., & Dhillon, S. (2016). Impact results of the eMINTS professional development validation study. *Educational Evaluation and Policy Analysis*, *38*(3), 455–476.

Murphy, R., Weinhardt, F., Wyness, G. & Rolfe, H. (2017). *Lesson Study: Evaluation Report and Executive Summary*. Education Endowment Foundation.

Neuman, S. B., Pinkham, A. & Kaefer, T. (2015). Supporting vocabulary teaching and learning in prekindergarten: The role of educative curriculum materials. *Early Education and Development*, *26*(7), 988-1011.

Newman, D., Finney, P. B., Bell, S., Turner, H., Jaciw, A. P., Zacamy, J. L., & Gould, L. F. (2012). *Evaluation of the Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI). Final Report* (NCEE 2012–4008). National Center for Education Evaluation and Regional Assistance.

O'Hare, L., Stark, P., Cockerill, M., Lloyd, K., McConnellogue, S., Gildea, A., Biggart, A., Connolly, P. & Bower, C. (2019). *Reciprocal Reading: Evaluation Report.* Education Endowment Foundation.

Olson, C. B., Kim, J. S., Scarcella, R., Kramer, J., Pearson, M., van Dyk, D. A., Collins, P. & Land, R. E. (2012). Enhancing the interpretive reading and analytical writing of mainstreamed English learners in secondary school: Results from a randomized field trial using a cognitive strategies approach. *American Educational Research Journal*, *49*(2), 323–355.

Olson, C. B., Woodworth, K., Arshan, N., Black, R., Chung, H. Q., D'Aoust, C., Dewar, T., Friedrich, L., Godfrey, L., Land, R., Matuchniak, T., Scarcella, R. & Stowell, L. (2020). The pathway to academic success: Scaling up a text-based analytical writing intervention for Latinos and English learners in secondary school. *Journal of Educational Psychology*, *112*(4), 701–717.

Pianta, R., Hamre, B., Downer, J., Burchinal, M., Williford, A., Locasale-Crouch, J., Howes, C., La Paro, K. & Scott-Little, C. (2017). Early childhood professional development: Coaching and coursework effects on indicators of children's school readiness. *Early Education and Development*, *28*(8), 956–975.

Piasta, S. B., Logan, J. A., Pelatti, C. Y., Capps, J. L. & Petrill, S. A. (2015). Professional development for early childhood educators: Efforts to improve math and science learning opportunities in early childhood classrooms. *Journal of educational psychology*, *107*(2), 407–422.

Portes, P. R., Canche, M. S. G. & Stollberg, R. (2016). *Early RCT Findings for ELL Elementary Student Learning Outcomes after a Two-Year Pedagogical Intervention.* Society for Research on Educational Effectiveness.

Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on head start teachers and children. *Journal of educational psychology*, *102*(2), 299–312.

Presser, A. L., Clements, M., Ginsburg, H. & Ertle, B. (2012). *Effects of a preschool and kindergarten mathematics curriculum: Big Math for Little Kids*. Education Development Center, Inc.

Randel, B., Beesley, A. D., Apthorp, H., Clark, T. F., Wang, X., Cicchinelli, L. F. & Williams, J. M. (2011). *Classroom Assessment for Student Learning: Impact on Elementary School Mathematics in the Central Region Final Report* (NCEE 2011–4005). National Center for Education Evaluation and Regional Assistance.

Reinke, W. M., Herman, K. C. & Dong, N. (2018). The incredible years teacher classroom management program: Outcomes from a group randomized trial. *Prevention Science*, *19*(8), 1043–1054.

Rienzo, C., Rolfe, H. & Wilkinson, D. (2015). *Changing Mindsets: Evaluation Report and Executive Summary*. Education Endowment Foundation.

Robinson-Smith, L., Fairhurst, C., Stone, G., Bell, K., Elliott, L., Gascoine, L., Hallett, S., Hewitt, C., Hugill, J., Torgerson, C., Torgerson, D., Menzies, V. & Ainsworth, H. (2018). *Maths Champions: 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.

Simmons, D., Hairrell, A., Edmonds, M., Vaughn, S., Larsen, R., Willson, V., Rupley, W. & Byrns, G. (2010). A comparison of multiple-strategy methods: Effects on fourth-grade students' general and content-specific reading comprehension and vocabulary development. *Journal of Research on Educational Effectiveness*, *3*(2), 121–156.

Sloan, S., Gildea, A., Miller, S. & Thurston, A. (2018). *Zippy's Friends: Evaluation report and executive summary*. Education Endowment Foundation.

Snow, P. C., Eadie, P. A., Connell, J., Dalheim, B., McCusker, H. J. & Munro, J. K. (2014). Oral language supports early literacy: A pilot cluster randomized trial in disadvantaged schools. *International Journal of Speech-Language Pathology*, *16*(5), 495–506.

Snyder, P., Hemmeter, M. L., McLean, M., Sandall, S., McLaughlin, T. & Algina, J. (2018). Effects of professional development on preschool teachers' use of embedded instruction practices. *Exceptional Children*, *84*(2), 213–232.

Stone III, J. R., Alfeld, C. & Pearson, D. (2008). Rigor and relevance: Enhancing high school students' math skills through career and technical education. *American Educational Research Journal*, *45*(3), 767–795.

Styles, B., Stevens, E., Bradshaw, S. & Clarkson, R. (2014). *Vocabulary Enrichment Intervention Programme: Evaluation Report and Executive Summary.* Education Endowment Foundation.

Sutherland, A., Broeks, M., Sim, M., Brown, E., Iakovidou, E., Ilie, S., Jarke, H. & Belanger, J. (2019). *Digital feedback in primary maths: Evaluation report and executive summary.* Education Endowment Foundation.

Taylor, J. A., Getty, S. R., Kowalski, S. M., Wilson, C. D., Carlson, J. & Van Scotter, P. (2015). An efficacy trial of research-based curriculum materials with curriculum-based professional development. *American Educational Research Journal*, *52*(5), 984–1017.

Tolan, P., Elreda, L. M., Bradshaw, C. P., Downer, J. T., & Ialongo, N. (2020). Randomized trial testing the integration of the Good Behavior Game and MyTeachingPartner™: The moderating role of distress among new teachers on student outcomes. *Journal of school psychology*, *78*, 75–95.

Torgerson, D., Torgerson, C., Mitchell, N., Buckley, H., Ainsworth, H., Heaps, C. & Jefferson, L. (2014). *Grammar for Writing: Evaluation Report and Executive Summary.* Education Endowment Foundation.

Tracey, L., Boehnke, J., Elliott, L., Thorley, K., Bowyer-Crane, C. & Ellison, S. (2019). *Grammar for Writing: Evaluation report and executive summary*. Education Endowment Foundation.

Speckesser, S., Runge, J., Foliano, F., Bursnall, M., Hudson-Sharp, N., Rolfe, H. & Anders, J. (2018). *Embedding formative assessment: Evaluation report and executive summary*. Education Endowment Foundation.

van der Scheer, E. A. & Visscher, A. J. (2018). Effects of a data-based decision-making intervention for teachers on students' mathematical achievement. *Journal of Teacher Education*, *69*(3), 307–320.

Vaughn, S., Roberts, G., Swanson, E. A., Wanzek, J., Fall, A. M. & Stillman-Spisak, S. J. (2015). Improving middle-school students' knowledge and comprehension in social studies: A replication. *Educational Psychology Review*, *27*(1), 31–50.

Vernon-Feagans, L., Kainz, K., Hedrick, A., Ginsberg, M. & Amendum, S. (2013). Live webcam coaching to help early elementary classroom teachers provide effective literacy instruction for struggling readers: The Targeted Reading Intervention. *Journal of Educational Psychology*, *105*(4), 1175–1187.

Wasik, B. A. & Hindman, A. H. (2011). Improving vocabulary and pre-literacy skills of at-risk preschoolers through teacher professional development. *Journal of Educational Psychology*, *103*(2), 455–469.

Wasik, B. A. & Hindman, A. H. (2020). Increasing preschoolers' vocabulary development through a streamlined teacher professional development intervention. *Early Childhood Research Quarterly*, *50*, 101–113.

Whittaker, J. V., Kinzie, M. B., Vitiello, V., DeCoster, J., Mulcahy, C. and Barton, E. A. (2020). Impacts of an early childhood mathematics and science intervention on teaching practices and child outcomes. *Journal of Research on Educational Effectiveness*, *13*(2), 177–212.

Wiggins, M., Jerrim, J., Tripney, J., Khatwa, M. & Gough, D. (2019). *The RISE project: Evidence-informed school improvement.* Education Endowment Foundation.

Wilcox, M. J., Gray, S. I., Guimond, A. B. & Lafferty, A. E. (2011). Efficacy of the TELL language and literacy curriculum for preschoolers with developmental speech and/or language impairment. *Early Childhood Research Quarterly*, *26*(3), 278–294.

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.

Worth, J., Sizmur, J., Walker, M., Bradshaw, S. & Styles, B. (2017). *Teacher Observation: Evaluation Report and Executive Summary*. Education Endowment Foundation.

Wright, H., Carr, D., Wiese, J., Stokes, L., Runge, J., Dorsett, R., Heal, J. & Anders., J. (2020). *URLEY Evaluation Report*. Education Endowment Foundation.

Appendix 1: Methods in detail

PICO statement

The PICO (Population, Intervention, Comparison, Outcome) framework can be used to define the scope of the project.

The **population** of interest is post-initial qualification teachers working in formal settings (for example, early years, schools, colleges), who provide instruction to pupils between the ages of 3 and 18.

The **intervention** of interest is teacher PD. We defined this as any structured, facilitated activity for teachers intended to increase their teaching ability. We focused on formal, structured programmes because informal and unstructured programmes are difficult to evaluate and difficult for school leaders or policymakers to promote. Note, however, that our definition does not exclude formal, structured approaches to peer collaboration or support (for example, Papay et al., 2016; West et al., 2017).

The **comparator** of interest is teachers who do not receive the PD intervention being evaluated. This may mean they are receiving no PD or that they continue to receive 'normal' PD under a 'business as usual' (BAU) condition.

The **primary outcome** is pupil achievement, as captured by scores on standardised (as opposed to research design or adapted) academic tests. The **secondary outcome** is change in teacher classroom practice, as captured by a validated observation instrument administered by a blind observer. Ultimately, however, we found too few studies using our secondary outcome measure and were forced to drop this from our analysis.

Information management

All records of research identified by searches were uploaded to the specialist systematic review software, EPPI-Reviewer 4, for duplicate stripping, screening, document management, data extraction and data analysis (Thomas et al., 2020).

Initial search strategy

The search process was designed and implemented by CS with contributions from the research team. The database search incorporated terms designed to capture three concepts that need to be present in each of the citations: (1) teachers (e.g. 'teachers', 'educators'); (2) professional development (e.g. 'inservice training', 'professional learning'); and (3) randomised controlled trials (RCTs) or terms to capture process evaluations of RCTs. Terms and phrases were used to search the title, abstract and controlled vocabulary fields of the databases. The searches were limited to English language and some terms were used to exclude on pre-service teachers, higher education teachers, and adult education teachers. Connolly et al. (2018) helped inform the development of search terms for RCTs.

An example of the database search, as applied in ERIC (EBSCO interface) can be found in Appendix 2. The following databases were searched: Australian Education Index (Proquest); British Education Index (BEI); EconLit (EBSCO); Education Resources Information Center (ERIC) (EBSCO); Education Abstracts (EBSCO); Educational Administration Abstracts (EBSCO); EPPI-Centre database of education research; ProQuest Dissertations & Theses; PsycINFO (OVID); Teacher Reference Center (EBSCO); Google Scholar.

We also searched the reference lists of previous systematic reviews (defined as reviews that themselves searched at least two databases) in this area, including: Cordingley et al., 2015; Desimone, 2009; Dunst et al., 2015; Kennedy, 2016a; Kraft et al., 2018; Lynch et al., 2019; Rogers et al., 2020; Timperley et al., 2007; Walter & Briggs, 2012; Wei et al., 2009; Yoon et al., 2007. These reviews were known to the review team and helped ensure the comprehensiveness of the database searches.

The following websites of education repositories were browsed: Center for Coordinated Education MRDC publications; CUREE—Centre for the use of evidence and research in education; Digital Education Resource Archive; Education Endowment Foundation (EEF); EIPEE search portal; EPPI-Centre database of education research; Institute of Education Studies What Works Clearinghouse; Nuffield Foundation.

Although the EEF website was browsed in the initial search process, a supplementary search of the EEF repository was conducted by SC after the evidence mapping stage to ensure no studies were missed.

Reference-checking and forward citation searching of studies were also undertaken for the studies that were included in the synthesis and that were available in Microsoft Academic (see next section for details).

Reference-checking and forward citation of full-text includes

This process was conducted by AO. For the 121 records included in the synthesis, we conducted reference-checking and forward citation checking in Microsoft Academic Search, within the EPPI-Reviewer interface. First, the records had to be identified in Microsoft Academic, of which we were able to match 108 of the records. Then, we applied the 'cited by' (for forward citation chasing) and 'bibliography' (for reference list checking) functions using the MAG searches. This yielded 2,348 and 1,549 records, respectively. Even after de-duplicating, there was a very large total number of records (3,614) to screen.

Given our time frame and resources available, we took the pragmatic approach to apply two machine learning classifiers within the EPPI-Reviewer interface on this corpus of records:

1. A classifier trained on all the included studies, with 'Exclude on PD intervention' at the title and abstract screening stage as the exclude comparator. This was to help identify studies that 'looked liked' our includes.
2. A classifier trained on a very large Cochrane RCT database. This was to help protect against 'more of the same' bias as the classifier was purely looking for terms related to RCTs.

Any record that was given a score of ≥ 50% on *either* classifier was then put forward for manual screening on title and abstract. From this process, two papers were ultimately included on full-text. See Appendix 3 for the flow of literature diagram for this process.

Screening and collation of studies

Screening for inclusion in the review occurred in two stages. First, the inclusion criteria were applied to titles and abstracts (T&A). The criteria were piloted on a sample of studies before being applied. The purpose of piloting is primarily to refine the coding tool and gain familiarity with the inclusion/exclusion criteria. We used a random sample of 387 records that were double screened (a few items were accidentally triple screened). Interrater agreement at this stage was 98.2%

agreement regarding whether to include versus exclude; all were able to be resolved easily through discussion. The inclusion rate for this batch was 8.8% (n = 43).

We planned to use text mining built into our in-house EPPI-Reviewer software to prioritise (rank order) all records identified through the search phase. The use of prioritised screening has been widely evaluated (see systematic review by O'Mara-Eves et al., 2015) and is an accepted approach to expediting the T&A screening process. Prioritising records has been found to significantly speed up the screening process (O'Mara-Eves et al., 2015; Shemilt et al., 2014). Using prioritised screening was important in order to complete the review by the deadline.

The approach to prioritised screening used in EPPI-Reviewer is known is an active learning approach, which is an iterative process whereby the accuracy of the predictions made by the machine is improved through interaction with reviewers (Brunton et al., 2017; Thomas et al., 2011). The human reviewer provides an initial sample of include/exclude decisions that the machine 'learns' from. The machine subsequently generates a ranked list and requests the reviewer to provide decisions on items high in the list, that it will learn the most from. The machine then adapts its decision rule including the information from the additional items and generates a new list of items for the reviewer to screen. An independent evaluation demonstrates that priority screening using EPPI-Reviewer can provide efficiency gains: in a retrospective evaluation of six systematic reviews, EPPI-Reviewer's priority screening algorithm estimated potential reductions in screening of between 9% and 60% (Tsou et al., 2020).

We used two indicators when deciding when to stop screening the ranked list. The simplest indicator that we have exhausted the relevant records in the ranked list is the observation that no new relevant records will be identified as screening down the list progresses. A graphical display of the inclusion rate over time is observable during screening in EPPI-Reviewer and reaching a long plateau (over 500 records with no new includes) may indicate that the remainder of the list is not relevant.

The more sophisticated indicator was a comparison with a predicted number of eligible studies based on the calculation of a *baseline inclusion rate* (Shemilt et al., 2014), which was established as follows. First, a power calculation was used to calculate the number of references required to be screened randomly in order to provide an estimate for attaining the initial predicted inclusion rate, based on a margin of error of 3% at 95% conference interval. The power calculation indicated how many records would need to be screened to confidently establish the baseline inclusion rate given these parameters. Based on the 387 studies screened in the piloting batch, with an inclusion rate of 8.8%, we determined we would need to screen at least a further 319 records to determine the baseline inclusion rate.

We went further than 319: a random sample of 476 citations was double-screened to calculate the baseline inclusion rate. Interrater agreement at this stage was 98.1% agreement regarding whether to include versus exclude. All disagreements were able to be resolved easily through discussion. This baseline inclusion rate should be a reasonable indicator of the likely number of eligible studies in the corpus of records had we randomly screened the records. Once that number of includable studies is met or exceeded, we can reasonably conclude that we have identified almost all the relevant records.

However, we observed a reduction in the inclusion rate from our initial pilot batch of studies to our sample for establishing the baseline inclusion rate: 8.8% versus 4.6%, respectively. This meant that the inclusion rate in this second batch of studies was outside the boundaries specified in our initial power calculation. It is a common phenomenon across reviews that screeners are less conservative as they progress through screening, meaning that they are less likely to include very ambiguous cases as they become more confident with the screening tool and what an include or exclude 'looks like'. As such, the decreased inclusion rate from batch 1 to batch 2 is not unexpected. Also, more records were screened in batch 2 than required by the power calculation, and a larger ME would have required an even smaller sample.

Combined, these pieces of information indicate that the inclusion rate for batch 2 is likely to be more accurate than batch 1. We therefore accepted the baseline inclusion rate of 4.6%, indicating a likely number of eligible includes of 253 (4.6% of 5506). However, there is a margin of error around this estimate.

At this point, having double screened a random selection of 15.4% of records with very high agreement, prioritised screening was switched on and we changed to single screening mode.

Single screening of titles and abstracts continued until our two conditions were met: not observing new includes in the prioritised list ('the plateau') and exceeding the initial predicted number of includes (prediction = 253; actual = 347). We also manually screened all records that did not have an abstract, as these may not be adequately dealt with by the prioritisation algorithm. In total, 3,140 records (57%) were manually screened at this stage; 2,373 of the 5,515 (lowest priority) records were not manually screened.

Full reports were obtained for those studies that appeared to meet the criteria or where there is insufficient information to be certain. These were screened on the basis of the full-text article.

Where multiple papers reported the same study, then we retrieved all versions and assigned one as the 'master' document. Master documents were usually the most recent or detailed journal article (for example, associated dissertations were treated as linked studies). Where information was not reported in the master document, then the linked documents were checked to see if they contained the relevant information.

Where one document reported results for more than one teacher PD intervention in addition to a comparator condition, then we used outcome data from the most intensive intervention in the analyses. By most intensive, we mean that the other versions of the intervention include (1) some but not all of the same components, and (2) no additional components. Where it was not possible to clearly distinguish more and less intensive versions, we picked a version at random.

Inclusion/exclusion criteria for the evidence map

To be included in the evidence map, studies needed to meet all of the following inclusion criteria:

1. Publication year: We restricted our search to studies published during or after 2002. Specifying a start date was necessary in order to ensure that we completed the project by the deadline and within the available budget. We chose this specific start date because it is the year in which the Institute of Education Sciences in the US was established, which marked the beginning of a new era in terms of the funding and conduct of rigorous experimental evaluations of PD in education (Hedges & Schauer, 2018). The main searches were undertaken between 12 and 16 November 2020.
2. Language: We included studies written in English and excluded studies in languages other than English. This was necessary due to resource constraints.
3. Reporting: We included studies reported in journal paper, working paper or institutional report formats, as well as doctoral theses that could be obtained via current UCL

subscriptions. We excluded conference papers or extended abstracts on the grounds that they do not contain enough information to assess quality or to extract sufficiently detailed information about intervention components.

4. Geography: We included programmes implemented in OECD countries.[11] This was necessary to ensure some level of comparability between the contexts in which the programmes were conducted.
5. Intervention: We included studies that evaluated teacher PD programmes, as defined above. We excluded programmes that incorporated a change in the pupil:teacher ratio (for example, training teachers in the delivery of small-group or one-to-one tuition). Our focus is on the relationship between certain characteristics of PD and the impact of that PD on pupil achievement. Small-group tuition is known to be highly effective in and of itself (Nickow, Oreopoulos & Quan, 2020) so including evaluations of PD that also incorporated small-group tuition would confound the relationship between the PD characteristics and the impact of the PD. In line with our definition of PD, we excluded programmes that aimed only to briefly familiarise teachers with educational technology or curriculum materials.
6. Intervention population: We included studies of qualified teachers working in formal settings (for example, early years settings, schools, colleges) who teach pupils between the ages of 3 and 18. We excluded studies in higher education (HE) settings.
7. Design: We included RCTs and excluded all other evaluation designs. The justification for this is that (well implemented) RCTs are the only design that reliably provide unbiased causal impact estimates—essential for our synthesis approach.
8. Outcomes: We included studies that measured pupil achievement using standardised tests, in any school curriculum subject. We excluded studies that only measured achievement using observational protocols (for example, Assessment of Scientific Argumentation in the Classroom; Sampson et al., 2012) or that used holistic teacher judgements, as opposed to quantitative aggregation of marks from multiple test items (for example, the Early Years Foundation Stage; DfE, 2020).
9. Comparison group: We included studies that provided no PD, business as usual PD, or waitlist intervention for the control group.

The actual codes as applied in EPPI-Reviewer software are contained within Box 1. For the evidence map, the same codes were applied for both the title and abstract stage and the full-text document stage of screening. At the end of this process, we included 186 records in the evidence map.

Box 1: Screening codes as applied in EPPI-Reviewer

- EX date <2002 or >Nov 2020
- EX not English language
- EX not journal, working paper, or doctoral thesis
  Include studies reported in journal paper or working paper format. Include doctoral theses that can be obtained via current UCL subscriptions. Exclude conference papers or extended abstracts and master's theses.

[11] https://www.oecd.org/about/document/list-oecd-member-countries.htm

- EX not OECD

  Study outside of list of included countries.

  OECD countries are Australia, Austria, Belgium, Canada, Chile, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, United Kingdom and United States.
- EX not PD intervention

  Exclude interventions that train teachers in the delivery of small-group or one-to-one tuition. Also exclude interventions that aim only to briefly familiarise teachers with educational technology or curriculum materials.
- EX not population

  The subjects of the intervention are NOT qualified teachers working in formal settings (e.g. early years settings, schools, colleges), who teach pupils between the ages of 3 and 18. Exclude studies in higher education (HE) settings.
- EX not an RCT
- EX no test score outcome

  Exclude studies that use observational protocols (e.g. Assessment of Scientific Argumentation in the Classroom; Sampson et al., 2012). In addition, exclude studies that use holistic teacher judgements, as opposed to quantitative aggregation of marks from multiple test items (e.g. the Early Years Foundation Stage; DfE, 2020).
- EX not comparator condition

  Include studies that provide no PD, business as usual PD, or waitlist intervention for the control group.
- INCLUDE for second opinion

  Unsure about this one. Need to speak to ___ for clarification.
- INCLUDE document

Inclusion/exclusion criteria for the synthesis

After the map was produced, we held a consultation meeting with the EEF to determine which studies to take through to data synthesis and assessment of study quality. We agreed to exclude those studies that did not use standardised tests (i.e. researcher developed or adapted tests) as their main outcome measures. In contrast, studies with outcomes measured through high stakes standardised exams (such as GCSE in the UK) and low stakes standardised exams were retained for inclusion in the in-depth synthesis.

In addition, at this stage, several records were removed from the dataset for being linked or duplicate records. A small number were excluded using the original criteria as in Box 1 after careful examination during the data extraction process.

Finally, studies could only be included in the statistical meta-analysis if we were able to calculate an effect size. See section on ‘Effect size calculation’ for details.

Data extraction process

The data extraction tool for the map was piloted to check whether additional guidance notes or definitions were needed. We then conducted independent double data extraction on 20 randomly selected studies. Two reviewers independently extracted study characteristics and numerical outcome data from studies meeting the eligibility criteria. In agreement meetings, the reviewers resolved discrepancies by discussion, and sought input from a third reviewer if the disagreement could not be resolved. Where necessary, the data extraction guidance was refined in order to ensure consistent understanding between reviewers. By this stage, we had set up a team Slack channel (an organised instant messaging app), so some agreements, discussions, and refinements occurred on Slack.

Each record in the 20-study sample needed to be resolved (i.e. full agreement on all assigned codes in each study) before reviewers could progress to independent extraction. In the end, 47 records were independently double extracted on the mechanisms code set, which is more than the minimum 20% (n = 37) double extraction that we had proposed in the protocol. Checks and consultations on many other studies after the official double screened set also occurred to assure quality of the codings.

Details of the data extraction tools are contained in the following sections.

Mapping data extraction tool

Evidence maps involve a systematic search of a broad field, followed by the coding of the studies identified based on study characteristics. We coded the studies that met the full-text inclusion criteria using the following codes:

- Research design RCT explicitly stated
  - Yes
- Age Group/s
  - Early years/pre-kindergarten
    Early years/pre-kindergarten
  - Primary/elementary
    Primary/elementary
  - Secondary/middle/high school/Sixth Form
    Secondary/middle/high school/Sixth Form
- Country
  - US
  - UK
  - Italy
  - France
  - Australia
  - Canada
  - Israel
  - Unclear or not reported
  - Sweden
  - Colombia

    - Chile
    - Denmark
    - New Zealand
    - Germany
    - Netherlands
    - Spain
    - Not UK or US
- Subject targeted
    - English/native language/speech/literacy
    - Maths
    - Science
    - Other subjects
    - Generic (cross-curricular)
- Test type
    - High stakes, standardised exam
    - Low stakes, standardised exam
    - Test developed or adapted by researchers or teachers
    - Unclear
- Randomisation
    - What was randomised?
        - Students
        - Teachers/classes
        - Classes
        - Grades/year groups
        - Schools
        - Other (please specify)
        - Not reported
        - Sections of content (within teacher)
        - Districts/local authorities/regions
    - Number of units randomised
- Process evaluation reported?
    - No
    - Yes
- Curricular reform component in intervention?
  PD intervention also includes curricular reform component
    - No
    - Yes
- New educational technology component in intervention?
  PD intervention also includes new educational technology component
    - No
    - Yes
- Linked document

Synthesis data extraction tool 1: Non-mechanism coding

After the mapping stage, and having excluded the studies without standardised attainment tests, we then applied two different code sets for the purposes of gathering data for the synthesis. The first was to code non-mechanisms information, as follows:

- Excluded on full-text
- Reports multi-arm trial (e.g. A v. B v. control)
- ~SECOND OPINION needed~ add explanatory note
- Kennedy's 'persistent problems' (select all that apply)
    - Portraying curriculum
    - Containing behaviour
    - Enlisting participation
    - Exposing student thinking
    - Unclear or other—add note
- Broad area of focus of the PD (select all that apply)
  **We will have already captured whether it has a focus on 'ed tech' or curricular change at the mapping stage*
    - Science of learning/cognitive science
    - Inquiry/discovery/problem based teaching
    - Formative assessment/responsive teaching
    - Data driven instruction
    - Unclear or other—add note
- Early-career/newly-qualified teachers?
    - Yes
    - No (mixed or experienced), OR unclear
- Dosage
    - No. of intended hours per participating teacher (specify)
    - Dosage unclear
- Overall pupil disadvantage
    - % eligible for free or reduced-price lunch
    - Pupil disadvantage unclear or not reported
- Attrition
    - Cluster level attrition (if clustered): *https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf*
        - Acceptable attrition
        - Not acceptable attrition
        - Attrition unclear
        - Not cluster randomised
    - Pupil level attrition: *https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf*
        - Acceptable attrition
        - Not acceptable attrition

        - Attrition unclear
- Analysis pre-specified?
    - Pre-registration/pre-specification/analysis plan
    - Not pre-registered
    - Not reported

Synthesis data extraction tool 2: Mechanism coding

The second synthesis data extraction tool was designed to extract information about the mechanisms in the interventions. This was coded as follows:

- Instil Insight
    - Manage cognitive load
    - Revisit material
- Motivate Goals
    - Goal setting
    - Credible source
    - Praise/reinforce
- Teach Techniques
    - Instruction
    - Practical soc. support
    - Feedback on practice
    - Modelling
    - Rehearsal
- Embed Practice
    - Prompts/cues
    - Action planning
    - Self-monitoring practice
    - Context-specific repetition
- Checked, but no mechanisms to extract

Effect size calculation

Effect sizes were calculated in the EPPI-Reviewer software. For student attainment outcome, the data were continuous test scores and were calculated using standardised mean difference effect sizes (Cohen's *d*). These were typically calculated from means and standard deviations (or SEs or CIs). In the absence of those statistics, we used (in decreasing order of preference) t- or F-statistics, or p-values. Any study results that could be converted to an effect size were used (see full list of formulae in Lipsey & Wilson, 2001). Cohen's *d* was selected as the review team knew that many studies in this literature report this as their effect size metric. Cohen's *d* can be corrected for potential small bias in studies with less than 50 participants using Hedges' *g* (Hedges, 1981). Hedges' *g* could not be calculated for two studies due to missing data. We therefore present all our results using Cohen's *d* but also present key results using Hedges' *g* for a slightly smaller sample.

Effect sizes were calculated in such a way that a positive effect size indicates a larger mean outcome for the PD intervention group, and a negative effect size indicates a larger mean outcome for the comparator.

We used the effect size for the overall experimental group, rather than for specific subgroups or cohorts. Where multiple attainment (or teacher practice change) outcomes were reported and a primary outcome measure was specified, then we used the primary outcome measure to calculate the effect size. If multiple attainment (or teacher practice change) outcomes were reported and no primary outcome was specified, we calculated multiple effect sizes and handled this using robust variance estimation to account for dependencies (Tanner-Smith & Tipton, 2013). The one exception to this is in our 'bubbleplots', which cannot be made using robust variance estimation. We therefore used either the specified primary outcome or else one randomly chosen outcome to make these plots.

Effect sizes were primarily calculated by SS. AOE checked the effect size calculations of at least 30 of the 121 studies included in the synthesis, usually in cases where SS was concerned that an effect size was incalculable. In total, 17 studies did not report sufficient or appropriate detail to calculate an effect size for inclusion in the review.

There were substantial dependencies in the data set: 219 effect sizes were calculated from 104 studies. This ranged from 1 to 17 effect sizes per study, with a median of three effect sizes. This is why the robust variance estimation method was used, as it explicitly handles non-independent effect sizes clustered within studies.

The distribution of the effect sizes indicated a moderate positive skewness. To test whether violations of the normality assumptions embedded in random effect meta-analysis were biasing our results, we also reported a sensitivity test based on a square root transformation of our data.

Unit of analysis issues

For studies in which randomisation occurred at the group level (for example, classes or schools) rather than the individual teacher level, we used White and Thomas's (2005) adjustment for clustering.

Data synthesis and sensitivity analysis

Statistical meta-analysis was conducted using a robust variance estimation random effects meta-analysis implemented using the robumeta command (Hedges, Tipton & Johnson, 2010; Tanner-Smith & Tipton, 2014) in the Stata software. This accounts for the dependencies between effect sizes coming from the same study and allows us to include all relevant effect sizes in the analyses.

Subgroup meta-analyses were reported to explore heterogeneity based on indicators of study quality and other methodological characteristics.

We conducted sensitivity analyses to check whether the results vary across sample restrictions based on indicators of study quality:

- high levels of attrition/missing data (defined using the NCEE 'cautious' standards[i])[12];
- number of units randomised (< 50); and
- pre-registration of the evaluation.

We conducted additional sensitivity checks to see if aspects of the systematic review methods influenced the observed effect size:

- transformation of the data to enhance normality of the distribution of effect sizes;
- the use of Hedges' *g* instead of Cohen's *d*; and
- whether the outcome was measured in the final year of the evaluation or the subsequent year.

We tested for publication bias using trim-and-fill (Duval & Tweedie, 2000), *p*-curve (Simonsohn et al., 2014a; Simonsohn et al., 2014b), and weight function methods (Vevea & Hedges, 1995), and graphically depicted a contour-enhanced funnel plot (Peters, Sutton, Jones, Abrams & Rushton, 2008). The robumeta command for the robust variance estimation models do not allow for these tests, so we conducted the trim-and-fill and the contour-enhanced funnel plot using the commands in Stata, and the *p*-curve and weight function methods were run in *R.*

We also used crisp set qualitative comparative analysis (QCA) to look for consistency of high impact across studies (Thomas et al., 2014). This was implemented using the 'fuzzy' package in Stata using cut-offs of .08, .10, and .12. The conditions in the QCA were the four mechanisms from the IGTP model: insights (I), goal-directed behaviours (G), techniques (T), and practice (P).

Protocol deviations

We deviated from the protocol in the following ways:

- The ERIC database search strategy was expanded upon slightly. Additional databases searched were Australian Education Index and British Education Index. We did not identify additional systematic reviews from our searches to screen their references as we had already identified a substantial number of reviews prior to screening and felt resources were better spent elsewhere.
- We proposed to use a truth table to identify commonly occurring clusters of mechanisms. However, we ended up with more mechanisms than we expected (14), which meant the truth table had more rows (representing every possible combination of the 14 mechanisms) than we had studies. This made using the truth table to identify clusters of mechanisms infeasible. Instead of using a truth table, we relied on existing literature and theory to identify clusters of mechanisms.
- The protocol did not specify how we would deal with studies identified through the supplementary citation chasing conducted after the list of eligible studies was identified. Given the large number of studies found from forward citation and reference list checks in MAG, at a late stage in the review, we took the pragmatic approach to apply two different

[12] https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

classifiers to narrow the screening sample down. This is described in the 'Reference checking' section of Appendix 1.

- In the protocol, we stated that we would run a sensitivity analysis on our main results by excluding studies with high levels of non-compliance (top ranked 5% of interventions on teacher non-attendance). However, we found that very few studies reported this, rendering the analysis infeasible.
- In the protocol, we stated that we would run a sensitivity analysis on our main results looking at studies randomised at the individual versus cluster level and studies in dissertations versus journal articles. However, in both cases we found very few studies in the former category, rendering the analysis infeasible.
- In the protocol, we stated our intention to use forest plots. However, the number of studies exceeded our expectations and the forest plots would have been prohibitively hard to display or to interpret.
- In addition to the three publication bias checks that we proposed in the protocol, we also produced and reported a contour-enhanced funnel plot (Peters et al., 2008). Given the number of studies included in the review for which we could not calculate an effect size (n = 17), we felt it was helpful to consider whether any hypothetically missing studies are likely to be due to non-reporting biases related to significance levels, which is a key assumption of publication bias.
- In the protocol, we planned to report heterogeneity statistics for our meta-analytic averages. However, we had not anticipated that these cannot be calculated using the robust variance estimation command in Stata.

Timeline

Dates	Activity	Staff responsible/leading
Sept 2020–Jan 2021	Writing protocol	Sam Sims Alison O’Mara-Eves Claire Stansfield Jo Van Herwegen
Oct 2020	Search	Claire Stansfield
Nov–Feb 2020	Screening	Sam Sims Harry Fletcher-Wood Sarah Cottingham Alison O’Mara-Eves
Jan–April 2021	Theory	Harry Fletcher-Wood Sam Sims
Feb–April 2021	Non-mechanism coding	Sarah Cottingham Sam Sims Alison O’Mara-Eves
Feb–April 2021	Mechanisms coding	Harry Fletcher-Wood Sam Sims Alison O’Mara-Eves
May 2021	Review of implementation	Harry Fletcher-Wood
April–June 2021	Data cleaning and analysis	Alison O’Mara-Eves Sam Sims Jake Anders
May–July 2021	Writing report	Sam Sims Harry Fletcher-Wood Alison O’Mara-Eves

Dates	Activity	Staff responsible/leading
		Sarah Cottingham Jo Van Herwegen

Appendix 2: Example search terms

Database search for ERIC (EBSCO interface)

No. of results run on 2/11/2020: 3,763

Key: N = within *n words* proximity to; *=truncation

#

S51 S48 OR S50

S50 S46 AND S49 Limiters - Date Published: 20020101-20201231; Language: English

S49 DE "Professional Education" OR DE "Teacher Improvement" OR DE "Continuing Education" OR DE "Mandatory Continuing Education" OR DE "Teacher Workshops" OR DE "Professional Training" OR DE "Inservice Education" OR DE "Professional Continuing Education" OR DE "Professional Development" OR DE "Faculty Development" OR DE "Inservice Teacher Education" OR DE "Teacher Competencies" OR DE "Teacher Collaboration" Limiters - Date Published: 20020101-20201231; Education Level: Early Childhood Education, Elementary Education, Elementary Secondary Education, Grade 1, Grade 2, Grade 3, Grade 4, Grade 5, Grade 6, Grade 7, Grade 8, Grade 9, Grade 10, Grade 11, Grade 12, High School Equivalency Programs, High Schools, Intermediate Grades, Junior High Schools, Kindergarten, Middle Schools, Preschool Education, Primary Education, Secondary Education; Language: English

S48 S47

S47 S27 AND S46 Limiters - Date Published: 20020101-20201231

S46 S28 OR S29 OR S30 OR S31 OR S32 OR S33 OR S34 OR S35 OR S36 OR S37 OR S38 OR S39 OR S40 OR S41 OR S42 OR S43 OR S44 OR S45

S45 AB (random* N5 (assign* OR conceal* OR allocat* OR comparison* OR control* OR experiment* OR trial OR treatment* OR intervention*)) OR AB((randomly* OR randomi*) N5 educators) OR AB((randomly* OR randomi*) N5 teachers)

S44 TI ((Random* AND ("controlled study" OR "impact evaluation" OR "program* evaluation" OR "impact assessment" OR "outcome evaluation" OR "process evaluation" OR "effectiveness evaluation" OR "outcome assessment" OR "outcome study" OR "process evaluative method*" OR "stage of change" OR "stages of change" or "process evaluation" OR (program* N2 implementation))) OR AB ((Random* AND ("controlled study" OR "impact evaluation" OR "program* evaluation" OR "impact assessment" OR "outcome evaluation" OR "process evaluation" OR "effectiveness evaluation" OR "outcome assessment" OR "outcome study" OR "process evaluative method*" OR "stage of change" OR "stages of change" or "process evaluation" OR (program* N2 implementation)))

S43 AB (prospective* N2 control*) OR (prospective* N2 comparison*) OR (prospective* N2 trial)

S42 AB (Random* N3 (evaluat* OR design OR study OR studies OR group OR groups OR trial OR trials OR comparison OR control OR controlled OR comparative OR intervention*))

S41 AB "Intervention group*" OR "Intervention arm" OR "Intervention arms" OR (Intervention N2 teachers) OR (intervention N2 educators) OR (intervention N2 classrooms) OR (intervention N2 classes) OR "Intervention participants" OR "Intervention condition"

S40 AB "Comparison group*" OR "Comparison arm" OR "Comparison arms" OR "Comparison teachers" OR "Comparison educators" OR "Comparison participants" OR "Comparison condition"

S39 AB "Experimental group*" OR "Experimental arm" OR "Experimental arms" OR "Experimental teachers" OR "Experimental educators" OR "Experimental participants" OR "Experimental condition"

S38 AB "Control group*" OR "Control arm" OR "Control arms" OR "Control teachers" OR "control educators" OR "Control participants" OR "control condition" OR " waitlist control"

S37 AB ("Treatment group*" OR "Treatment arm" OR "Treatment arms" OR (Treatment N2 teachers) OR (Treatment N2 educators) OR "Treatment participants" Or "treatment condition")

S36 AB (Control OR Comparison OR Intervention OR Experiment*) AND ("Treatment as Usual" OR TAU OR "business as usual" OR BAU)

S35 AB Random* AND (Comparison N2 (trial OR trials OR study OR studies OR experiment OR design OR evaluat*))

S34 AB Random* AND (controlled N2 (trial OR trials OR study OR studies OR experiment OR design OR evaluat*))

S33 AB (Random* AND ("Controlled before and after" OR "natural experiment" OR "single-blind" OR "double-blind" OR blinded OR "effectiv* study" OR "noninferiority trial*" OR "parallel trials" OR "parallel-group*" OR placebo OR "trial subjects" OR "triple-blind" OR "two group*" OR crossover OR factorial))

S32 AB "trial registration"

S31 TI random* OR "cluster random*" OR trial

S30 TI RCT OR AB RCT

S29 TI ("randomised clinical trial" OR "randomised comparative trial" OR "randomised controlled trial" OR "randomised experiment*" OR "randomised study" OR "randomised trial" OR "randomized clinical trial" OR "randomized comparative trial" OR "randomized controlled trial" OR "randomized experiment*" OR "randomized study" OR "randomized trial") OR AB ("randomised clinical trial" OR "randomised comparative trial" OR "randomised controlled trial" OR "randomised experiment*" OR "randomised study" OR "randomised trial" OR "randomized clinical trial" OR "randomized comparative trial" OR "randomized controlled trial" OR "randomized experiment*" OR "randomized study" OR "randomized trial")

S28 DE "Randomized Controlled Trials" OR DE "Outcome Measures" OR DE "Program Evaluation" OR DE "control groups" OR DE "experimental groups"

S27 S18 OR S26

S26 S19 OR S20 OR S21 OR S22 OR S23 OR S24 OR S25

S25 TI ((teacher OR teachers OR educator OR educators) AND "continuing education") OR AB ((teacher OR teachers OR educator OR educators) N5 "continuing education")

S24 TI ((teacher OR teachers OR educator OR educators) AND "continuing development") OR AB ((teacher OR teachers OR educator OR educators) N5 "continuing development")

S23 TI ((teacher OR teachers OR educator OR educators) AND "skills program*") OR AB ((teacher OR teachers OR educator OR educators) N5 "skills program*")

S22 TI ((teacher OR teachers OR educator OR educators) AND "continuing training") OR AB ((teacher OR teachers OR educator OR educators) N5 "continuing training")

S21 TI ((teacher OR teachers OR educator OR educators) AND "continuing learning") OR AB ((teacher OR teachers OR educator OR educators) N5 "continuing learning")

S20 TI ("teacher improvement" AND (program* OR workshop* OR training)) OR AB ("teacher improvement" N5 (program* OR workshop* OR training OR intervention* OR strateg* OR scheme*))

S19 TI ("teaching improvement" AND (program* OR workshop* OR training OR intervention* OR strateg* OR scheme*)) OR AB ("teaching improvement" N5 (program* OR workshop* OR training OR intervention* OR strateg* OR scheme*))

S18 S17 AND S15

S17 S7 NOT S16

S16 TI ("pre service" OR "pre-service" OR preservice OR "prospective teacher*") NOT TI("in service" OR inservice OR "in-service")

S15 S8 OR S9 OR S10 OR S11 OR S12 OR S13 OR S14

S14 DE "Professional Education" OR DE "Teacher Improvement" OR DE "Continuing Education" OR DE "Improvement Programs" OR DE "Mandatory Continuing Education" OR DE "Teacher Workshops" OR DE "Professional Training" OR (DE "Teacher Education" NOT (DE "Preservice Teacher Education" OR "Preservice Teachers")) OR DE "Inservice Education" OR DE "Professional Continuing Education" OR DE "Professional Development" OR DE "Faculty Development" OR DE "Inservice Teacher Education" OR DE "Instructional Improvement" OR DE "Teacher Competencies" OR DE "Coaching (Performance)" OR DE "Teacher Collaboration"

S13 TI ("peer coaching" OR "peer mentoring" OR "peer collaboration") OR AB ("peer coaching" OR "peer mentoring" OR "peer collaboration") OR TI ("professional coursework") OR AB ("professional coursework") OR TI ("training workshop*") OR AB ("training workshop*") OR TI ("training program*") OR AB ("training program*") OR TI ((coaching OR mentoring) N2 program*) OR AB ((coaching OR mentoring) N2 program*)

S12 TI (inservice AND (development OR training OR learning OR education OR program* OR workshop*)) OR TI (("in service") AND (development OR training OR learning OR education OR program* OR workshop*)) OR AB (inservice N3 (development OR training OR learning OR education OR program* OR workshop*)) OR AB (("in service") N3 (development OR training OR learning OR education OR program* OR workshop*))

S11 AB ((PD N2 program*) OR (PD N2 teacher*) OR (PD N2 educator*) OR (PD N2 Intervention*) OR (PD N2 strateg*) OR (PD N2 workshop*) OR (PD N2 scheme*) OR (PD N2 initiative*) OR (PD N2 mentor*) OR (PD N2 coaching*) OR (PD N2 collabor*) OR (PD N2 formal*))

S10 TI PD

S9 TI CPD OR AB CPD

S8 TI (professional N3 (development OR training OR learning OR education)) OR AB (professional N3 (development OR training OR learning OR education))

S7 S5 OR S6

S6 (S3 NOT S4)

S5 S3 Limiters - Education Level: Early Childhood Education, Elementary Education, Elementary Secondary Education, Grade 1, Grade 2, Grade 3, Grade 4, Grade 5, Grade 6, Grade 7, Grade 8, Grade 9, Grade 10, Grade 11, Grade 12, High School Equivalency Programs, High Schools, Intermediate Grades, Junior High Schools, Kindergarten, Middle Schools, Preschool Education, Primary Education, Secondary Education, Two Year Colleges

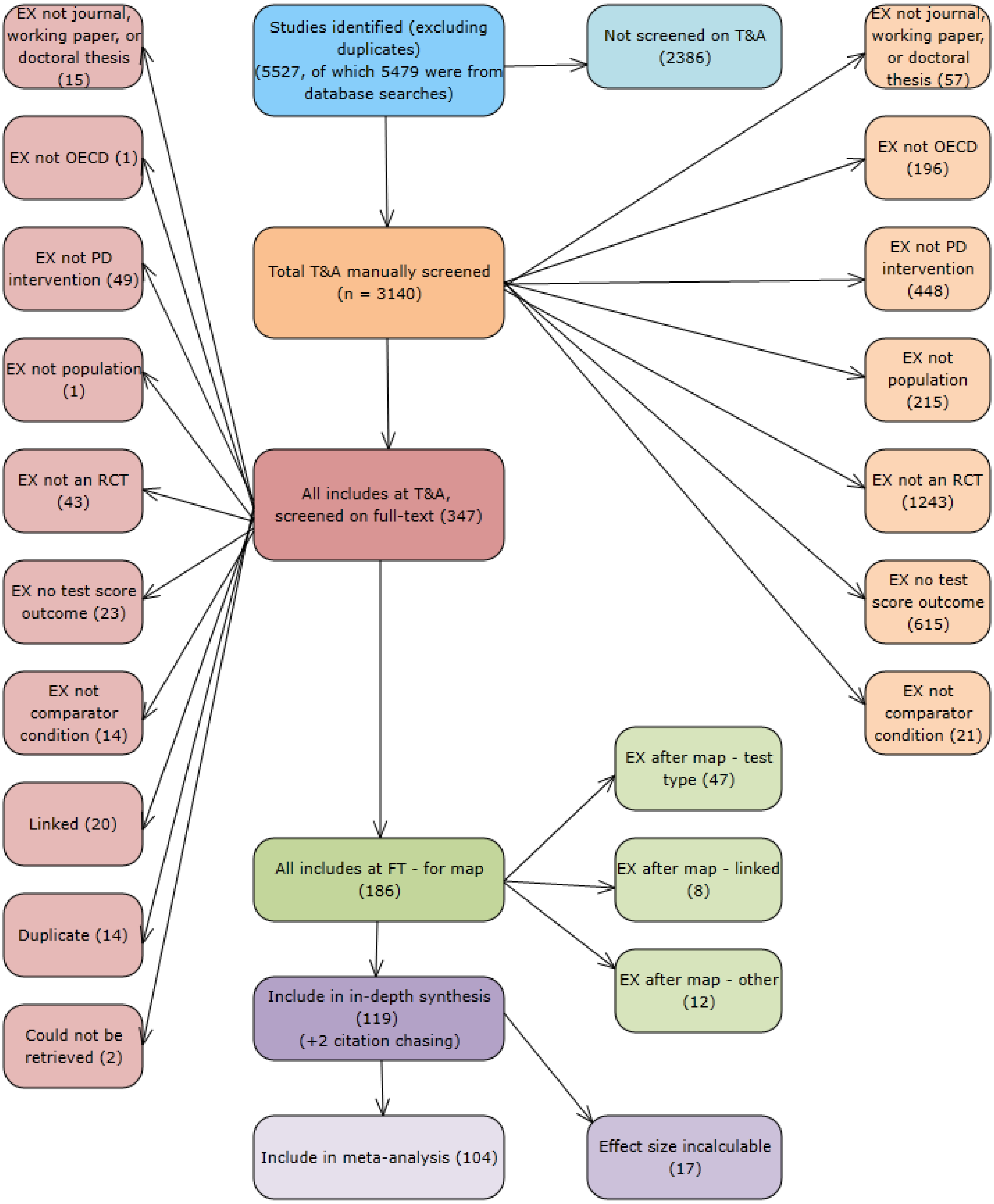
S4 S3 Limiters - Education Level: Adult Basic Education, Adult Education, Higher Education

S3 S1 OR S2

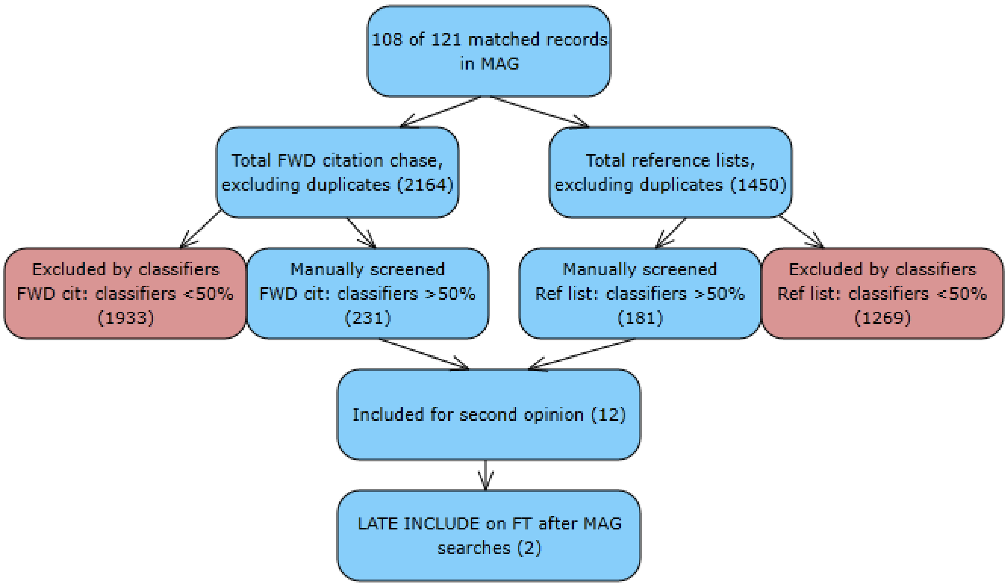
S2 TI (teacher OR teachers) OR AB (teacher OR teachers) OR TI (educators OR educator) OR AB (educator OR educators)

S1 DE "Teachers" OR DE "Secondary School Teachers" OR DE "Science Teachers" OR DE "Preschool Teachers" OR DE "Experienced Teachers" OR DE "Middle School Teachers" OR DE "Elementary School Teachers"

Appendix 3: PRISMA flow diagram


EX not journal, working paper, or doctoral thesis (15)
Studies identified (excluding duplicates) (5527, of which 5479 were from database searches)
Not screened on T&A (2386)
EX not journal, working paper, or doctoral thesis (57)
EX not OECD (1)
EX not OECD (196)
EX not PD intervention (49)
Total T&A manually screened (n = 3140)
EX not PD intervention (448)
EX not population (1)
EX not population (215)
EX not an RCT (43)
All includes at T&A, screened on full-text (347)
EX not an RCT (1243)
EX no test score outcome (23)
EX no test score outcome (615)
EX not comparator condition (14)
EX not comparator condition (21)
EX after map - test type (47)
Linked (20)
All includes at FT - for map (186)
EX after map - linked (8)
Duplicate (14)
EX after map - other (12)
Include in in-depth synthesis (119) (+2 citation chasing)
Could not be retrieved (2)
Include in meta-analysis (104)
Effect size incalculable (17)

Flow of literature diagram for the supplementary search (citation chasing of full-text documents)


108 of 121 matched records in MAG
Total FWD citation chase, excluding duplicates (2164)
Total reference lists, excluding duplicates (1450)
Excluded by classifiers FWD cit: classifiers <50% (1933)
Manually screened FWD cit: classifiers >50% (231)
Manually screened Ref list: classifiers >50% (181)
Excluded by classifiers Ref list: classifiers <50% (1269)
Included for second opinion (12)
LATE INCLUDE on FT after MAG searches (2)

Notes. MAG = Microsoft Academic Graph, FWD = forward, FWD cit = forward citation chasing, Ref list = reference list checking, FT = full-text.

Appendix 4: Characteristics of included studies table

For the 121 reports included in the synthesis

Short Title	Population	Intervention	Mechanisms	Research methods	Outcomes
Abe (2012)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Not clear for some of the sample* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note *None of the above specifically* **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *295*	**Instil Insight** Revisit material **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling **Embed Practice** Context-specific repetition	**Randomisation** Number of units randomised *45* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Allen (2011)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth	**Broad area of focus of the PD** Unclear or other - add note *Not one of these in*	**Motivate Goals** Praise/reinforce **Teach Techniques** Instruction	**Randomisation** Number of units randomised *78*	**Outcome types** Test score - Other (specify) Practice change outcome

	Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Not reported* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	*particular* **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Portraying curriculum Containing behaviour Enlisting participation **Dosage (No. of intended hours per participating teacher)** *20*	Practical soc. support Feedback on practice Modelling	**Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Include in meta-analysis?** Yes
Allen (2015)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *They report poverty levels but not free lunch levels* **Early-career / newly-qualified teachers?**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Containing behaviour Enlisting participation **Dosage (No. of intended hours per participating**	**Instil Insight** Manage cognitive load Revisit material **Motivate Goals** Goal setting Praise/reinforce **Teach Techniques** Instruction Feedback on practice Modelling **Embed Practice** Self-monitoring practice	**Randomisation** Number of units randomised *86* **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Other (specify) **Include in meta-analysis?** Yes

	No (mixed or experienced), OR unclear	**teacher)** *20*			
Ansari (2018)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Not reported* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum Containing behaviour Enlisting participation **Dosage (No. of intended hours per participating teacher)**	**Motivate Goals** Praise/reinforce **Teach Techniques** Instruction Feedback on practice Modelling **Embed Practice** Self-monitoring practice	**Randomisation** Number of units randomised *401* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy Test score - Vocabulary **Include in meta-analysis?** Yes
Arens (2012)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Unclear*	**Broad area of focus of the PD** Formative assessment / responsive teaching *A few mentions of assessment e.g. "Such changes are related to improved teacher responsiveness and the use of diverse assessment practices," RISE =*	**Motivate Goals** Goal setting Credible source **Teach Techniques** Instruction Modelling **Embed Practice** Self-monitoring practice	**Randomisation** Number of units randomised *53* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	Early-career / newly-qualified teachers? No (mixed or experienced), OR unclear	*responsive instruction for success in English.* **Subject targeted** Other subjects **Kennedy's 'persistent problems'** Portraying curriculum Exposing student thinking **Dosage (No. of intended hours per participating teacher)** *unclear*		**Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	
Argentin (2014)	**Country** Italy Not UK or US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *67*	**Motivate Goals** Goal setting **Teach Techniques** Practical soc. support	**Randomisation** Number of units randomised *175* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

				Analysis pre-specified? Not reported	
Arteaga (2019)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** % eligible for free or reduced-price lunch *>75%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** English/native language/speech/literacy Generic (cross-curricular) **Kennedy's 'persistent problems'** Portraying curriculum Containing behaviour Exposing student thinking **Dosage (No. of intended hours per participating teacher)**	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling **Embed Practice** Action planning	**Randomisation** Number of units randomised *24* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy Test score - Vocabulary **Include in meta-analysis?** Yes
August (2014)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Science **Kennedy's 'persistent**	**Motivate Goals** Praise/reinforce **Teach Techniques** Instruction Practical soc. support Feedback on practice Rehearsal	**Randomisation** Number of units randomised *60 sections (classes) for the 15 teachers* **Cluster level attrition (if clustered)** Acceptable attrition	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)**		**Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	
Ault (2017)	**Country** US **Age Group/s** Primary/elementary Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *p.21 not clear what the total is. Does say earlier 'about two thirds' but figure is not definite* **Early-career / newly-qualified teachers?** Yes	**Broad area of focus of the PD** Unclear or other - add note *General - mentors used CLASS observation framework* **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Portraying curriculum Containing behaviour **Dosage (No. of intended hours per participating**	**Teach Techniques** Practical soc. support Feedback on practice Modelling	**Randomisation** Number of units randomised *556* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Include in meta-analysis?** No

		teacher) *p.8 'at least 3.5hrs'*			
Babinski (2018)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *83%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note *None specifically* **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)**	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support **Embed Practice** Action planning	**Randomisation** Number of units randomised *15* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Biggart (2015)	**Country** UK **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *Original sample: 40.3% Analysed*	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'**	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *25* **Cluster level attrition (if clustered)** Not acceptable attrition **Pupil level attrition** Not acceptable attrition	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	sample: 39.0% **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	Portraying curriculum Exposing student thinking **Dosage (No. of intended hours per participating teacher)** *2 days (2 x 8 = 16hrs)*		**Test type** Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	
Boardman (2015)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *Calculated for the strongest intervention (full CSR) and control only: 717/813 = 88.2%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** Science Other subjects **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)** *Unclear how long booster sessions were*	**Instil Insight** Revisit material **Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling **Embed Practice** Action planning Context-specific repetition	**Randomisation** Number of units randomised *Very unclear* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Bos (2012)	**Country** US	**Broad area of focus of the PD** Unclear or other - add	**Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised	**Outcome types** Test score - English/native

	Age Group/s Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Only reported at school level not sample level (p.33)* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**	Feedback on practice **Embed Practice** Action planning	*52* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	language/speech/literacy **Include in meta-analysis?** Yes
Brendefur (2013)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note *none of these specifically* **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating**	**Instil Insight** Revisit material **Teach Techniques** Instruction Practical soc. support Rehearsal	**Randomisation** Number of units randomised *6* Notes - for discussion *Randomising 6 schools is useless. We need a way to exclude these studies.* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

		teacher) *8*		**Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	
Buysse (2010)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Unreported* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note *Not specifically related to any of these* **Subject targeted** Other subjects **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *Unclear how long the reflective supervision was in hours*	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice **Embed Practice** Action planning Self-monitoring practice Context-specific repetition	**Randomisation** Number of units randomised *55* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy Test score - Vocabulary Practice change outcome **Include in meta-analysis?** Yes
Campbell (2011)	**Country** US **Age Group/s** Primary/elementary	**Broad area of focus of the PD** Unclear or other - add note	**Teach Techniques** Practical soc. support Modelling	**Randomisation** Number of units randomised *36*	**Outcome types** Test score - Maths **Include in meta-**

	Overall pupil disadvantage Pupil disadvantage unclear or not reported *It has the coefficient but not how many?* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum Exposing student thinking **Dosage (No. of intended hours per participating teacher)** *Quote various hours but I'm not sure!*		**Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**analysis?** Yes
Castro (2017)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy Maths **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**	**Motivate Goals** Credible source **Teach Techniques** Instruction Practical soc. support Feedback on practice **Embed Practice** Context-specific repetition	**Randomisation** Number of units randomised *33* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy Test score - Maths Test score - Vocabulary

Cavalluzzo (2014)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *At student level % not reported* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Data driven instruction **Subject targeted** Maths **Kennedy's 'persistent problems'** Unclear or other - add note **Dosage (No. of intended hours per participating teacher)** *9 days for teachers (9x8 = 72hrs)*	**Instil Insight** Revisit material **Teach Techniques** Instruction Practical soc. support Modelling Rehearsal	**Randomisation** Number of units randomised *60 schools put into 4 blocks and half randomly assigned to treatment, half to control* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes
Chuang (2020)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *61.0%*	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths **Kennedy's 'persistent problems'** Containing behaviour	**Teach Techniques** Modelling	**Randomisation** Number of units randomised *105* **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not pre-registered	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

	Early-career / newly-qualified teachers? No (mixed or experienced), OR unclear	**Dosage (No. of intended hours per participating teacher)** *48*			
Clements (2013)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Reported at school level* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *32hrs of delivery but unclear how long the mentors spent with teachers*	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *42* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Maths Practice change outcome **Include in meta-analysis?** Yes
Connor (2007)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage**	**Broad area of focus of the PD** Data driven instruction *using the recommendation of A2i to plan instruction*	**Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *47* **Cluster level attrition (if clustered)**	**Outcome types** Test score - English/native language/speech/literacy Test score - Maths

	Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Subject targeted** English/native language/speech/literacy **Kennedy’s ‘persistent problems’** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**		Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Include in meta-analysis?** Yes
Cordray (2012)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** % eligible for free or reduced-price lunch *99.4%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Data driven instruction **Subject targeted** English/native language/speech/literacy **Kennedy’s ‘persistent problems’** Unclear or other - add note **Dosage (No. of intended hours per participating teacher)** *4 x 8 = 32hrs*	**Instil Insight** Revisit material **Teach Techniques** Instruction	**Randomisation** Number of units randomised *32* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Correnti (2020)	**Country** US	**Broad area of focus of the PD**	**Instil Insight** Manage cognitive load	**Randomisation** Number of units	**Outcome types** Test score -

	Age Group/s Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Says 'economically disadvantaged' but not lunch* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**	Revisit material **Motivate Goals** Goal setting Credible source **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Action planning Self-monitoring practice Context-specific repetition	randomised *31* **Cluster level attrition (if clustered)** Not acceptable attrition **Pupil level attrition** Not acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	English/native language/speech/literacy **Include in meta-analysis?** Yes
Cromley (2016)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?**	**Broad area of focus of the PD** Science of learning / cognitive science **Subject targeted** Science **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating**	**Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *193 (96 randomised for one curriculum and 97 for the other)* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition	**Outcome types** Test score - Maths Test score - Science **Include in meta-analysis?** No

	No (mixed or experienced), OR unclear	**teacher)** *Unclear how long in total*		**Test type** High stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	
Culliney 2019	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *50*	**Motivate Goals** Goal setting **Teach Techniques** Instruction **Embed Practice** Action planning Context-specific repetition	**Randomisation** Number of units randomised *91* **Test type** Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
DeCesare (2017)	**Country** US **Age Group/s** Primary/elementary	**Broad area of focus of the PD** Unclear or other - add note *Mentors can provide support of their discretion*	**Teach Techniques** Practical soc. support	**Randomisation** Number of units randomised *77* **Cluster level attrition (if**	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

	Overall pupil disadvantage Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** Yes	**Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Unclear or other - add note **Dosage (No. of intended hours per participating teacher)** *Unclear as per our standards*		**clustered)** Not acceptable attrition **Pupil level attrition** Not acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	
Dix (2018)	**Country** Australia Not UK or US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation Exposing student thinking **Dosage (No. of intended hours per participating teacher)**	**Teach Techniques** Instruction Practical soc. support Modelling **Embed Practice** Context-specific repetition	**Randomisation** Number of units randomised *167* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?**	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

		Unclear how long the professional reading, journaling etc was		Pre-registration / pre-specification / analysis plan	
Dolfin (2019)	**Country** US **Age Group/s** Primary/elementary Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** Yes	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *would require some estimation*	**Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling **Embed Practice** Context-specific repetition	**Randomisation** Number of units randomised *70* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Engelstad (2020)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Not reported in terms of free lunch*	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent**	**Instil Insight** Manage cognitive load Revisit material **Motivate Goals** Goal setting Praise/reinforce **Teach Techniques** Instruction	**Randomisation** Number of units randomised *6* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition**	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	(referred to as family disadvantage) **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**problems'** Portraying curriculum Containing behaviour Enlisting participation **Dosage (No. of intended hours per participating teacher)** *42hrs*	Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Action planning Context-specific repetition	Not acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	
Finkelstein (2011)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *No reference to "lunch"* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Other subjects **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *No total reported*	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Modelling Rehearsal	**Randomisation** Number of units randomised *128* **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not pre-registered	**Outcome types** Test score - Other (specify) **Include in meta-analysis?** Yes
Foliano (2019)	**Country** UK **Age Group/s** Primary/elementary	**Broad area of focus of the PD** Unclear or other - add note *Growth mindset*	**Motivate Goals** Credible source **Teach Techniques** Instruction	**Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

	Overall pupil disadvantage % eligible for free or reduced-price lunch *35.7* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Unclear or other - add note **Dosage (No. of intended hours per participating teacher)**		**Test type** High stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	
Garet (2008)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *78* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *108 (48+60)*	**Instil Insight** Manage cognitive load Revisit material **Motivate Goals** Credible source **Teach Techniques** Instruction Feedback on practice Rehearsal	**Randomisation** Number of units randomised *90* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Garet (2016)	**Country** US	**Broad area of focus of the PD**	**Instil Insight** Revisit material	**Randomisation** Number of units	**Outcome types** Test score - Maths

	Age Group/s Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Not reported for all in the sample* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	Unclear or other - add note *Doesn't fall easily into a category* **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *93*	**Motivate Goals** Praise/reinforce **Teach Techniques** Instruction Practical soc. support Feedback on practice Rehearsal	randomised *221* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	Practice change outcome **Include in meta-analysis?** Yes
Gerde (2014)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note *Doesn't fit into a category* **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended**	**Teach Techniques** Instruction Modelling	**Randomisation** Number of units randomised *34* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

		hours per participating teacher) *10*		**Analysis pre-specified?** Not reported	
Gersten (2010)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**	**Instil Insight** Manage cognitive load Revisit material **Teach Techniques** Instruction Practical soc. support Modelling **Embed Practice** Action planning	**Randomisation** Number of units randomised *84* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Practice change outcome **Include in meta-analysis?** Yes
Glazerman (2010)	**Country** US **Age Group/s** Primary/elementary	**Subject targeted** English/native language/speech/literacy Maths	**Teach Techniques** Practical soc. support Feedback on practice Modelling	**Randomisation** Number of units randomised *200* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition**	**Include in meta-analysis?** No

				Attrition unclear **Test type** High stakes, standardised exam	
Goodson (2010)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** % eligible for free or reduced-price lunch *85.2%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** Unclear	**Instil Insight** Manage cognitive load Revisit material **Motivate Goals** Goal setting Praise/reinforce **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal	**Randomisation** Number of units randomised *70* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Vocabulary **Include in meta-analysis?** Yes
Gorard (2015)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Generic (cross-curricular)	**Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling	**Randomisation** Number of units randomised *48* **Cluster level attrition (if clustered)** Acceptable attrition	**Outcome types** Test score - English/native language/speech/literacy Test score - Maths **Include in meta-**

	lunch *46.8%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Kennedy's 'persistent problems'** Enlisting participation **Dosage (No. of intended hours per participating teacher)** *Unclear how much support schools would have had*		**Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**analysis?** Yes
Greenleaf (2011)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Total is unclear* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** English/native language/speech/literacy Science **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)** *80*	**Instil Insight** Revisit material **Teach Techniques** Instruction Practical soc. support Modelling Rehearsal **Embed Practice** Action planning Self-monitoring practice Context-specific repetition	**Randomisation** Number of units randomised *83* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Science **Include in meta-analysis?** Yes

Hamm (2010)	**Country** US **Age Group/s** Primary/elementary Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *53.8%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Containing behaviour Enlisting participation **Dosage (No. of intended hours per participating teacher)** *27*	**Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *4* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Include in meta-analysis?** No
Hanley (2016)	**Country** UK **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *24.5%* **Early-career / newly-qualified**	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Science **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended**	**Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling	**Randomisation** Number of units randomised *53* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Not acceptable attrition **Test type**	**Outcome types** Test score - Science **Include in meta-analysis?** Yes

	teachers? No (mixed or experienced), OR unclear	**hours per participating teacher)**		Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	
Haring (2016)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *93.5* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *Unclear how much coaching*	**Motivate Goals** Goal setting Praise/reinforce **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Action planning Context-specific repetition	**Randomisation** Number of units randomised *14* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Hitchcock (2011)	**Country** US **Age Group/s** Primary/elementary	**Broad area of focus of the PD** Unclear or other - add note	**Motivate Goals** Praise/reinforce **Teach Techniques**	**Randomisation** Number of units randomised *86*	**Outcome types** Test score - English/native language/speech/literacy

	Overall pupil disadvantage % eligible for free or reduced-price lunch *73.1%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *10*	Instruction Feedback on practice	**Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not pre-registered	**Include in meta-analysis?** Yes
Humphrey (2018)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *98.4*	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Containing behaviour **Dosage (No. of intended hours per participating teacher)** *26*	**Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling	**Randomisation** Number of units randomised *77* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?**	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

				Pre-registration / pre-specification / analysis plan	
IEE (2016)	**Country** UK **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Portraying curriculum Containing behaviour Exposing student thinking **Dosage (No. of intended hours per participating teacher)**	**Teach Techniques** Instruction Feedback on practice	**Randomisation** Number of units randomised *45* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Jaciw (2016a)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *60%*	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Maths **Kennedy's 'persistent problems'**	**Teach Techniques** Instruction Practical soc. support Modelling **Embed Practice** Action planning	**Randomisation** Number of units randomised *22* **Cluster level attrition (if clustered)** Not acceptable attrition **Test type**	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

	Early-career / newly-qualified teachers? No (mixed or experienced), OR unclear	Portraying curriculum **Dosage (No. of intended hours per participating teacher)**		Low stakes, standardised exam **Analysis pre-specified?** Not reported	
Jaciw (2016b)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *39.4* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** English/native language/speech/literacy Science **Kennedy's 'persistent problems'** Enlisting participation **Dosage (No. of intended hours per participating teacher)** *Unclear how long the ongoing support is*	**Teach Techniques** Instruction Practical soc. support Modelling	**Randomisation** Number of units randomised *82* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Jacob (2017)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native	**Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal	**Randomisation** Number of units randomised *63* **Cluster level attrition (if clustered)**	**Outcome types** Test score - English/native language/speech/literacy Test score - Maths

	Pupil disadvantage unclear or not reported *At school level only* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *35*		Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	**Include in meta-analysis?** Yes
James-Burdumy (2010)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Unclear for just cohort 1 readabout + control* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *24*	**Instil Insight** Revisit material **Teach Techniques** Instruction	**Randomisation** Number of units randomised *61* **Cluster level attrition (if clustered)** Not acceptable attrition **Pupil level attrition** Not acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Include in meta-analysis?** No

Jay (2017)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Only reported separately by group* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)**	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Modelling **Embed Practice** Action planning Self-monitoring practice Context-specific repetition	**Randomisation** Number of units randomised *76* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Jayanthi (2017)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Only at school level* **Early-career / newly-qualified**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum	**Teach Techniques** Practical soc. support Modelling **Embed Practice** Self-monitoring practice	**Randomisation** Number of units randomised *84* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

	teachers? No (mixed or experienced), OR unclear	**Dosage (No. of intended hours per participating teacher)** *24*		**Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	
Jayanthi (2018)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Only at school level not for the sample* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *12.5*	**Motivate Goals** Credible source **Teach Techniques** Instruction Practical soc. support **Embed Practice** Action planning Context-specific repetition	**Randomisation** Number of units randomised *62* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Vocabulary **Include in meta-analysis?** Yes
Jerrim (2016)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths	**Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *127 = 83 (primary) + 44 (secondary)* **Cluster level attrition (if**	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

	Pupil disadvantage unclear or not reported *Unclear combined %* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**		**clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	
Johanson (2016)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)**	**Instil Insight** Manage cognitive load Revisit material **Motivate Goals** Praise/reinforce **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Action planning Self-monitoring practice Context-specific repetition	**Randomisation** Number of units randomised *49* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Vocabulary **Include in meta-analysis?** Yes

Kinzie (2014)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Maths Science **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**	**Instil Insight** Manage cognitive load **Teach Techniques** Instruction Practical soc. support Modelling	**Randomisation** Number of units randomised *24* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes
Kitmitto (2018)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *26.1%*	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Science **Kennedy's 'persistent problems'** Portraying curriculum	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Modelling **Embed Practice** Prompts/cues Self-monitoring practice	**Randomisation** Number of units randomised *205* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

	Early-career / newly-qualified teachers? No (mixed or experienced), OR unclear	Enlisting participation **Dosage (No. of intended hours per participating teacher)** *32*		**Test type** Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	
Kraft (2020)	**Country** US **Age Group/s** Primary/elementary Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)**	**Instil Insight** Revisit material **Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling **Embed Practice** Action planning Self-monitoring practice Context-specific repetition	**Randomisation** Number of units randomised *142* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** High stakes, standardised exam Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes
Kushman (2011)	**Country** US	**Broad area of focus of the PD**	**Instil Insight** Revisit material	**Randomisation** Number of units	**Outcome types** Test score - Maths

	Age Group/s Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	Science of learning / cognitive science *Later mentions of explicit instruction* **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)**	**Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling **Embed Practice** Action planning	randomised *52* **Cluster level attrition (if clustered)** Not acceptable attrition **Pupil level attrition** Not acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Include in meta-analysis?** Yes
Landry (2009)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note *Some responsive teaching but not the broad area* **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum	**Instil Insight** Manage cognitive load Revisit material **Motivate Goals** Praise/reinforce **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal	**Randomisation** Number of units randomised *158* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised	**Include in meta-analysis?** No

		Containing behaviour Exposing student thinking **Dosage (No. of intended hours per participating teacher)**	**Embed Practice** Action planning Self-monitoring practice Context-specific repetition	exam **Analysis pre-specified?** Not reported	
Landry (2014)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Unable to work it out for just the intensive intervention vs. control* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Portraying curriculum Containing behaviour Enlisting participation Exposing student thinking **Dosage (No. of intended hours per participating teacher)** *66 coaching hrs + 34.5 pd = 100.5hrs*	**Motivate Goals** Goal setting Praise/reinforce **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Self-monitoring practice Context-specific repetition	**Randomisation** Number of units randomised *65* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy Test score - Maths Test score - Vocabulary **Include in meta-analysis?** Yes
Landry (2017)	**Country** US **Age Group/s** Early years / pre-kindergarten	**Broad area of focus of the PD** Unclear or other - add note	**Motivate Goals** Praise/reinforce **Teach Techniques** Instruction	**Randomisation** Number of units randomised *77*	**Include in meta-analysis?** No

	Overall pupil disadvantage % eligible for free or reduced-price lunch *97.3%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**	Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Self-monitoring practice Context-specific repetition	**Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	
Lewis (2015)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)** *56*	**Instil Insight** Revisit material **Teach Techniques** Instruction Practical soc. support Modelling Rehearsal	**Randomisation** Number of units randomised *16* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

Llosa (2016)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *68.0%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Science **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)** *40*	**Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *33* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	**Outcome types** Test score - Science **Include in meta-analysis?** Yes
Lonigan (2011)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified**	**Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum Containing behaviour **Dosage (No. of intended**	**Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal	**Randomisation** Number of units randomised *33* Notes - for discussion *48 but across two conditions, with 18 in control and 15 in most intensive condition* **Cluster level attrition (if**	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	teachers? No (mixed or experienced), OR unclear	**hours per participating teacher)**		**clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam	
Lord (2017)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Can't seem to work this out for arm 7 + control* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *Can add up some of it but then ongoing engagement is 0.5 hr per month - unclear how many months*	**Instil Insight** Revisit material **Motivate Goals** Credible source **Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *823* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Ludwig (2015)	**Country** US	**Broad area of focus of the PD**	**Teach Techniques** Practical soc. support	**Randomisation** Number of units	**Outcome types** Test score - Maths

	Age Group/s Early years / pre-kindergarten **Overall pupil disadvantage** % eligible for free or reduced-price lunch *44.0%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	Unclear or other - add note **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**		randomised *22* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Include in meta-analysis?** No
Maerten-Rivera (2016)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *79.7* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Science **Kennedy's 'persistent problems'** Portraying curriculum Exposing student thinking **Dosage (No. of intended hours per participating**	**Teach Techniques** Instruction Practical soc. support Modelling	**Randomisation** Number of units randomised *64* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam	**Include in meta-analysis?** No

		teacher) *Depended upon if you were a new or returning teacher*		**Analysis pre-specified?** Not reported	
Martin (2012)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Only average school-wide %* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation Exposing student thinking **Dosage (No. of intended hours per participating teacher)** *40*	**Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *70* **Cluster level attrition (if clustered)** Not acceptable attrition **Pupil level attrition** Not acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes
Matsumura (2013)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch	**Broad area of focus of the PD** Unclear or other - add note *Discussion approach to reading comprehension - doesn't fit into any of these categories well*	**Teach Techniques** Instruction Practical soc. support Modelling Rehearsal **Embed Practice** Action planning Context-specific repetition	**Randomisation** Number of units randomised *32* **Cluster level attrition (if clustered)** Acceptable attrition	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	91% **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum Exposing student thinking **Dosage (No. of intended hours per participating teacher)**		**Pupil level attrition** Attrition unclear **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	
Mattera (2018)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *coaching dosage unclear*	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice **Embed Practice** Action planning	**Randomisation** Number of units randomised *47 (34 continue BAU; 12 schools implement Making Pre-K count across Pre-K and K; (excludes those which did Making Pre-K count only at Pre-K and those with added tutoring - see page 11).* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition	**Outcome types** Test score - Maths Practice change outcome **Include in meta-analysis?** Yes

				Test type Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	
Mazzie (2009)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *73.5* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note *Developing assessments - not really about the instruction afterwards* **Subject targeted** Science **Kennedy's 'persistent problems'** Unclear or other - add note **Dosage (No. of intended hours per participating teacher)** *unclear specific hrs of homework*	**Instil Insight** Revisit material **Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice Rehearsal	**Randomisation** Number of units randomised *33* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Include in meta-analysis?** No
McMaster (2019)	**Country** US **Age Group/s**	**Broad area of focus of the PD** Data driven instruction	**Instil Insight** Revisit material **Motivate Goals**	**Randomisation** Number of units randomised *19 classes (20 teachers but*	**Outcome types** Test score - English/native language/speech/literac

	Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *69.8* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Subject targeted** English/native language/speech/literacy **Kennedy’s ‘persistent problems’** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**	Goal setting Credible source Praise/reinforce **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Prompts/cues Action planning Self-monitoring practice Context-specific repetition	*1 classroom had 2 teachers and was treated as '1' for the purposes of assignment)* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	y **Include in meta-analysis?** Yes
McNally (2014)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?**	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** Generic (cross-curricular) **Kennedy’s ‘persistent problems’** Exposing student thinking	**Instil Insight** Revisit material **Motivate Goals** Credible source	**Randomisation** Number of units randomised *36* **Cluster level attrition (if clustered)** Not acceptable attrition **Test type** Low stakes, standardised exam	**Outcome types** Test score - Other (specify) **Include in meta-analysis?** Yes

	No (mixed or experienced), OR unclear	**Dosage (No. of intended hours per participating teacher)**			
Meyers (2016)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *82.2* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** English/native language/speech/literacy Maths **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *This is for eMINTs but unclear how much for eMINTs + intel (more intense PD)*	**Instil Insight** Revisit material **Teach Techniques** Instruction Practical soc. support Modelling **Embed Practice** Context-specific repetition	**Randomisation** Number of units randomised *60* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy Test score - Maths **Include in meta-analysis?** Yes
Murphy (2017)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native	**Instil Insight** Revisit material **Teach Techniques** Instruction Practical soc. support Feedback on practice	**Randomisation** Number of units randomised *181* **Cluster level attrition (if clustered)**	**Outcome types** Test score - Other (specify) **Include in meta-analysis?** Yes

	% eligible for free or reduced-price lunch *51.5* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	language/speech/literacy Maths **Kennedy’s ‘persistent problems’** Enlisting participation **Dosage (No. of intended hours per participating teacher)** *40*	**Embed Practice** Action planning Self-monitoring practice	Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	
Neuman (2015)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** % eligible for free or reduced-price lunch *66.4* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy’s ‘persistent problems’** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *8*	**Instil Insight** Manage cognitive load **Teach Techniques** Instruction **Embed Practice** Prompts/cues	**Randomisation** Number of units randomised *10* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Vocabulary **Include in meta-analysis?** Yes

Newman (2012)	**Country** US **Age Group/s** Primary/elementary Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *56* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** Maths Science **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**	**Teach Techniques** Instruction Practical soc. support Modelling Rehearsal	**Randomisation** Number of units randomised *82 schools across 2 sub experiments (40 schools and then 42 schools) were paired and then randomly allocated* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes
O'Hare (2019)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy	**Teach Techniques** Instruction Practical soc. support Modelling Rehearsal	**Randomisation** Number of units randomised *98* **Cluster level attrition (if clustered)** Acceptable attrition	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	50.3 **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *Unclear how long follow up sessions were*		**Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	
Olson (2012)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Would require calculation* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)** *46*	**Teach Techniques** Instruction Practical soc. support Feedback on practice	**Randomisation** Number of units randomised *72* **Cluster level attrition (if clustered)** Not acceptable attrition **Pupil level attrition** Not acceptable attrition **Test type** High stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

Olson (2020)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Enlisting participation Exposing student thinking **Dosage (No. of intended hours per participating teacher)** *46*	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling **Embed Practice** Self-monitoring practice Context-specific repetition	**Randomisation** Number of units randomised *230* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Parkinson (2015)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Not reported at pupil sample level*	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'**	**Instil Insight** Manage cognitive load **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling **Embed Practice**	**Randomisation** Number of units randomised *78* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** No

	Early-career / newly-qualified teachers? No (mixed or experienced), OR unclear	Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *Most teachers: 186 but more for some individuals designated as model teachers (not analysed separately)*	Action planning Context-specific repetition	**Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	
Pianta (2017)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum Containing behaviour Enlisting participation **Dosage (No. of intended hours per participating teacher)** *dosage of coaching element unclear*	**Motivate Goals** Praise/reinforce **Teach Techniques** Instruction Feedback on practice **Embed Practice** Self-monitoring practice Context-specific repetition	**Randomisation** Number of units randomised *91* Notes - for discussion *Most intensive (course + coaching) vs pure control (no course, no coaching, not supplemental). Page 5* **Cluster level attrition (if clustered)** Not acceptable attrition **Pupil level attrition** Not acceptable attrition **Test type** Low stakes, standardised exam Test developed or adapted	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

				by researchers or teachers **Analysis pre-specified?** Not reported	
Piasta (2015)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths Science **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *64*	**Instil Insight** Revisit material **Teach Techniques** Instruction Modelling	**Randomisation** Number of units randomised *65* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes
Piasta (2020)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native	**Instil Insight** Revisit material **Motivate Goals** Goal setting **Teach Techniques**	**Randomisation** Number of units randomised *Unclear* **Cluster level attrition (if clustered)**	**Include in meta-analysis?** No

	Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *Not clear how many hrs from the coaching element*	Instruction Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Action planning Context-specific repetition	Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	
Portes (2016)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *100*	**Teach Techniques** Practical soc. support Feedback on practice **Embed Practice** Context-specific repetition	**Randomisation** Number of units randomised *74* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

Powell (2010)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)** *Unclear exactly how many hrs of coaching*	**Motivate Goals** Praise/reinforce **Teach Techniques** Instruction Feedback on practice Modelling **Embed Practice** Context-specific repetition	**Randomisation** Number of units randomised *88* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Prast (2018)	**Country** Netherlands **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?**	**Broad area of focus of the PD** Unclear or other - add note *Differentiation* **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Context-specific repetition	**Randomisation** Number of units randomised *32* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear	**Outcome types** Test score - Maths

	No (mixed or experienced), OR unclear	**Dosage (No. of intended hours per participating teacher)** *30*		**Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	
Presser (2012)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *28*	**Instil Insight** Revisit material **Teach Techniques** Instruction Modelling Rehearsal	**Randomisation** Number of units randomised *16* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes
Randel (2011)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage**	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** Maths	**Teach Techniques** Instruction Practical soc. support Rehearsal **Embed Practice** Context-specific repetition	**Randomisation** Number of units randomised *67* **Cluster level attrition (if clustered)**	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

	Pupil disadvantage unclear or not reported *Only at school level* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Kennedy's 'persistent problems'** Enlisting participation Exposing student thinking **Dosage (No. of intended hours per participating teacher)**		Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	
Reinke (2018)	**Country** US **Age Group/s** Early years / pre-kindergarten Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *61* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy Maths **Kennedy's 'persistent problems'** Containing behaviour **Dosage (No. of intended hours per participating teacher)** *Coaching dosage unclear/varied*	**Instil Insight** Revisit material **Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal	**Randomisation** Number of units randomised *105* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

Rienzo (2015)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *21.1%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Enlisting participation **Dosage (No. of intended hours per participating teacher)** *8*	**Instil Insight** Manage cognitive load **Motivate Goals** Credible source **Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *30* **Cluster level attrition (if clustered)** Not acceptable attrition **Pupil level attrition** Not acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes
Robinson-Smith (2018)	**Country** UK **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** % eligible for free or reduced-price lunch *pre-school* **Early-career / newly-qualified teachers?**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths **Kennedy's 'persistent problems'** Unclear or other - add note	**Teach Techniques** Instruction Practical soc. support Feedback on practice	**Randomisation** Number of units randomised *108* **Cluster level attrition (if clustered)** Not acceptable attrition **Pupil level attrition** Not acceptable attrition **Test type**	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

	No (mixed or experienced), OR unclear	**Dosage (No. of intended hours per participating teacher)**		Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	
Savage (2013)	**Country** Canada Not UK or US **Age Group/s** Early years / pre-kindergarten Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *8 hrs*	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice Rehearsal **Embed Practice** Action planning	**Randomisation** Number of units randomised *74* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Simmons (2010)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native	**Instil Insight** Manage cognitive load **Teach Techniques** Instruction Practical soc. support Rehearsal	**Randomisation** Number of units randomised *48* **Cluster level attrition (if clustered)**	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-**

	Pupil disadvantage unclear or not reported *Unclear at sample level* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	language/speech/literacy Other subjects **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *18*		Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	**analysis?** Yes
Sloan (2018)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *15.2* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Containing behaviour **Dosage (No. of intended hours per participating teacher)**	**Teach Techniques** Instruction Practical soc. support Feedback on practice **Embed Practice** Action planning	**Randomisation** Number of units randomised *83* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?**	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

				Pre-registration / pre-specification / analysis plan	
Snow (2014)	**Country** Australia Not UK or US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *Continuing contact unclear*	**Instil Insight** Revisit material **Teach Techniques** Instruction Modelling **Embed Practice** Action planning	**Randomisation** Number of units randomised *14* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Snyder (2018)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** % eligible for free or reduced-price lunch *39%*	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'**	**Motivate Goals** Goal setting **Teach Techniques** Instruction Feedback on practice Modelling Rehearsal **Embed Practice**	**Randomisation** Number of units randomised *36* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	Early-career / newly-qualified teachers? No (mixed or experienced), OR unclear	Portraying curriculum Containing behaviour **Dosage (No. of intended hours per participating teacher)** *What is 'alternate' is not clear.*	Action planning Self-monitoring practice	**Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	
Stone (2005)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Reported as mean % at school level not pupil level* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note *Whilst it has check for understanding element as assess what they know, it isn't clear that they are responding to this* **Subject targeted** Maths Other subjects **Kennedy's 'persistent problems'** Portraying curriculum Exposing student thinking **Dosage (No. of intended hours per participating teacher)**	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Modelling **Embed Practice** Action planning Self-monitoring practice	**Randomisation** Number of units randomised *274* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Include in meta-analysis?** No

Stone (2008)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Modelling Rehearsal **Embed Practice** Action planning Context-specific repetition	**Randomisation** Number of units randomised *131* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes
Styles (2014)	**Country** UK **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *28.7%* **Early-career / newly-qualified**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum	**Instil Insight** Revisit material **Teach Techniques** Instruction Practical soc. support Modelling Rehearsal **Embed Practice** Action planning	**Randomisation** Number of units randomised *649* **Cluster level attrition (if clustered)** Not cluster randomised **Pupil level attrition** Acceptable attrition **Test type**	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	teachers? No (mixed or experienced), OR unclear	**Dosage (No. of intended hours per participating teacher)**		Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	
Supovitz (2018)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *69* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Formative assessment / responsive teaching Data driven instruction **Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum Exposing student thinking **Dosage (No. of intended hours per participating teacher)**	**Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *61* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Not acceptable attrition **Test type** High stakes, standardised exam Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Include in meta-analysis?** No
Supovitz (2018)	**Country** US **Age Group/s** Primary/elementary	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching Data driven instruction	**Motivate Goals** Praise/reinforce **Teach Techniques** Practical soc. support	**Randomisation** Number of units randomised *18*	**Outcome types** Test score - English/native language/speech/literacy

	Overall pupil disadvantage Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Subject targeted** Maths **Kennedy's 'persistent problems'** Portraying curriculum Exposing student thinking **Dosage (No. of intended hours per participating teacher)**	Feedback on practice Modelling **Embed Practice** Self-monitoring practice	**Cluster level attrition (if clustered)** Not acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	Practice change outcome **Include in meta-analysis?** No
Sutherland (2019)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *30%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** Maths **Kennedy's 'persistent problems'** Unclear or other - add note **Dosage (No. of intended hours per participating teacher)** *8*	**Teach Techniques** Instruction Practical soc. support Feedback on practice Rehearsal **Embed Practice** Action planning Context-specific repetition	**Randomisation** Number of units randomised *108 classes* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?**	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

				Pre-registration / pre-specification / analysis plan	
Taylor (2015)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *45.5* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Inquiry / discovery / problem based teaching Formative assessment / responsive teaching **Subject targeted** Science **Kennedy’s ‘persistent problems’** Portraying curriculum Exposing student thinking **Dosage (No. of intended hours per participating teacher)** *112*	**Instil Insight** Manage cognitive load **Teach Techniques** Instruction Practical soc. support Modelling	**Randomisation** Number of units randomised *18* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Science **Include in meta-analysis?** Yes
Thiede (2018)	**Country** US **Age Group/s** Primary/elementary **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Formative assessment / responsive teaching Unclear or other - add note *ambitious instruction* **Subject targeted** Maths	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice **Embed Practice** Action planning	**Randomisation** Number of units randomised *Bit confused - 2 sets of 4 schools per year. Appears to be 3 years in the study so presumably 8 schools per year, 24 in total.* **Cluster level attrition (if**	**Outcome types** Test score - Maths **Include in meta-analysis?** No

		Kennedy's 'persistent problems' Exposing student thinking **Dosage (No. of intended hours per participating teacher)** *No hours for second 'component'*		**clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	
Tolan (2020)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *52* **Early-career / newly-qualified teachers?** Yes	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Containing behaviour **Dosage (No. of intended hours per participating teacher)** *2 days training but unclear for the dosage of coaching*	**Motivate Goals** Goal setting Praise/reinforce **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Action planning Self-monitoring practice Context-specific repetition	**Randomisation** Number of units randomised *228* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy Test score - Maths **Include in meta-analysis?** Yes

Torgerson (2014)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *31.4* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *24*	**Teach Techniques** Instruction	**Randomisation** Number of units randomised *110* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Tracey (2019)	**Country** UK **Age Group/s** Primary/elementary **Overall pupil disadvantage** % eligible for free or reduced-price lunch *45.5%* **Early-career / newly-qualified**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum	**Motivate Goals** Goal setting **Teach Techniques** Instruction	**Randomisation** Number of units randomised *155* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	teachers? No (mixed or experienced), OR unclear	**Dosage (No. of intended hours per participating teacher)** *32*		**Test type** Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	
Trial of Embedding Formative... (2018)	**Country** UK **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *29.5%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Exposing student thinking **Dosage (No. of intended hours per participating teacher)** *Not specific*	**Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling **Embed Practice** Action planning	**Randomisation** Number of units randomised *140* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	**Outcome types** Test score - Other (specify) **Include in meta-analysis?** Yes
van der Scheer (2018)	**Country** Netherlands Not UK or US **Age Group/s**	**Broad area of focus of the PD** Data driven instruction **Subject targeted**	**Instil Insight** Revisit material **Motivate Goals** Goal setting	**Randomisation** Number of units randomised *60* Notes - for discussion	**Outcome types** Test score - Maths **Include in meta-**

	Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	Maths **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *45*	Praise/reinforce **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Action planning Self-monitoring practice Context-specific repetition	*60 classes in 60 schools.* **Cluster level attrition (if clustered)** Not acceptable attrition **Pupil level attrition** Not acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Not reported	**analysis?** Yes
Vaughn (2015)	**Country** US **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *39* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Other subjects **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**	**Motivate Goals** Goal setting **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling **Embed Practice** Context-specific repetition	**Randomisation** Number of units randomised *85* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam Test developed or adapted	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

				by researchers or teachers **Analysis pre-specified?** Not reported	
Vernon-Feagans (2010)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *Only reported at school not pupil level* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Formative assessment / responsive teaching **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)**	**Teach Techniques** Instruction Practical soc. support Modelling Rehearsal	**Randomisation** Number of units randomised *4* Notes - for discussion *99% sure this has been misanalysed* **Cluster level attrition (if clustered)** Not acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Include in meta-analysis?** No
Vernon-Feagans (2013)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported	**Broad area of focus of the PD** Formative assessment / responsive teaching Unclear or other - add note **Subject targeted** English/native	**Instil Insight** Revisit material **Teach Techniques** Instruction Feedback on practice Modelling Rehearsal	**Randomisation** Number of units randomised *15* **Cluster level attrition (if clustered)** Acceptable attrition	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	No mention of lunch % **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum Exposing student thinking **Dosage (No. of intended hours per participating teacher)** *No total given*		**Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	
Wasik (2011)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** % eligible for free or reduced-price lunch *100%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note *not specifically one of these* **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum **Dosage (No. of intended hours per participating teacher)** *exact hours unclear*	**Instil Insight** Manage cognitive load Revisit material **Motivate Goals** Credible source Praise/reinforce **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling **Embed Practice** Context-specific repetition	**Randomisation** Number of units randomised *3* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy Practice change outcome **Include in meta-analysis?** Yes

Wasik (2020)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** % eligible for free or reduced-price lunch *87%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Data driven instruction **Subject targeted** English/native language/speech/literacy **Kennedy's 'persistent problems'** Portraying curriculum Enlisting participation **Dosage (No. of intended hours per participating teacher)** *Not clear how long the coaching was altogether to then add to the 12hrs of training.*	**Motivate Goals** Praise/reinforce **Teach Techniques** Instruction Feedback on practice Modelling	**Randomisation** Number of units randomised *35* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - Vocabulary **Include in meta-analysis?** Yes
Whittaker (2020)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Maths Science **Kennedy's 'persistent problems'**	**Instil Insight** Manage cognitive load **Teach Techniques** Instruction Practical soc. support Modelling	**Randomisation** Number of units randomised *140* Notes - for discussion *p.187* **Cluster level attrition (if clustered)** Not acceptable attrition	**Outcome types** Test score - Science **Include in meta-analysis?** Yes

	teachers? No (mixed or experienced), OR unclear	Portraying curriculum **Dosage (No. of intended hours per participating teacher)**		**Pupil level attrition** Not acceptable attrition **Test type** Low stakes, standardised exam Test developed or adapted by researchers or teachers **Analysis pre-specified?** Not reported	
Wiggins (2019)	**Country** UK **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *28%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Unclear or other - add note **Dosage (No. of intended hours per participating teacher)** *64*	**Teach Techniques** Instruction	**Randomisation** Number of units randomised *40* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes

Wilcox (2011)	**Country** US **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage** Pupil disadvantage unclear or not reported *No lunch stuff* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy’s ‘persistent problems’** Portraying curriculum	**Instil Insight** Revisit material **Teach Techniques** Instruction Practical soc. support Feedback on practice Modelling Rehearsal **Embed Practice** Action planning Context-specific repetition	**Randomisation** Number of units randomised *29* **Cluster level attrition (if clustered)** Attrition unclear **Pupil level attrition** Attrition unclear **Test type** Low stakes, standardised exam **Analysis pre-specified?** Not reported	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes
Wolf (2018)	**Country** US **Age Group/s** Primary/elementary **Overall pupil disadvantage** Pupil disadvantage unclear or not reported **Early-career / newly-qualified teachers?**	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** English/native language/speech/literacy **Kennedy’s ‘persistent problems’** Portraying curriculum	**Teach Techniques** Instruction Practical soc. support Feedback on practice Rehearsal	**Randomisation** Number of units randomised *79* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Attrition unclear **Test type**	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-analysis?** Yes

	No (mixed or experienced), OR unclear	**Dosage (No. of intended hours per participating teacher)** *Coaching hrs unclear*		Low stakes, standardised exam **Analysis pre-specified?** Not reported	
Worth (2017)	**Country** UK **Age Group/s** Secondary/middle/high school/Sixth Form **Overall pupil disadvantage** % eligible for free or reduced-price lunch *39.0* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Broad area of focus of the PD** Unclear or other - add note **Subject targeted** Generic (cross-curricular) **Kennedy's 'persistent problems'** Containing behaviour Enlisting participation **Dosage (No. of intended hours per participating teacher)**		**Randomisation** Number of units randomised *82* **Cluster level attrition (if clustered)** Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** High stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	**Outcome types** Test score - Maths **Include in meta-analysis?** Yes
Wright (2020)	**Country** UK **Age Group/s** Early years / pre-kindergarten **Overall pupil disadvantage**	**Broad area of focus of the PD** Data driven instruction **Subject targeted** English/native language/speech/literacy	**Motivate Goals** Credible source **Teach Techniques** Instruction Practical soc. support	**Randomisation** Number of units randomised *120* **Cluster level attrition (if clustered)**	**Outcome types** Test score - English/native language/speech/literacy **Include in meta-**

	% eligible for free or reduced-price lunch *31.1%* **Early-career / newly-qualified teachers?** No (mixed or experienced), OR unclear	**Kennedy’s ‘persistent problems’** Exposing student thinking **Dosage (No. of intended hours per participating teacher)** *6days + 17.5 hrs = (6 x8) + 17.5 = 65.5*	Feedback on practice Modelling	Acceptable attrition **Pupil level attrition** Acceptable attrition **Test type** Low stakes, standardised exam **Analysis pre-specified?** Pre-registration / pre-specification / analysis plan	**analysis?** Yes

Appendix 5: Full mechanisms coding frame

Mechanism group	Instil insight
Mechanism	0.1 Manage cognitive load
Coding guidance (Michie et al., 2013)	None—this mechanism is newly-described for this study
Additional coding guidance for this study	Present new information in ways which support understanding by reducing cognitive load and promoting encoding: code if designers explicitly do one of: a) removing less relevant content/focusing only on most relevant content; b) varying presentation and promoting encoding by providing multiple examples; c) employing dual coding.
Example	'[Training] focused on methods for helping Latinos and mainstreamed ELs to develop the academic literacy necessary to meet the CCSS-ELA, with special emphasis on interpretive reading and analytical writing.' (Olson et al., 2017, p.6) 'Our student-thinking work played out initially and quite significantly understanding of the equal sign. For example, we posed a + 5 and asked teachers to generate the range of strategies that to solve the problem. Teachers detailed incorrect solutions such 5) and 17 (adding all the numbers). They also discussed ways at the correct solution of 7. For instance, students could add 8 mine what number added to 5 equals 12, or they could look between 4 and 5 and reason that the number in the box must asked teachers to talk about how these approaches were similar what each might indicate about a child's understanding of the was to help teachers create for themselves organized ways connecting student responses.' (Jacobs et al., 2017, p.270)
Non-example	
Summary of the evidence base	Sweller et al. (2019) provide a narrative review of experimental evidence in education demonstrating the importance of cognitive load for learning. The paper sets out a range of empirical effects which promote the comprehension and learning of new ideas, including removing redundant information, employing the modality effect (dual coding) and providing completion problems or gradually reducing (fading) guidance. For evidence from another domain, see the narrative review focused on medical education by Fraser et al. (2015).
Statement of theory explaining the operation of the mechanism	'Human cognitive processing is heavily constrained by our limited working memory which can only process a limited number of information elements at a time. Cognitive load is increased when unnecessary demands are imposed on the cognitive system. If cognitive load becomes too high, it hampers learning and transfer. Such demands include inadequate instructional methods to educate students about a subject as well as unnecessary distractions of the environment. Cognitive load may also be increased by processes that are germane to learning, such as instructional methods that emphasise subject information that is intrinsically complex. In order to promote learning and transfer, cognitive load is best managed in such a way that cognitive processing irrelevant to learning is minimised and cognitive processing germane to learning is optimised, always within the limits of available cognitive capacity.' (Sweller et al., 2019, p.262)

	‘Expertise, reliant on information held in long-term memory, transforms our ability to process information in working memory and transforms us, reflecting the transformational consequences of education on individuals and societies. It follows that the major function of instruction is to allow learners to accumulate critical information in long-term memory. Because it is novel, that information must be presented in a manner that takes into account the limitations of working memory when dealing with novel information.’ (Sweller et al., 2019, p.263)

Mechanism group	Instil insight
Mechanism	0.2 Revisit prior learning
Coding guidance (Michie et al., 2013)	None—this mechanism is newly-described for this study
Additional coding guidance for this study	Revisit learning to support lasting learning: code if designers explicitly do one of: a) revisiting previous topics/techniques later in a professional development programme; b) testing participants on information provided in past sessions; c) use synoptic tasks which cause participants to draw on past learning. Do not code if synoptic task is not explicitly used for the purpose of revisiting past learning. Do not code if activity is cyclical/repeated (e.g. weekly meetings) unless past content is also revisited. Follow-up visits by coaches are insufficient to code for revisiting (most likely code for feedback): there must be an explicit reference to revisiting past learning.
Example	‘Each work-group meeting was planned to focus on one particular but ideas from earlier sessions were regularly revisited.’ (Jacobs et al., 2007, p.268) ‘The literacy coach also met with each school team for 30 min bi-weekly through webcam technology to further reinforce the strategies and problem solve about individual children.’ (Vernon-Feagans et al., 2013, p.4) ‘Program staff provide the program content—the three reading strategy components and the three instructional strategy components—to teachers through these activities using a spiral approach, in which teachers revisit the content repeatedly and in greater depth over the course of the two years.’ (Abe et al., 2012, p.8)
Non-example	
Summary of the evidence base	A large number of experiments have been conducted in both lab and real classroom settings showing that quizzing (one form of revisiting) promotes learning (Moreira et al., 2019; Pastotter et al., 2014). Importantly, while quizzing is the most effective form of revisiting, simply revising, re-reading or being re-exposed to material over multiple occasions does increase long-term retention (Adesope et al., 2017; Rohrer, 2015; Yang et al., 2021).
Statement of theory explaining the	‘Thus, because recollection consists in a more elaborative retrieval of studied information, and involves the reinstatement of episodic/contextual features of the original study event, according to

operation of the mechanism	current testing effect theories free- and cued-recall tests are expected to produce greater testing effects than recognition.' (Moreira et al., 2019, p.2) Both encoding and retrieval explanations have been put forth to account for the forward effect of testing on list learning in laboratory studies. Retrieval explanations typically assume that recall testing between the study of lists promotes contextual list segregation, which may enhance list differentiation and reduce interference between lists at test … [E]ncoding explanations of the forward effect of testing assume that recall testing of prior non-target materials improves encoding of the subsequently studied target material. Specifically, it has been suggested that testing induces a reset of the encoding process, making the encoding of the later lists as effective as the encoding of the earlier lists (Pastötter et al., 2011), or a change in participants' encoding strategy, enhancing elaborative encoding for the later lists compared to the earlier lists (Wissman et al., 2011).' [Pastotter et al., 2014, p.2]

Mechanism group	Motivate goals
Mechanism	1.1 Goal setting
Coding guidance (Michie et al., 2013)	Set or agree on a goal defined in terms of the behavior to be achieved. Note: only code goal setting if there is sufficient evidence that goal set as part of intervention; if goal unspecified or a behavioral outcome, code 1.3, Goal setting (outcome); if the goal defines a specific context, frequency, duration, or intensity for the behavior, also code 1.4, Action planning.
Additional coding guidance for this study	Required/expected (action planning if lesson planning). Must be a behaviour for the teacher to do (not just a behaviour for students).
Example	'Teachers left each session with a gap task to try out and evaluate with their own classes prior to the next CPD session.' (Kitmitto et al., 2018, p.9) 'Participating teachers agreed to teach core concepts in economics, as identified by national economics standards, using the curricular materials provided.' (Finkelstein, 2010, p.ix)
Non-example	'Pacific CHILD focuses on interactive tasks and encourages teachers to adopt group-based activities in the classroom.' (Abe et al., 2012, p.6) *'Encourages' is too weak to represent a goal; the focus is insufficiently clear on what teachers are to adopt.* 'The DDI intervention did not require that teachers implement specific instructional approaches, but they were generally expected to make greater use of evidence-based instructional strategies, such as reviewing and adjusting students' small-group assignments, using differentiated instruction, re-teaching difficult material, and increasing time spent on Instruction.' (Gleason et al., 2019, p.4)

Summary of the evidence base	A systematic review tested the effect of setting goals compared to otherwise identical interventions without goal setting (Epton et al., 2017), finding 384 independent effect sizes from randomised studies. The studies included set goals in a range of settings: cognitive, sporting, productive (such as building a Lego model), health-related, and educational. Meta-analysis demonstrated that setting goals substantially increased the likelihood of behaviour change.
Statement of theory explaining the operation of the mechanism	'Goal setting theory (Locke & Latham, 2002; Locke & Latham, 2006) was derived from a series of industrial/organizational psychology experiments regarding work-related task performance (Locke & Latham, 2002). The original theory posits that goal setting will promote behavior change when two conditions are met: (a) the goal must be conscious and specific, and (b) the goal must be sufficiently difficult (i.e., over and above what is usually achieved).' (Epton, 2017, p.1183)

Mechanism group	Motivate goals
Mechanism	9.1 Credible source
Coding guidance (Michie et al., 2013)	'Present verbal or visual communication from a credible source in favour of or against the behavior. Note: code this BCT if source generally agreed on as credible e.g., health professionals, celebrities or words used to indicate expertise or leader in field and if the communication has the aim of persuading.'
Additional coding guidance for this study	Must state that source offers evidence/statements in favour of/against specified behaviour (not just that credible source was present). Having teacher educators in the room insufficient—they must appeal to a third party (published research, education celebrities, expert teachers) to justify behaviour.
Example	'Cases also provided additional text pages, including research-based rationales for teaching strategies highlighted in the case.' (Powell et al., 2010, p.303)
Non-example	'Training was led by the developers of the Pathway Project and supported by literacy coaches who participated as Pathway teachers in a previous quasi-experimental research study.' (Olson et al., 2017, p.6) *This is not explicit as to what the coaches said or did.* 'The front matter of the teachers' guide explains how the curriculum is designed to promote students' mastery of the state science standards, why science inquiry is crucial for facilitating students' understanding of the big ideas in the state science standards, how teachers can help students progress toward student-initiated inquiry, and how teachers can engage all students, especially ELLs, in language and literacy development.' (Llosa et al., 2016, p.403) *No independent source of supporting evidence is offered to teachers.*
Summary of the evidence base	The evidence for the influence of a credible source primarily focuses on the power of evidence to sway people's attitudes, beliefs and intentions. For example, a narrative review (Hornikx, 2015) examined the persuasiveness of different forms of evidence (such as statistical and anecdotal), again in terms of acceptance of an idea, or more consequential actions such as intentions, showing more powerful

	influences for statistical and factual evidence than for anecdotal in domains such as health behaviour, belief about crime and administrative changes like the introduction of a new science programme. A meta-analysis found that a complete argument with an explicit justification was consistently more persuasive than not giving such justifications (O'Keefe, 1998), for messages around crime, the value of education, risk communication, and medication.
Statement of theory explaining the operation of the mechanism	'The persuasive function of evidence has been the object of several direct and indirect theoretic probes. One perspective holds that evidence improves the persuasiveness of a message since audiences may expect effective communicators to demonstrate their credentials by its prominent use … Since people may want sources of information to be in command of the facts, it may be a part of the Western cultural tradition that advocates are expected to prove their worthiness to speak by backing up what they say.' (Reinard, 1988)

Mechanism group	Motivate goals
Mechanism	10.4 Praise/reinforce
Coding guidance (Michie et al., 2013)	Arrange verbal or non-verbal reward if and only if there has been effort and/or progress in performing the behavior (includes 'Positive reinforcement').
Additional coding guidance for this study	Must be subsequent to action (can't be prior encouragement).
Example	'These coaches conducted three informal, nonevaluative classroom observations and provided teachers with detailed written feedback identifying areas of strength, areas for improvement, and specific suggestions of classroom practices teachers could implement.' (Olson et al., 2017, p.10)
Non-example	
Summary of the evidence base	A narrative review (Delin & Baumeister, 1998) drew on both laboratory and field experiments to illustrate the positive effects of praise on people's self-concept, the standards they sought to achieve, their motivation, and their liking for the person praising them. These experiments were conducted on children and adults, and across a range of domains including mathematics, game performance, and art.
Statement of theory explaining the operation of the mechanism	'Praise is an interpersonal communication in which one person expresses a positive evaluation of another person. The potential effects thus include cognitive responses, emotional responses, motivational responses, and interpersonal consequences.' (Delin and Baumeister, 1998, p.223)

Mechanism group	Teach techniques
Mechanism	4.1 Instruction

Coding guidance (Michie et al., 2013)	Advise or agree on how to perform the behavior (includes 'Skills training').
Additional coding guidance for this study	Michie et al. state that 'when the person attends classes such as exercise or cookery, code 4.1, Instruction on how to perform the behavior, 8.1, Behavioral practice/rehearsal and 6.1, Demonstration of the behavior.' We will not follow her in making these assumptions: we will only code practice/rehearsal and modelling if these are mentioned explicitly.
Example	'The teaching guide is the cornerstone of each module. It lays out for teachers the problem statement, introduction, placement in curriculum, concepts taught, objectives, content standards, time required, lesson description, resource materials, sequence of the unit, procedures, and do's and don'ts.' (Finkelstein, 2010, p.5)
Non-example	
Summary of the evidence base	A recent narrative review sets out the importance of clear instruction, designed to maintain a manageable cognitive load, in educational settings (Sweller et al., 2019). This builds on the results of a prior review that provides experimental evidence demonstrating the consistent superiority of clear, guided and explicit instruction over unguided, discovery-based and problem-based learning in science learning and medical education (Kirschner et al., 2006).
Statement of theory explaining the operation of the mechanism	'Direct instructional guidance is defined as providing information that fully explains the concepts and procedures that students are required to learn as well as learning strategy support that is compatible with human cognitive architecture.' (Kirschner et al., 2006, p.75)

Mechanism group	Teach techniques
Mechanism	3.2 Practical social support
Coding guidance (Michie et al., 2013)	Advise on, arrange, or provide practical help (e.g. from friends, relatives, colleagues, 'buddies' or staff) for performance of the behavior.
Additional coding guidance for this study	None
Example	'The intervention team requested that at least two teachers from each school participate in the CPD sessions to facilitate ongoing collaboration and mutual support.' (Hanley et al., 2015, p.9) 'This was followed by a 20- to 30-minute phone conference between the teacher and the coach to discuss instructional strategies that would foster positive teacher–student relationships and teachers' ability to sensitively engage all students.' (Allen et al., 2015, p.481) 'On four occasions—at the start of the semester and then roughly timed to the completion of the curriculum modules—teachers participated in a group conference call with the developers and the study team to discuss progress. Teacher participation on these calls varied from a high of 24 to a low of 14. Teachers raised issues of pacing, handling particular content,

	and juggling other curricular requirements of their schools and districts. They also raised challenges that they had faced and asked for feedback and support. The calls were collegial and afforded the teachers a professional community to discuss challenges.' (Finkelstein, 2010, p.50)
Non-example	
Summary of the evidence base	A systematic review of the effects of peer support on health promotion and disease prevention (Ramchand et al., 2017) found that peer educators and facilitators supported change in knowledge, attitudes, and beliefs, while individual peer support influenced behaviour change. Reviews of sub-domains in health have shown evidence of the benefits of peer support in increasing breastfeeding (Jolly et al., 2012) and improving outcomes for diabetes patients (Dale et al., 2012). Experiments have also found that collaborative social support helps with skill acquisition when added to other activities in medical training (Grierson et al., 2012).
Statement of theory explaining the operation of the mechanism	'Peers often share a common culture, language, and knowledge about the problems that their community experiences. Moreover, in the research context, peers must share a health problem (e.g. newly diagnosed tuberculosis) or a potential for change in their health status (e.g. breastfeeding for new mothers).' (Webel et al., 2010, p.247) 'Peer support, within the health care context, is the provision of emotional, appraisal, and informational assistance by a created social network member who possesses experiential knowledge of a specific behaviour or stressor and similar characteristics as the target population, to address a health-related issue of a potentially or actually stressed focal person..' (Dennis, 2003, p.329)

Mechanism group	Teach techniques
Mechanism	6.1 Modelling
Coding guidance **(Michie et al., 2013)**	Provide an observable sample of the performance of the behaviour, directly in person or indirectly, for example, via film, pictures, for the person to aspire to or imitate (includes 'Modelling').
Additional coding guidance for this study	If teachers are watching videos of each other as part of a community of practice, do not code as modelling (video is for analysis, not—necessarily—a model).
Example	'[Teachers] watched exemplar videos of teachers employing these principles.' (Allen et al., 2015, p.480) 'Delivery of the curriculum modules is modeled by master teachers with years of experience delivering the curriculum, thus melding content and pedagogical practice.' (Finkelstein, 2010, p.8)
Non-example	

Summary of the evidence base	Renkl (2014) offers a narrative review drawing from different research areas to show the effects of providing examples on cognitive skill development. Narrative reviews have also found evidence for the causal effect of modelling (particularly relative to practice-alone) for motor skill acquisition in medical and surgical education (Cordovani & Cordovani, 2016; Harris et al., 2018).
Statement of theory explaining the operation of the mechanism	'Learning from [worked examples] means that learners study (usually several) problems for which the solution is given (Renkl, 2005) before they are confronted with problem-solving demands. Learners should acquire a basic understanding of domain principles while studying examples, which provides a basis for later meaningful problem solving.' (Renkl, 2014, p.4)

Mechanism group	Teach techniques
Mechanism	2.2, 2.7 Feedback
Coding guidance **(Michie et al., 2013)**	Monitor and provide feedback on the outcome of performance of the behavior.
Additional coding guidance for this study	Michie distinguishes between feedback on practice and on outcomes; we do not.
Example	'These coaches conducted three informal, nonevaluative classroom observations and provided teachers with detailed written feedback identifying areas of strength, areas for improvement, and specific suggestions of classroom practices teachers could implement.' (Olson et al., 2010, p.10) 'Trained coaches review recordings that teachers submit and select brief video segments that illustrate either positive teacher interactions or areas for growth in one of the dimensions in the CLASS-S.' (Allen et al., 2015, p.476)
Non-example	
Summary of the evidence base	A meta-analysis (Kluger and DeNisi, 1996) compared the effects of feedback on performance in experimental and quasi-experimental studies, incorporating 607 effects. Overall, feedback had a substantial positive effect on performance. With respect to motor skills, causal evidence on the positive effects of expert feedback comes from studies in dental and medical education (Al-Saud et al., 2017; Hatala et al., 2014).
Statement of theory explaining the operation of the mechanism	FIT has five basic arguments: (a) Behavior is regulated by comparisons of feedback to goals or standards, (b) goals or standards are organized hierarchically, (c) attention is limited and therefore only feedback-standard gaps that receive attention actively participate in behavior

	regulation, (d) attention is normally directed to a moderate level of the hierarchy, and (e) FIs change the locus of attention and therefore affect behavior. (Kluger and DeNisi, 1996, p.259)

Mechanism group	Teach techniques
Mechanism	8.1 Rehearsal
Coding guidance **(Michie et al., 2013)**	Prompt practice or rehearsal of the performance of the behavior one or more times in a context or at a time when the performance may not be necessary, in order to increase habit and skill. Note: if aiming to associate performance with the context, also code 8.3, Habit formation.
Additional coding guidance for this study	
Example	'Training materials from the Learning Language and Loving It leader's guide were utilized to include … teacher roleplaying activities.' (Cabell, 2011, p.320)
Non-example	'The purpose of the first training session was to introduce the project and first unit of curriculum and to give teachers practice with project techniques and materials.' (August et al., 2014) *This does not specify what is being practised.*
Summary of the evidence base	Two meta-analyses investigating the relationship between accumulated hours of practice and performance showed high levels of deliberate practice were associated with higher performance in a range of domains (Ericsson and Harwell, 2019; Macnamara et al., 2016). A meta-analysis of causal studies in medical education found that practice in simulations consistently outperformed medical education without practice in simulations (McGaghie et al., 2011).
Statement of theory explaining the operation of the mechanism	'Practice improves accuracy and speed of performance on cognitive, perceptual, and motor tasks.' (Ericsson et al., 1993, p.367)

Mechanism group	Embed practice
Mechanism	7.1 Prompts/cues
Coding guidance **(Michie et al., 2013)**	Introduce or define environmental or social stimulus with the purpose of prompting or cueing the behavior. The prompt or cue would normally occur at the time or place of performance. Note: when a stimulus is linked to a specific action in an if-then plan including one or more of frequency, duration or intensity also code 1.4, Action planning.

Additional coding guidance for this study	Must be a specific action AND at specific time to qualify.
Example	'Trainers contacted teachers between sessions only to send a reminder about the gap tasks.' (Hanley et al., 2015, p.9) 'Diagnostic tools consist of a series of checklists that align with the three main components of the simple view of writing (transcription, text generation, and self-regulation). Each checklist encourages the teacher to observe both the student's writing process (behavior during writing) and product (the result of the student's work) to gain insights into the student's strengths and needs. For example, the transcription checklist includes processes and products related to both handwriting (e.g. "Does the student hold the pen or pencil comfortably? Write fluidly without excessive scribbling or erasing?") and spelling (e.g. "Does the student consistently use the correct consonant at the beginning and end of words?"). The teacher can then use this information to determine the focus and content of instruction and match minilessons to the student's needs (e.g. the teacher might select spelling lessons for a student with relative strengths in handwriting and weaknesses in spelling).' (McMaster et al., 2020, p.3)
Non-example	
Summary of the evidence base	Prompts/cues have been shown to prompt increased goal-directed behaviour in experimental research on gym attendance (Calzoari & Nardetto, 2017), changing medical doctors' clinical practice (Shojania et al., 2010), and getting patients to attend medical appointments (Guy et al., 2012). With respect to habitual behaviour, multiple reviews have found that environmental cues help embed automatic behaviours in a range of settings (Gardner & Rebar, 2019; Neal & Wood, 2009; Wood & Neal, 2007).
Statement of theory explaining the operation of the mechanism	With respect to goal-directed behaviour, prompts often work by simply reminding somebody to follow through on an intention. With respect to habitual behaviours, 'perception of the relevant context cues automatically activates the mental representation of the habitual response.' (Wood and Runger, 2016, p.292)

Mechanism group	Embed practice
Mechanism	1.4 Action planning
Coding guidance **(Michie et al., 2013)**	Prompt detailed planning of performance of the behavior (must include at least one of context, frequency, duration and intensity). Context may be environmental (physical or social) or internal (physical, emotional or cognitive) (includes 'Implementation Intentions'). Note:

	evidence of action planning does not necessarily imply goal setting, only code latter if sufficient evidence.
Additional coding guidance for this study	Includes lesson planning, if planning is to apply techniques in lessons.
Example	'Staff teams in each of the eight research schools worked to develop their plans to implement teaching initiatives focused on enhancing student literacy outcomes via a focus on two of the four aspects of oral language competence targeted by ICPALER.' (Snow, 2014, p.500)
Non-example	'Teachers were given time to meet in grade-level groups and as school teams to discuss how to modify the materials to meet their specific students' needs.' (Olson et al., 2017, p.7) *No specific context/frequency for action.*
Summary of the evidence base	A review of 94 experimental studies found strong evidence that setting implementation intentions about how to act and when improved performance for a range of populations in health, goal pursuit, academic and laboratory tasks. The review included both correlational and experimental research, but found that implementation intentions were similarly effective in both cases (Gollwitzer and Sheeran, 2006).
Statement of theory explaining the operation of the mechanism	'Implementation intentions should enhance people's ability to initiate, maintain, disengage from, and undertake further goal striving and thereby increase the likelihood that strong goal intentions are realized successfully. In other words, this form of planning is expected to bridge the intention–behavior gap.' (Gollwitzer and Sheeran, 2006, p.82)

Mechanism group	Embed practice
Mechanism	2.3, 2.4 Self-monitoring
Coding guidance **(Michie et al., 2013)**	Establish a method for the person to monitor and record their behavior(s) as part of a behavior change strategy. Note: if monitoring is part of a data collection procedure rather than a strategy aimed at changing behavior, do not code.
Additional coding guidance for this study	Must be a mechanism to record actions towards a specific goal—not just reflection upon them. Michie distinguishes between self-monitoring of outcomes and self-monitoring of behaviour, but we can include self-monitoring of outcomes and behaviour under this code. May include examining student learning if the search is for specific outcomes of teacher actions (not just looking at student work generally).
Example	'Teachers were asked to record the scenario that they presented to pupils, the pupil responses, and reflections on how it worked (for

	example, how it impacted learning, engagement, or pupil management).’ (Hanley et al., 2015, p.9) ‘Each teacher is asked to observe his/her behavior and subsequent student reactions, and to then respond to coach prompts that call attention to the connection between teacher behavior and student responses.’ (Allen et al., 2015, p.476)
Non-example	‘Central to our professional development work were these tasks that were generated in the work-group meetings, used in teachers' classrooms, back to work-group meetings in the form of written student work classroom interactions.’ (Jacobs et al., 2007, p.268) *This is generalised reflection without an explicit monitoring process.*
Summary of the evidence base	A review of weight loss interventions found a consistent association between self-monitoring and weight loss, in both correlational and experimental research (Burke et al., 2011). Self-monitoring has also been shown to change behaviour in experiments on sleep hygiene (Todd & Mullan, 2014) and in meta-analysis of experimental studies on reducing sedentary behaviours (Compernolle et al., 2019).
Statement of theory explaining the operation of the mechanism	‘Self-regulation theory posits that self-monitoring precedes self-evaluation of progress made towards one’s goal and self-reinforcement for the progress made (2); thus, the process of changing habits requires well developed self-regulatory skills (3, 4). Self-monitoring is central to this process and includes deliberate attention to some aspect of an individual’s behavior and recording some details of that behavior. In order to change behaviors, individuals need to pay adequate attention to their own actions, as well as the conditions under which they occur and their immediate and long-term effects (5). Thus, successful self-regulation depends in part on the truthfulness, consistency and timeliness of self-monitoring in relation to the performance of the target behavior, e.g. eating (5).’ (Burke et al., 2011, p.2)

Mechanism group	Embed practice
Mechanism	8.3 Context-specific repetition
Coding guidance **(Michie et al., 2013)**	Prompt rehearsal and repetition of the behavior in the same context repeatedly so that the context elicits the behavior.
Additional coding guidance for this study	SAME action in SAME realistic context, twice or more. Michie states ‘also code 8.1, Behavioral practice/rehearsal’—our distinction here is more nuanced: code only 'rehearsal' if out of context (practising in a training room, for example); code only 'context-dependent repetition' if practice is in realistic environment (e.g. teacher's own classroom).

	Code if a coach/observer visits your lesson to watch you enacting the intervention (provided this is not solely for data collection purposes).
Example	'Our coaching model reflects these ideas, in that we hold coaches accountable for enacting a well-specified routine during their coaching conversations with teachers (described below) and embed this conversation in a wider set of routines for teachers to follow. We also set expectations, to the extent their schedules allow, that teachers will engage in a coaching cycle as frequently as every 2 weeks. Our goal is to increase the interpersonal accountability between teachers and coaches by ensuring that teachers know they must take action as the next meeting with the coach draws nearer.' (Kraft & Hill, 2020, p.2383)
Non-example	
Summary of the evidence base	See the evidence base for rehearsal above.
Statement of theory explaining the operation of the mechanism	'Interventions based on the habit-formation model differ from non-habit-based interventions in that they include elements that promote reliable context-dependent repetition of the target behavior, with the aim of establishing learned context–action associations that manifest in automatically cued behavioral responses.' (Gardner & Rebar, 2019, p.1)

Appendix 6: Sensitivity tests and subgroups

Table 7b: Sensitivity checks based on indicators of study quality

	Full Sample	Low attrit.	High attrit.	>50 units	<51 units	Pre-reg.	Not pre-reg.
Estimate	0.05**	0.018*	0.082**	0.036**	0.096**	0.005	0.074**
Std. Error	(0.009)	(0.007)	(0.014)	(0.009)	(0.018)	(0.007)	(0.012)
k[n]	205[104]	49[36]	156 [68]	106[65]	104[39]	32[26]	173[78]
Difference	NA	*p*=0.006		*p*=0.008		*p*=0.0001	

Note. Low/High attrit. (attrition) is based on the What Works Clearinghouse 'cautious' standards for acceptable attrition at both the cluster and pupil level. >50 units means that the trial randomised more than 50 units to treatment and control. Pre-reg. = the trial was pre-registered before it was conducted. Numbers in round parentheses are standard errors. k is number of effect sizes and n is number of experimental studies. **p<0.01 different from zero. *p<0.05 different from zero. Calculated using random effects meta-analysis, incorporating all standardised test score outcome measures using robust variance estimation.

Table 7c: Further sensitivity checks

	Transformed	Hedges' *g*	Final year of intervention	Subsequent year
Estimate	0.026**	0.048**	0.044**	0.031
Std. deviation	(0.001)	(0.009)	(0.008)	(0.023)
k[n]	205[104]	201[102]	174[87]	11[21]

Note. Transformed = analysed using a square root transformation (to enhance normality) then backward transformed for presentation of numerical results. Two studies get dropped from Hedges' *g* analysis. First, Wasick (2011) because the n=3 units randomised yields an illogical value for the standard error of Hedges' *g*. Second, Boardman (2015) because they randomise sections of teaching but do not report the number of sections that were randomised, which makes it impossible to calculate Hedges' *g*. **p<0.01 different from zero. *p<0.05 different from zero. Calculated using random effects meta-analysis, incorporating all standardised test score outcome measures using robust variance estimation.

Table 7d: Overall impact of PD on pupil achievement, by pupil and teacher demographics

	Full sample	Early-career teachers	Not early-career teachers	High deprivation pupils	Low deprivation pupils
Estimate	0.05**	0.014	0.05**	0.054*	0.003
Std. deviation	(0.009)	(0.013)	(0.009)	(0.015)	(0.009)
k[n]	205[104]	5[3]	200[101]	42[22]	29[20]
Difference		*p*=0.001		*p*=0.001	

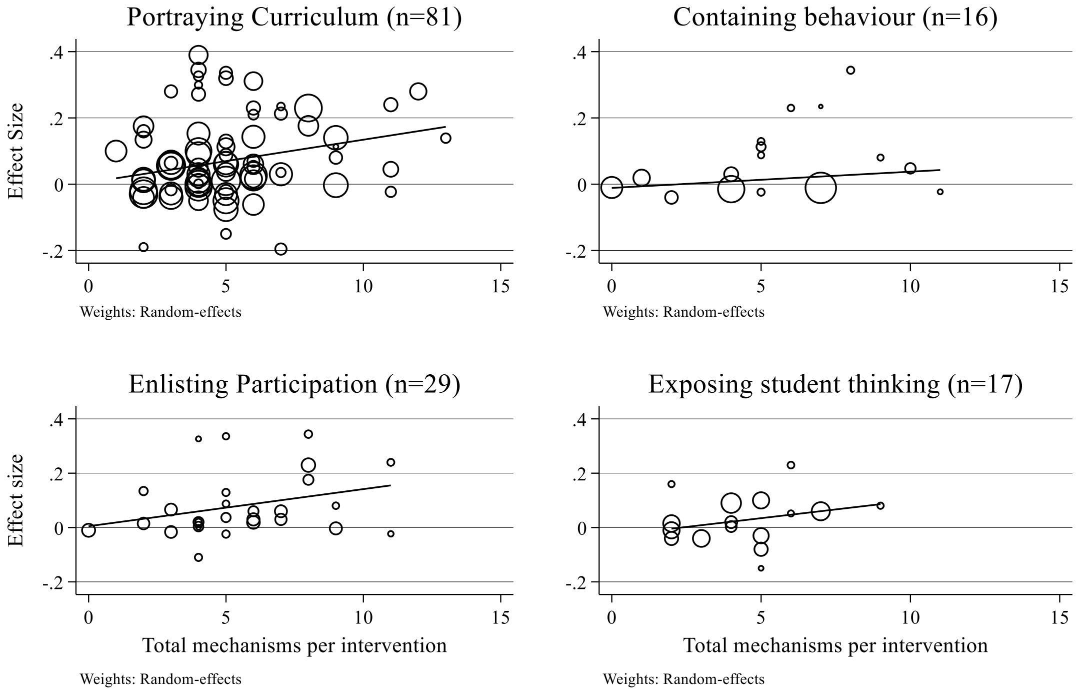
Note. Early career teachers = less than three years of experience. High deprivation pupils = greater than 51% (study-level median) eligibility for any free or reduced price lunch. Note that eligibility for criteria for free or reduced price lunch probably varies substantially across studies. Numbers in round parentheses are standard errors. k[n]: k is number of effect sizes and n is number of experimental studies. **p<0.01 different from zero. *p<0.05 different from zero. Calculated using random effects meta-analysis, incorporating all standardised test score outcome measures using robust variance estimation.

Table 7e: Investigating potential explanations for smaller effect sizes in pre-registered trials

	Indicators of higher methods standards			Indicators of less effective PD		
	High stakes test score	'Acceptable' attrition	No. of units randomised	PD + curric/tech	No. of mechanisms	I, G, T & P mechanisms
Pre-reg	34.6%	65.4%	149	42.3%	4.2	7.7%
Not pre-reg	25.6%	24.4%	67.9	44.9%	5.2	8.9%

Note. 'Acceptable' attrition is defined in line with the What Works Clearinghouse standards. 'PD + curric/tech' implies the PD programme also had a curriculum reform or educational technology element. 'I, G, T & P mechanisms' implies that a PD programme has at least one mechanism in each of the Insight, Goals, Technique and (embed) Practice categories.

Figure 13: Association between the number of mechanisms included in a PD programme and the impact of the programme on pupil test scores, by whether PD addresses each of Kennedy's persistent challenges



Note. n = number of studies. Line of best fit derived from a meta-regression using random effect weights and either the primary outcome if specified in the study, or else one randomly selected outcome per study. Effect sizes >.5 or <-.2 are used in the underlying meta-regression but not shown in the chart in order to aid visual clarity. A single PD programme may address more than one of Kennedy's persistent challenges of teaching.

Appendix 7: Publication bias

In our protocol, we committed to conducting three publication bias tests: trim-and-fill, *p*-curve, and selection modelling. The rationale for using all three is that they test for different types of publication bias and rely on different assumptions (see Harrer et al., 2019). Our primary interest is in understanding whether and how much the overall meta-analytic impact estimates vary after adjustment for possible publication bias using each method. The results of these tests are presented in this section.

We start by presenting a contour-enhanced funnel plot (Figure 14) to show the distribution of effect sizes in the dataset in the context of their statistical significance; this is because one of the key assumptions of publication bias is that publication decisions are influenced by the statistical significance of the study findings. Given the number of studies included in the review for which we could not calculate an effect size (n = 17), it is helpful to consider whether any hypothetically missing studies are likely to be due to non-reporting biases related to significance levels.

Contour-enhanced funnel plots (Peters et al., 2008) show the distribution of effect sizes plotted against their standard errors, with contour lines to denote popular statistical significance levels (typically *p* < .01, .05, and .10). The plotted contour regions (shaded areas of the plot) show the regions of non-significance at each of the specified levels for the effect size estimates. Although the plot (Figure 14) suggests that there are more effect sizes in the white region to the right of the null effect (effect size = 0) than the left, we still see a very high proportion of the effect sizes falling in the shaded areas of non-significance at the three levels (*p* < .01, .05, and .10), which is encouraging.

Figure 14: Contour-enhanced funnel plot

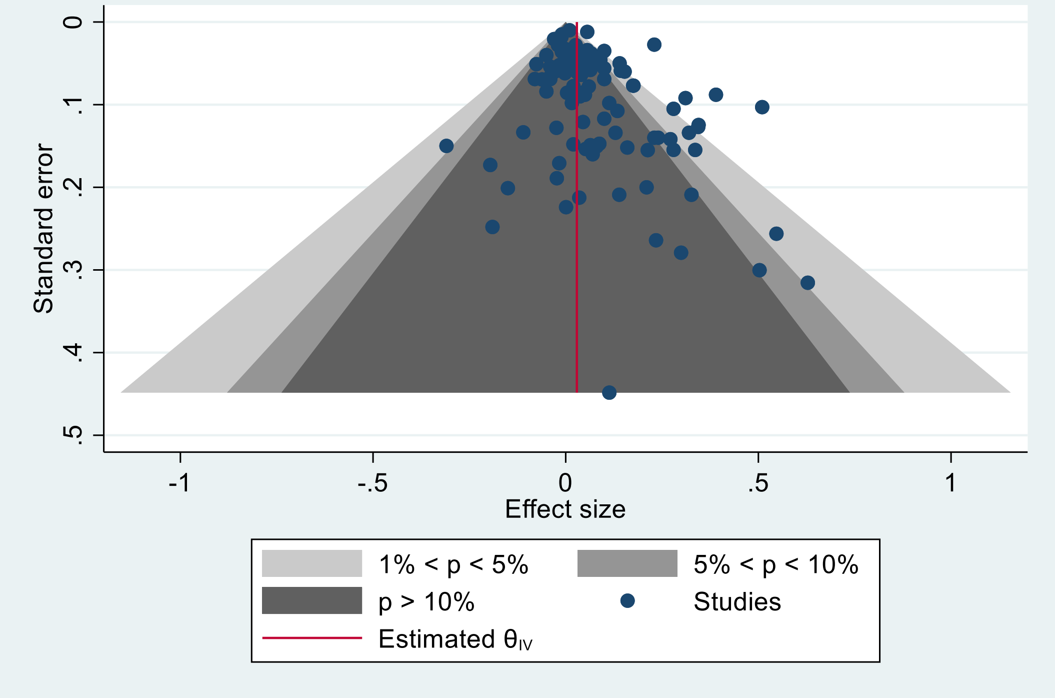


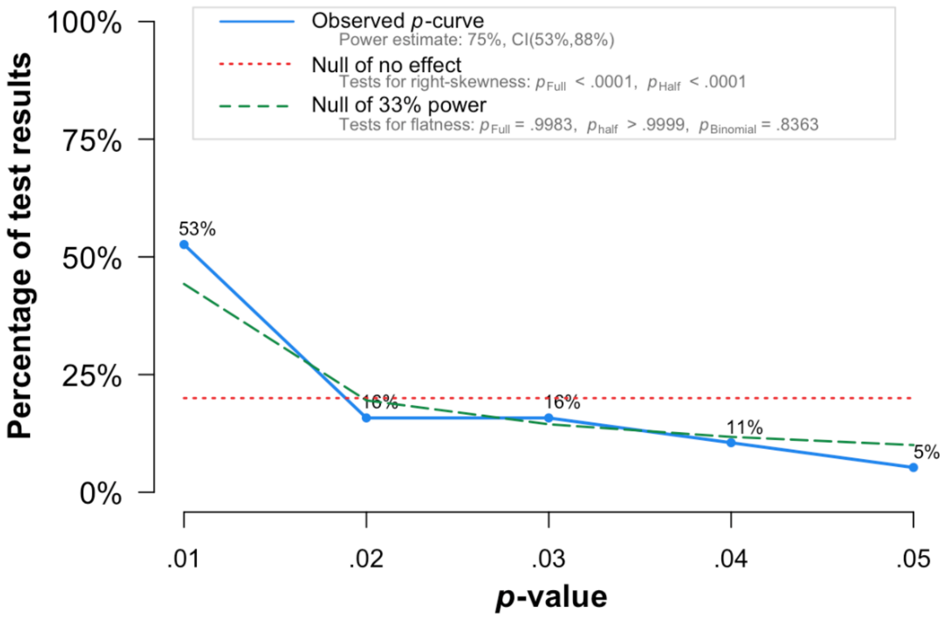
Table 17 shows the overall meta-analytic average effect sizes following the Duval & Tweedie (2000) trim and fill method. Odd rows show the meta-analytic average effect sizes based on the studies/outcomes that we actually observe in our sample. Even rows show the version in which theoretically missing studies/outcomes are imputed to restore any imbalance in the funnel plot. Rows 1 and 2 use one randomly selected effect size and associated standard error per study. Rows 3 and 4 use the average of the effect sizes and standard errors within each study. The table shows that imputing theoretically missing effect sizes attenuates the meta-analytic average effect size by about half. However, all of the 95% confidence intervals still exclude zero and therefore would not change the conclusions of the review.

Table 17: Trim and fill

	Effect size	Lower limit 95% CI	Upper limit 95% CI
1. Observed effects: one random outcome per study	0.044	0.027	0.062
2. Observe & imputed: one random outcome per study	0.033	0.014	0.051
3. Observed effects: averaging outcomes in studies	0.049	0.032	0.066
4. Observed & imputed: averaging outcomes in studies	0.023	0.003	0.043

Figure 15 shows the *p*-curve (Simonsohn et al., 2014a; Simonsohn et al., 2014b) for 104 effect sizes, using one random outcome per study. The *p*-curve is right skewed, which is what would be expected in the presence of a real (non-zero) effect of PD on pupil test scores. *P*-curve analysis can also be used to produce an adjusted estimate of the meta-analytic average effect. However, *p*-curve has been shown to overestimate the true meta-analytic average effect when heterogeneity is high (Van Aert et al., 2016), and Van Aert et al (2016) recommend not using *p*-curve when the I^2 is above 50%. In our sample, the I^2 is 60%. For transparency, our estimate of adjusted meta-analytic effect is 0.58. However, we place no weight on this estimate in the presence of such high heterogeneity.

Figure 15: *P*-curve



Note: The observed *p*-curve includes 19 statistically significant ($p < .05$) results, of which 15 are $p < .025$. There were 85 additional results entered but excluded from *p*-curve because they were $p > .05$.

Note. Calculated using the default method for estimating the between-study variance in the R software: DerSimonian and Laird.

Finally, we tested for publication bias using a selection modelling approach (Vevea & Hedges, 1995). In particular, we used a weight function with two cut-points: the first at *p* = 0.1 and the second at *p* = 0.05. We chose to model publication probability in these intervals as they correspond with alpha levels commonly used or reported in social science journals (see Sutton, 2009). The unadjusted meta-analytic impact estimate was .049 and the adjusted estimate was .051. The likelihood ratio test for the difference in adjusted and unadjusted impact estimates was *p* =.90, suggesting no significant difference at conventional level.

Summarising this section, there is a slight indication from the contour-enhanced funnel plot and the trim-and-fill analysis that there may be some 'missing studies' with small sample sizes and large negative effects, but it is impossible to know if this is the case. The *p*-curve analysis and trim-and-fill both suggest that taking publication bias into account in different ways would not change the substantive conclusions of the review, while the enhanced contour funnel plot shows a very high proportion of effect sizes that are not statistically significant, thereby reassuring us that substantial non-reporting bias is unlikely in this dataset.

Appendix 8: Examples of balanced PD designs

This appendix illustrates different ways in which studies can achieve a balanced design—a design which incorporates mechanisms to increase insight, set goals, develop techniques and embed practice. Twelve studies in our sample achieved this balance. Here, we illustrate three contrasting ways in which this balance can be achieved.

Allen et al. (2015)

The 'My Teaching Partner-Secondary' programme aimed to improve student-teacher interaction by offering teachers online coaching. The programme is generic, rather than subject-specific: it does not seek to improve teacher subject knowledge. The core of the intervention was regular cycles of coaching. In each cycle, the teacher filmed a lesson; a coach watched the video, highlighting strong and weak elements of student-teacher interaction; the teacher watched highlighted sections of video; then the coach and teacher met to discuss potential improvements.

Instil insight

The coach helped the teacher to focus on specific improvement (*Insight 1: Manage cognitive load*) by selecting a few video segments to review. The segments chosen were intended to emphasise specific dimensions of student-teacher interactions: relationships, classroom organisation, and instructional support. Coaches returned to key ideas (*Insight 2: Revisit material)* by focusing on each dimension more than once, and revisiting each dimension in the second year of the programme.

Motivate goal-directed behaviour

The coach and the teacher set goals (*Motivate 1—Goal setting)* by agreeing ways to improve student-teacher relationships during their meeting. Coaches also encouraged teachers to keep pursuing successful actions (*Motivate 3—Praise/reinforce*) by highlighting teachers' strengths in the videos they reviewed.

Teach techniques

During the introductory training session, teachers were introduced to the dimensions of high-quality student-teacher interaction (*Techniques 1—Instruction).* They were shown videos of effective teaching during this session (*Techniques 3—Modelling*), and had ongoing access to a library of videos, annotated to highlight particularly effective practices. Through highlighting videos and during meetings with teachers, coaches suggested areas of improvement (*Techniques 4—Feedback).*

Embed practice

Once the coach had watched the video, the teacher was asked to review the sections the coach had highlighted. Teachers were asked to observe their own behaviour (*Practice 3—Self-monitoring*) and to consider the connection between their actions and students' reactions.

Boardman et al. (2015)

This programme aimed to improve students' reading in science and social studies lessons, using Collaborative Strategic Reading, an approach to reading comprehension which teaches students reading strategies, then encourages them to use the strategies in small cooperative learning groups. Teachers received two days of initial training in the model, followed by two after-school booster sessions tailored to their needs, support from leaders, and individual coaching. This study illustrates how a design with relatively few mechanisms (eight in total) can still achieve balance.

Instil insight

Developers revisited key ideas through the booster sessions during the school year, depending on teacher need (*Insight 2—Revisit material*). These booster sessions incorporated reminders that Collaborative Strategic Reading resources could be used in particular contexts, such as helping English language learners.

Motivate goal-directed behaviour

Developers gave teachers a specific goal (*Motivate 1—Goal setting)*: to use Collaborative Strategic Reading in at least one lesson a week, with a suitable text.

Teach techniques

Developers taught teachers the Collaborative Strategic Reading model during the initial two-day training sessions (*Techniques 1—Instruction*). Teachers received ongoing support from teacher leaders in school and from coaches (*Techniques 2—Practical social support*). Coaches modelled effective practice to teachers (*Techniques 3—Modelling*) and spent time debriefing lessons with them (*Techniques 4—Feedback).*

Embed practice

Coaches helped teachers plan the use of Collaborative Strategic Reading in future lessons (*Practice 2—Action planning*). Finally, coaches encouraged repeated application of Collaborative Strategic Reading by offering coaching weekly (*Practice 4—Context-specific repetition*).

McMaster et al. (2019)

This programme aimed to improve the writing of children with special educational needs, through Data-Based Instruction—the use of formative assessment to adapt teaching. Developers provided teachers with assessment and decision-making tools alongside instructional resources, such as lesson plans. Teachers attended a series of workshops introducing them to key ideas from the programme, and received ongoing coaching supporting them to apply their learning in the classroom. This programme applied every behaviour change mechanism we were seeking, with one exception (managing cognitive load).

Instil insight

Developers revisited material in two ways (*Insight 2—Revisit material*): first, they assessed mastery of new learning after each taught module, with individualised in-person and online support where necessary. Second, during each coaching session they reviewed the steps in the Data-Based Instruction process, 'to build on previous learning–build mastery, emphasize alignment' (McMaster et al., 2019, p.5).

Motivate goal-directed behaviour

The programme encouraged teachers to set goals in multiple ways (*Motivate 1—Goal setting).* The development team gave teachers how much time to spend on writing interventions each week. Coaches also set goals with teachers, both immediate objectives (during coaching meetings) and long-term goals. Developers provided teachers with an overview of the research of the research supporting specific Data-Based Instruction steps (*Motivate 2—Credible Source*). Each coaching session began by celebrating teachers' successes (*Motivate 3—Praise/reinforce*).

Teach techniques

Developers gave teachers step-by-step guidance (*Techniques 1—Instruction*) in how to implement Data-Based Instruction. They provided extensive opportunities to discuss assessments, lessons, and instructional materials, both during initial training and ongoing coaching sessions (*Techniques 2—Practical social support*). Each step was illustrated to teachers (*Techniques 3—Modelling*) with both videos and live demonstrations. Coaches offered teachers feedback (*Techniques 4—Feedback)* about both their understanding of key ideas and their classroom implementation of Data-Based Instruction. During training, teachers practised (*Techniques 5—Rehearsal)* the steps of Data-Based Instruction, and practised using the tools provided.

Embed practice

Developers provided teachers with checklists (*Practice 1—Prompts/cues*) to prompt their thinking and action in the classroom. They supported teachers to plan their next steps at the end of each coaching session (*Practice 2—Action planning*). The intervention encouraged teachers to examine and reflect upon their own practice (*Practice 3—Self-monitoring*), inviting them to make hypotheses, make changes, document their changes in a decision log, and bring the results to a workshop. Finally, *c*oaches observed teachers' ongoing application of Data-Based Instruction through biweekly meetings (*Practice 4—Context-specific repetition*).

Mechanisms of three programmes with balanced designs

Purpose	Mechanism	Allen et al. (2015)	Boardman et al. (2015)	McMaster et al. (2019)
Instil insight	0.1 Manage cognitive load	✓		
	0.2 Revisit prior learning	✓	✓	✓
Motivate goals	1.1 Goal setting	✓	✓	✓
	9.1 Credible source			✓
	10.4 Praise/reinforce	✓		✓
Teach techniques	4.1 Instruction	✓	✓	✓
	3.2 Practical social support		✓	✓
	6.1 Modelling	✓	✓	✓
	2.2, 2.7 Feedback	✓	✓	✓
	8.1 Rehearsal			✓
Embed practice	7.1 Prompts/cues			✓
	1.4 Action planning		✓	✓
	2.3, 2.4 Self-monitoring	✓		✓
	8.3 Context-specific repetition		✓	✓

Note. Numbers in the mechanism column represent the codes from the Behaviour Change Taxonomy. See notes to Table 3.

Appendix 9: Team, conflict of interest statement, registration

Team

- Sam Sims, UCL Institute of Education—conception of study, writing protocol, screening, theory review, coding, data cleaning, analysis, writing report
- Harry Fletcher-Wood, Ambition Institute—conception of study, writing protocol, screening, theory review, coding, implementation review, writing report
- Alison O'Mara-Eves, UCL Institute of Education—writing protocol, screening, coding, data cleaning, analysis, writing report
- Sarah Cottingham, Ambition Institute—screening, coding, writing report
- Claire Stansfield, UCL Institute of Education—writing protocol, search, review of final manuscript
- Jo Van Herwegen, UCL Institute of Education—writing protocol, writing report
- Jake Anders, UCL Institute of Education—implementing publication bias checks

Conflicts of interest

- Harry Fletcher-Wood and Sarah Cottingham declare that they work for a charity that provide PD to teachers and schools in return for fees.
- Alison O'Mara-Eves, Claire Stansfield, Jo Van Herwegen, Sam Sims, and Jake Anders declare that they work for a university that also provides PD to teachers in return for fees.
- None of the authors are shareholders or otherwise directly financially benefit (beyond their ongoing employment) from their employers' activity providing PD to teachers.
- All authors declare no other conflicts of interest.

Registration

The protocol for this trial was published on the EEF website on 14 January 2021, and can be accessed here:

https://educationendowmentfoundation.org.uk/public/files/Publications/EEF._Systematic_Review_of_Professional_Development._Dr_Sam_Sims._Protocol._.pdf.

[i] Figure II.2 https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf