What are the characteristics of teacher professional development that increase pupil achievement? Protocol for a systematic review

Sam Sims, Harry Fletcher-Wood, Alison O'Mara-Eves, Claire Stansfield, Jo Van Herwegen, Sarah Cottingham, John Higton. **January 2021.**

Table of contents

Background and review 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 (e.g., Allen et al., 2011, 2015), many schools do not have access to such programmes, due to either cost or location. School leaders and teacher educators 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, 2016; Timperley et al., 2007; Walter & Briggs, 2012; Wei et al., 2009; Yoon et al., 2007). Indeed, many of 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, 2020). In particular, existing reviews have no way of distinguishing causally redundant components of interventions from the 'active ingredients' that lead to improved teaching and learning (Sims & Fletcher-Wood, 2020). 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, 2016), with one stating only that professional development is "hard to define by aggregation and generalities" (Opfer and Pedder, 2011, p. 379). A recent review adopted a relatively 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). This is a useful starting point, however, the breadth of this definition seems 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, that 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, respectively, 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 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 – or indeed *not used* – in previous reviews.

This review will employ a systematic search of the literature evaluating teacher PD. This will be used to develop a map of the relevant literature, which will inform the development of the final inclusion criteria. Results will then be extracted from each of the included studies. Crucially, each of the interventions in each of the included studies will also be coded based on the 'mechanisms' they incorporate, defined as "*entities and activities organized in such a way that they are responsible for the phenomenon"* (Illari & Williamson, 2012, p14; see also Sims & Fletcher-Wood, 2019). In the social sciences, mechanisms can be thought of as domain-general empirical regularities related to what motivates individuals, how they learn, and why they act in certain ways. Coding for such mechanisms helps distinguish the causally active from the causally redundant components of the interventions, in a way that previous reviews have not. Meta-analysis and qualitative comparative analysis will then be used to investigate the relationships between (groups of) mechanisms and the impact of the interventions on pupil achievement. The results of this analysis will directly inform the development of recommendations for a subsequent EEF guidance report on the characteristics of effective PD.

Initial conceptual framework

The design of PD can 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 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 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 essential characteristics of that form.

Another term for these 'essential characteristics' is PD *mechanisms.* These were defined in the previous section as "entities and activities organized in such a way that they are responsible for the phenomenon" (Illari & Williamson, 2012, p14). To establish that something is indeed a mechanism requires 'evidence of mechanism', which comes "*from basic research describing fundamental characteristics of human motivation or learning, which hold across diverse contexts"* (Sims & Fletcher Wood, 2020). PD mechanisms can be thought of as the 'active ingredients' of PD, which means that they could not be removed without making it less effective. In that sense, mechanisms are the basic building blocks of PD. One example of a PD mechanism might be to establish new habit of practice, through repetition of a new technique in a realistic classroom setting (Hobbiss et al., 2020).

It follows from the above that PD programmes can be thought of as belonging to the same form if they share the same set of mechanisms. Conversely, PD forms and programmes can in turn 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: provide an observable example of a technique, provide 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 can also now be expressed in terms of identifying effective forms of PD, and their constituent mechanisms.

Having conceptualised PD in this way, how do we theorise it being 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' knowledge/insight, goals (for instruction), techniques/skills, and habits of practice. These two factors in turn are thought to 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 final stage of this review – the review of practice – is intended to probe the importance of school environment, in order to understand how this interacts with the design of the PD. 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 in the top left of Figure 1 ('PD Forms' and 'Changes within the teacher') and their relationship with the top right box (pupil achievement). In particular, we hypothesise that effective PD has to incorporate a set of mechanisms that are able to bring about four important changes (Goodrich, 2021). These are:

1. To instil new, true, and relevant insight e.g., 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 e.g., I will manage the amount of cognitive load on any one of these two systems, at any one time.
3. To develop new techniques that teachers use to put these insights to work e.g., do not read aloud text that is also being presented visually, in order to not to overload the phonological loop.
4. To help embed this new practice e.g., by repeatedly practicing 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 things are missing. If PD brings about the necessary changes to Insight (I), Goals (G), and Techniques (T) but does not embed this in practice (P), then teachers will tend to revert to established ways of working (Hobbiss et al., 2020). This is depicted in row 2 of the table. If PD brings about the necessary changes to I and G, but not the 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, Turner, & Dekkers, 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 even an intention 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 changing all of I, G, T and P, it is much more likely to work.

An important point here is that different mechanisms will be suitable for bringing about change in each column. We previously gave the example (mechanism) of repeatedly practicing a new technique in a realistic environment, in order to help develop a new habit. That might help embed practice (P) but would clearly be unsuited to help instil insight (I). An additional aim of this review is therefore to test the hypothesis that PD incorporating mechanisms theorised to change all of I, G, T and P is more likely to be effective.

Table 1: Theorising how PD fails, using the IGTP model

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

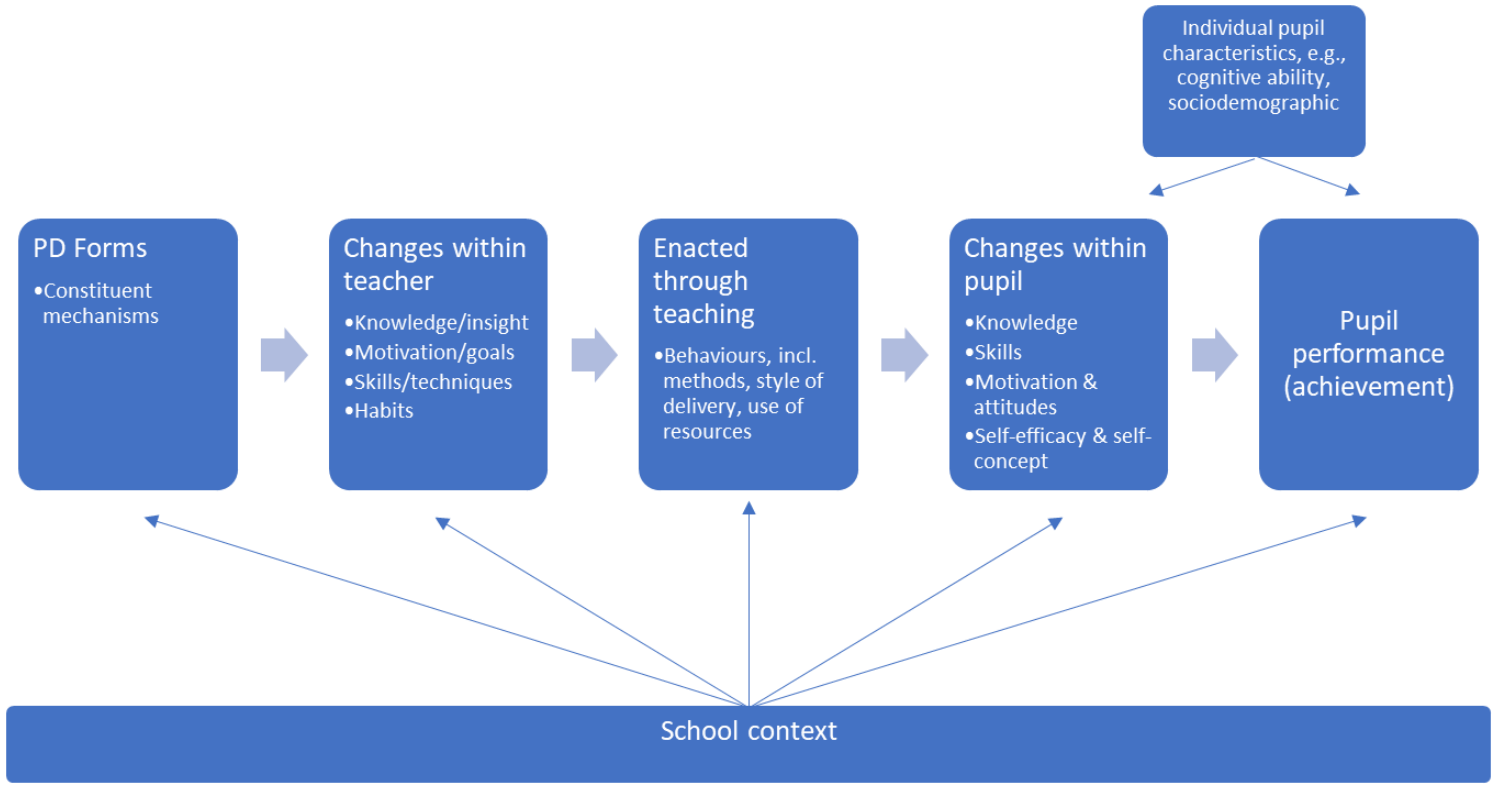


Figure 1. Logic model showing pathway from professional development interventions to pupil achievement

Objectives

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 will address three specific research questions:

1. What are the characteristics of the studies and interventions in the experimental impact evaluation literature on teacher PD?
2. a. Overall, are teacher PD interventions effective at increasing pupil achievement compared to business as usual?
    b. Does this vary based on study characteristics (features of the evaluation not specific to the intervention itself)?
    c. Does this vary by study-level pupil disadvantage?
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 of: instil insights (I); motivate goals (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?

The PICO (Population, Intervention, Comparison, Outcome) framework provides a useful way to further clarify the scope of the project.

- The population of interest is post-initial qualification teachers working in formal settings (e.g. early years centres, schools, colleges), who provide instruction to pupils between the ages of 3 and 18.
- The intervention of interest is teacher PD. We define this as any structured, facilitated activity for teachers intended to increase their teaching ability. We focus here on formal, structured programmes as opposed to informal or unstructured collaboration or support, because, while such support may well play an important role in helping teachers improve, by its nature (unplanned, ad hoc, idiosyncratic), it is extremely difficult to evaluate, and difficult for school leaders or policymakers to promote. Note, however, that our definition does not exclude *formal, structured* approaches to improve peer collaboration or support, such as attempts to encourage greater teacher collaboration between groups (e.g., West et al., 2017) or individual support (e.g., Papay et al., 2016).
- The comparison group 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.

- In general, something can be said to be effective if it achieves its purpose. Teachers' main purpose is to help pupils learn. By extension, teacher professional development is effective if helps teachers increase pupil achievement. The outcome of interest in this review is therefore pupil achievement, as captured by scores on relevant tests.

Methodology

Logic of the review

The review will follow a four-step process:

1. Review of theory. This will help us determine the coding framework to be used in Step 3.
2. Systematic search for empirical studies. This will identify the studies to be used in Step 3.
3. Synthesis of the evidence from the empirical studies. This will produce the findings that will be explored in Step 4.
4. Review of practice. This will explore barriers and enablers for schools accessing PD that is aligned with the findings from Step 3.

This protocol follows the standard EEF protocol structure, rather than reflecting these four steps. However, it is hoped that the above provides a sense of the underlying logic of our approach. For information about our review of theory see pages 15-16, for information about our search for empirical studies see pages 9-14, for information about our synthesis of evidence from the empirical studies see pages 16-22, and for more information about the review of practice see pages 24 onwards.

Information management

All records of research identified by searches will be 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).

Inclusion and exclusion criteria for the review

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

1. Design: we will include randomised controlled trials and exclude all other evaluation designs. The justification for this is that RCTs are the only design that provide unbiased causal impact estimates – which is essential for our synthesis approach. Our initial searches suggest that there will easily be enough RCTs to support our synthesis approach, so there is no need to include designs that provide potentially biased impact estimates. We will also utilise any process evaluations for such RCTs by capturing any that appear in our searches and by searching the websites for the two organisations that commission such process evaluations

for all of their trials: EEF and National Center for Education Evaluation and Regional Assistance (NCEE).

2. Publication year: we will restrict our search to studies published during or after 2002. It is necessary to specify some restriction on the earliest date for our searches in order to ensure that we can complete the project by the deadline and within the available budget. We have chosen 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 will be the 2nd of November 2020.
3. Language: include studies published in English; exclude studies published in languages other than English. This is necessary due to resource constraints.
4. Geography: include interventions implemented in OECD countries.[1]
5. Intervention: include studies that are teacher PD programmes, as defined above. We will also exclude interventions that train teachers in the delivery of small-group or one-to-one tuition. This is necessary, given that our focus is on the relationship between the presence of 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). If we were to include evaluations of PD that also incorporated small-group tuition, then the impact of the small-group aspect would confound the relationship between the PD characteristics and the impact of the PD. We will also exclude interventions that aim only to briefly familiarise teachers with educational technology or curriculum materials.
6. Intervention population: Include studies of 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.
7. Outcomes: Include studies that measure pupil achievement using tests, in any school curriculum subject (e.g., GCSE exams in England). Tests that are administered verbally by teachers, or marked by teachers are not excluded (e.g., The Research-Based Elementary Math Assessment; Clements, Sarama, & Liu, 2008). 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 Ealy Years Foundation Stage; DfE, 2020).
8. Comparison group: Include studies that provide no PD, business as usual PD, or waitlist intervention for the control group.
9. Reporting: Include studies reported in journal paper or working paper format. Include doctoral theses that can be obtained via current UCL subscriptions. Exclude conference

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

papers or extended abstracts - the rationale for this is that short versions do not contain enough information to assess quality or to extract information about intervention components. Both are essential for answering our research question. Exclude master's thesis on the basis that they are generally of lower quality and often harder to access.

Search strategy for identification of studies

- We will search the following databases: EconLit; Education Resources Information Center (ERIC); Education Abstracts; Educational Administration Abstracts; Google Scholar; ProQuest Dissertations & Theses; PsycINFO; Teacher Reference Center.
- We will also search 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, 2016; 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). We will identify such reviews by searching for them. This will be undertaken on Microsoft Academic Search, within the EPPI-Reviewer interface.
- Browsing and searching websites of education repositories: Center for Coordinated Education MRDC publications; CUREE - Centre for the use of evidence and research in education; Digital Education Resource Archive; Education Endowment Foundation; EIPEE search portal; EPPI-Centre database of education research; Institute of Education Studies What Works Clearinghouse; Nuffield Foundation.
- Reference-checking and forward citation searching of studies will be undertaken for the studies that remain in the review following assessment of study quality. This will be undertaken on Microsoft Academic Search, within the EPPI-Reviewer interface.
- The search process will be designed and implemented by Claire Stansfield, who is an experienced information scientist. Bibliographic information will be managed in EPPI Reviewer software.
- The database search incorporates 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') 3) randomised controlled trials or terms to capture process evaluations of randomised controlled trials. Terms and phrases will be used to search the title, abstract and controlled vocabulary fields of the databases. The searches will be limited to English language and use some terms to exclude on preservice teachers, higher education teachers, and adult education teachers. Appendix A of this document contains a sample search strategy for the ERIC database.

Selection of studies

Screening for inclusion in the review will occur in two stages. First, the inclusion criteria will be applied to titles and abstracts (T&A). The criteria will be piloted on a sample of studies before being

applied. In the EPPI-Reviewer software, we will apply screening prioritisation to improve the efficiency of the T&A screening process. The records will be ranked by relevance so that we see examples of includable studies early on (see section "Prioritised screening" for more details).

Once piloting is completed, we will begin screening the titles and abstracts to narrow down the pool. Partial double screening will be used. This will consist of double screening 15% of records and then switching to single screening if 95% agreement has been achieved; if not, then screening will continue in allocations of 100 records until at least 95% agreement is attained for that allocation. The reviewers will regularly discuss screening to ensure consistency in the way that studies are being included and excluded.

Full reports will be obtained for those studies that appear to meet the criteria or where there is insufficient information to be certain. In the second screening stage, those marked for inclusion will be retrieved and screened on the basis of the full-text article. Those that pass the inclusion criteria on the basis of full-text screening will be included in the review. The results of this process will be documented in a PRISMA flow chart (Moher et al., 2009).

If multiple papers report the same study, then we will retrieve all versions and assign one as the 'master' document. Master documents will usually be the most recent or detailed journal article, with e.g. associated dissertations treated as linked studies. If information is not reported in the master document, then the linked documents will be checked to see if they contain the relevant information.

If one document reports results for more than one teacher PD intervention in addition to a comparator condition, then we will use 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 is not possible to clearly distinguish more and less intensive versions, we will pick a version at random. Any less intense interventions will be included in the map but will not contribute an effect size to the analyses to avoid dependencies in the dataset.

Prioritised screening

We will use text mining built into the 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 is important in order to complete the review by the deadline.

The approach to prioritised screening used in EPPI-Reviewer is known as 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. Priority screening using EPPI-Reviewer can provide efficiency gains - in a retrospective evaluation of 6 systematic reviews, EPPI-Reviewer's priority screening algorithm estimated potential reductions in screening of between 9% and 60% (Tsou et al., 2020). We will use the built-in classifiers in EPPI-Reviewer to exclude systematic reviews from the prioritised screening. This is important in order to avoid confusing the prioritisation algorithm.

We will use two indicators when deciding when to stop screening the ranked list. The first is comparison with a predicted number of eligible studies based on the calculation of a *baseline inclusion rate* (Shemilt et al., 2014), which is established as follows. First, a random sample of 400 citations will be screened to provide an initial predicted inclusion rate. Second, a power calculation will be 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 2% at 95% conference interval. The power calculation will indicate how many records would need to be screened to confidently establish the baseline inclusion rate given these parameters. Third, once this quantity of references has been screened, we will calculate the baseline inclusion rate. 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 of all relevant records. The second 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. Combined, these two indicators will help ensure we do not stop manual screening before almost all relevant studies have been identified.

Evidence mapping

Recall that our first research question is: What are the characteristics of the studies and interventions in the experimental impact evaluation literature on teacher PD? This will be explored through a systematic evidence map. Evidence maps involve a systematic search of a broad field, followed by the coding of the studies identified based on study characteristics. We will present the results in a user-friendly tabular format (Miake-Lye et al., 2016). We will then code the studies that have met the full-text inclusion criteria using the following codes:

Full APA reference	Alphanumeric
Publication year	Numeric
Age group / phase	1 – Early years / pre-kindergarten 2 – Primary/elementary 3 – Secondary/middle/high school/Sixth Form
Location of intervention	Two-letter country code, e.g. US, EN
Academic subject targeted	1 – English/native language/speech/literacy 2 – Maths 3 – Science 4 – Other subjects 5 – Generic (cross-curricular)
Test type	1 – High-stakes, standardised test 2 – Low-stakes, standardised test 3 – Test developed or adapted by researchers or teachers
Number of units randomised	Numeric
Process evaluation reported?	0 – No 1 – Yes
PD intervention also includes curricular reform component	0 – No 1 – Yes
PD intervention also includes new educational technology component	0 – No 1 – Yes

The data extraction tool for the map will be piloted to check whether additional guidance notes or definitions are needed. We will then conduct independent double data extraction on 20 randomly selected studies. Two reviewers will independently extract study characteristics and numerical outcome data from studies meeting the eligibility criteria. In agreement meetings, the reviewers will resolve discrepancies by discussion, and seek input from a third reviewer if the disagreement cannot be resolved. In some cases, the data extraction tool or guidance may need to be refined in order to ensure consistent understanding between reviewers. Each record in the 20-study sample will need to be resolved (i.e., full agreement on all assigned codes in each study) before reviewers can progress to independent extraction. If, after completing the 20-study sample, concerns about interpretation of the extraction tool remain, then the team will meet to decide whether more double extraction is required, or whether further refinement to the tool will clarify the differences in interpretation.

After the map is produced, we will hold a consultation meeting with EEF to determine which studies to take through to data synthesis and assessment of study quality. We will include as many studies

as possible in the review, consistent with being able to complete the review within the time and budget available. If it is necessary to exclude some studies then we will exclude those that are methodologically weakest i.e., those with low stakes test score outcome measures.

Review of theory

Our review of theory is intended to develop a coding frame that will be used to capture the mechanisms from each intervention. This information will then be used in the synthesis stage.

Our review of theory will begin with the Michie et al. (2013) taxonomy of domain-general practice-change techniques, which was developed through an exhaustive expert consultation exercise. A sample of this taxonomy can be found in Appendix B. This represents a longlist of candidate mechanisms, potentially relevant to understanding the: motivate goals (G), develop techniques (T), and embed practice (P) part of our theory of change (Figure 1, Table 1). Two members of the team will code up twenty randomly chosen studies from our evidence map in order to filter out the subset of such mechanisms that are relevant to the teacher PD setting. Codes that are not used by either researcher will be discarded. For each potential mechanism on this shortlist, we will then carefully search through the associated empirical literature, to check whether there is supporting 'evidence of mechanism' – *basic research describing fundamental characteristics of human motivation or learning, which hold across diverse contexts* (Sims & Fletcher-Wood, 2020). In order to look for such evidence, we will use Google scholar to conduct targeted searches. In particular, we will look for evidence from 1) a range of academic disciplines, 2) lab-based studies, 3) field studies in a range of settings (inside and outside schools), 4) biological/neuroscientific evidence. Where possible, we will rely on published reviews bringing together this type of evidence for each mechanism. Where such supporting evidence can be found, the mechanism will be entered into our coding frame. Where such supporting evidence cannot be found, the potential mechanism will be discarded.

While the Michie et al. (2013) taxonomy provides a good starting point for our review of theory (and coding framework), it does not adequately address certain parts of our theory of change (Figure 1, Table 1). In particular, it does not address the instil insight (I) part of our theory of change. By this, we mean that it does not include mechanisms that address how people learn declarative knowledge – which is an important part of many teacher PD programmes. An example of a candidate mechanism here might be the well-established finding that people are more likely to learn new knowledge if it can be related to pre-existing knowledge (Brown et al., 2014). In order to address this shortcoming, we will search in the cognitive psychology literature to identify a set of further mechanisms related to the way in which people assimilate new knowledge. Again, we will use Google scholar to conduct targeted searches. In particular, we will look for evidence from 1) a range of academic disciplines, 2) lab-based studies, 3) field studies in a range of settings (inside and outside schools), 4) biological/neuroscientific evidence. Again, where possible, we will rely on published reviews bringing together this type of evidence for each mechanism. Where such supporting

evidence can be found, the mechanism will be entered into our coding frame. Where such supporting evidence cannot be found, the potential mechanism will be discarded.

It should be noted that - by contrast with our typology of mechanisms - our conceptualisation of PD forms will not emerge from our review of theory. Rather, this will in part be derived empirically from our data using the truth table that we will construct in our qualitative comparative analysis. For an example of this approach, see Fiss (2011). Even so, the truth table output is likely to need refining in order to develop our final taxonomy of forms. This is because some interventions will differ on the presence/absence of just a single mechanism, and we may therefore want to create a more parsimonious solution by grouping certain rows of the truth table. We will be guided in this process by the three heuristics for configurational theorising outlined by Furnari et al. (2020).

Coding studies for the synthesis

We will then use both the main evaluation reports and the associated process evaluations to code studies/interventions that have met the agreed final inclusion criteria, for the following:

- Mechanisms (including those from the Michie et al. [2013] taxonomy), incorporated in each intervention (see Appendix B for a sample).
- Kennedy's (2016) four 'persistent problems' of teaching: portraying curriculum; containing behaviour; enlisting participation; exposing student thinking. If a particular intervention has multiple foci then it will be coded as such.
- Broad area of focus of the PD: science of learning / cognitive science; discovery / problem-based / inquiry teaching methods; formative assessment; supporting curricular change; supporting new technology in the classroom; data-driven instruction. If a particular intervention has multiple foci then it will be coded as such.
- Teaching experience of those receiving and providing the intervention (in years). Our hypothesis here is that more experienced teachers might benefit from different forms of PD (Kalyuga, 2007).
- Length or "dosage" (average number of intended hours per participating teacher).
- Effect size for achievement outcomes, or other statistics necessary to calculate it (see below).
- Effect size for teacher practice change outcomes, or other statistics necessary to calculate it.
- Level of pupil disadvantage (% eligible for free or reduced-price lunch) at the study level.

We will double code a random sample of 20% of the interventions. As with the data extraction for the map, full agreement on the double-coded sample will need to be achieved before embarking on independent single coding. If, after completing the 20% sample, concerns remain, then the team will meet to decide whether more double extraction is required, or whether further refinement to the tool will clarify the differences in interpretation.

Assessment of risks of bias

We will also code the following details for each study, which will be useful in identifying the potential for biased effect size estimates. This list was developed in part from the team's knowledge and part by consulting the Cochrane Risk of Bias tool (RoB 2; Higgins et al., 2019). We did not adopt the full RoB2 or another off-the-shelf tool for two reasons. First, many existing tools rely on opaque judgements such as 'low', 'medium' and 'high'. Second, many existing tools contain criteria that are inappropriate for experimental studies in this topic area, such as 'allocation concealment'. Instead, we use the following criteria on the grounds that they are transparent and appropriate in our setting:

- Attrition: outcome data missing for any reason by allocated group (% reported separately for intervention and comparator groups, or 'not reported')
- Non-compliance with the intervention (% attendance, or 'not reported')
- Whether the analysis was pre-specified (0/1, or 'not reported')
- Whether the experiment has been analysed at the same level as it was randomised (0/1)
- Number of units randomised (sample size)

We will not code for baseline equivalence, since this misunderstands the nature of experimental research (Senn, 2013). In addition, we will not code for blinding since this is infeasible in almost all education research.

Effect size calculation

Effect sizes will be calculated in the EPPI-Reviewer software. For our student attainment outcome, the data will be continuous test scores and we will therefore calculate standardised mean difference effect sizes using Cohen's *d*. These will be typically calculated from means and standard deviations (or SEs or CIs). In the absence of those statistics, we will use (in decreasing order of preference) t- or F-statistics, or p-values. Any study results that can be converted to an effect size will be used (see full list of formulae in Lipsey & Wilson, 2001). Cohen's *d* has been selected as the review team know that many studies in this literature report this as their effect size metric. We can correct Cohen's *d* for small bias in studies with less than 50 participants using Hedges' *g* (Hedges, 1981).

We will only use the effect size for the overall experimental group, rather than for specific subgroups or cohorts. If multiple attainment (or teacher practice change) outcomes are reported and a primary outcome measure is specified, then we will use the primary outcome measure to calculate the effect size. If multiple attainment (or teacher practice change) outcomes are reported and no primary outcome is specified, we will calculate multiple effect sizes and handle this using robust variance estimation to account for dependencies (Tanner-Smith & Tipton, 2013). For studies in which

randomisation occurs at the group level (e.g., classes or schools) rather than the individual teacher level, we will use the White and Thomas (2005) adjustment for clustering.

Dealing with missing data

Missing (unreported) outcome data will be recorded. Where values can be calculated or imputed form the available data (e.g., by combining separate results from two subgroups to yield an outcome for the total sample), then we will do so in preference to treating the effect as missing. We will keep a record of any such calculations/imputations and then conduct a sensitivity analysis to determine if this process might lead to a bias in the observed effects.

Data synthesis

Different synthesis methods will be used to address research questions 2a-3c.

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

Put simply, this research question aims to explore whether PD interventions generally 'work'. This will be assessed through a random effects statistical meta-analysis, with effect sizes as the unit of analysis. We will use an inverse variance weighting approach to account for sample size (Deeks, Higgins, & Altman, 2020). The results will be visually displayed in a forest plot. Given the aim to give an overall picture of the evidence base, this analysis will include all studies that report sufficient data to calculate an effect size for student achievement. If there is statistically significant heterogeneity as indicated by the measures indicated in the section 'Investigation of heterogeneity', then we will not report the overall mean effect size (i.e., the 'diamond' in the forest plot), but just include the effect sizes and confidence intervals for each study and the heterogeneity statistics in the forest plot. Irrespective of whether there is statistically significant heterogeneity, we will conduct analyses to explore possible sources of variation in the studies as determined in the research questions below.

Research question 2b: Does this vary based on study characteristics (features of the evaluation not specific to the intervention itself)?

This research question seeks to explore whether aspects of the studies, including the way in which the evaluation is designed, and the analyses are conducted, explain variation in the observed effect sizes. This is to assess the extent to which the estimated effect sizes might show heterogeneity due to different design choices, rather than to in the intervention or population characteristics. For example, different types of achievement tests tend to produce different effect sizes. Note that this is not an issue of bias, but rather one of design.

Since all our study characteristics are categorical variables, we will use subgroup meta-analyses with inverse variance weighting to explore whether there is statistical variation between effect sizes

grouped according to key study characteristics. We anticipate the following to be of interest, but these may be refined as we interact with the studies and either identify new important issues to consider, or discover insufficient data to conduct a given subgroup analysis:

i. Publication type (journal, working paper, doctoral thesis)
ii. Form of the comparison group (no PD, business as usual PD, waitlist, other)
iii. Randomisation to conditions (individual or group randomisation)
iv. Whether the analysis was pre-specified (yes, no, or 'not reported')
v. The type of test used to measure the outcome (High-stakes standardised; Low-stakes standardised; Test developed, adapted, or marked by researchers or teachers)

As with the subgroup analyses, this list of analyses may be refined once we have a full view of the evidence base. Any changes to the analysis plan will be reported in the final report as a deviation from the protocol.

Research question 2c) Does this vary by study-level pupil disadvantage?

This research question aims to explore variation in the observed effect sizes due to differences in the level of pupil disadvantage in the populations targeted by the interventions. Given that the unit of analysis in the review is study-level data (as opposed to individual participant data), we cannot analyse variation between sub-populations *within* a study. Even across studies, there are difficulties in that we expect pupil disadvantage to be measured inconsistently. However, at the request of the funder we will investigate this by recording the proportion of pupil eligible for free or reduced-price lunch (which is a metric common to the US and UK setting) and then looking for heterogeneity using subgroup meta-analyses.

Research question 3a) Which forms of PD (clusters of mechanisms) do we observe in the literature?

We will identify the forms of PD present in the experimental literature by using some of the tools of qualitative comparative analysis (QCA). QCA is predicated upon set theory (Ragin, 2008) and, when applied to systematic reviews, it is a research synthesis method that allows identification of configurations of 'conditions' that are associated with successful outcomes (Thomas, O'Mara-Eves, Brunton, 2014). For introductions to QCA, see Rihoux & De Meur (2012), Thiem (2017), and Thomann and Maggetti (2020). Our goal in research question 3a is restricted to identifying the configurations of mechanisms that co-occur in interventions in the PD literature.

For the purposes of the analysis, the mechanisms will be treated as a crisp set. That is, they will be coded as either present or absent in the intervention. Using this code set, we will generate a truth table, which shows the number of cases (interventions) that display each of the possible unique combinations of conditions (mechanisms) and the outcome for each combination (Thomas, O'Mara-Eves, Brunton, 2014).

We will then investigate the truth table, considering any logical remainders, and applying Boolean minimisation. This helps purge causally redundant components from the forms (Baumgartner, 2015). Doing so requires us to dichotomise our outcome measure to make it suitable for crisp-set analysis. There is currently no generally agreed upon approach for categorising an intervention as successful, as it depends on the context and setting under study. The cut-off must therefore be determined on a review-by-review basis, with reference to theory and empirical evidence in the substantive domain (see Ragin, 2008).

We will draw on two sources of information to develop and justify our cut-off: considerations about what comprises a substantively meaningful increase in learning and the size of the effect relative to what we know about interventions in this domain. With respect to the first, we will anchor our intuition for a substantively meaningful effect size based on the additional learning that would be necessary for an average student to overtake one other pupil in a classroom of average size. Average class sizes in England vary between 22 (secondary) and 27 (primary) (DfE, 2019). The increase in learning required to move the median pupil ahead of one other pupil in a class of this size is equivalent to an effect size of 0.08-0.1.[i] Our second source of information for determining our cut-off comes from a meta-review of experimental effect sizes in education settings (Kraft, 2020). This study summarised 750 experimental effect sizes and found a median effect size in studies using standardised tests of 0.1 (Math) and 0.1 (Reading). We therefore adopt 0.1 as our cut-off for dichotomising our outcome, which can be interpreted as either 'the smallest substantively meaningful effect' or 'the average effect that might be expected in education'. While we acknowledge that any decision about a cut-off for 'effective' will contain some element of arbitrariness, we believe out 0.1 cut-off is defensible and contextually grounded. We will also check for sensitivity of these results by varying our cut-off for effectiveness between 0.08 and 0.12. In addition, our use of meta-analysis (research question 3b) does not rely on this assumption, which mitigates the extent to which our overall conclusions will be sensitive to this assumption.

After this process, we may still be left with configurations that differ only very marginally on e.g. one or two conditions. We will therefore combine very similar configurations, drawing on the 'three heuristics for configurational theorising' outlined by Furnari et al. (2020). The end result of this will be an empirical taxonomy of PD forms defined as combinations of mechanisms. This will provide a complete picture of the frequency with which different forms of PD appear in the experimental PD literature.

Research question 3b) Are forms more likely to be effective when they incorporate mechanisms addressing all four of: instil insights (I); motivate goals (G); develop techniques (T); and embed practice (P)?

We will also use tools from QCA to address question 3b. QCA aims to identify necessary and sufficient conditions (defined as Boolean 'difference-makers' within the sample) in order to identify the 'active ingredients' in interventions (Brunton, O'Mara-Eves, Thomas, 2014). QCA differs from

meta-analytic approaches in that it takes a case-based (rather than variable-based) approach. This allows multiple observed characteristics of an intervention to be studied *in combination* in small-medium sample size settings.

In order to answer question 3b, we will aggregate the mechanisms in each intervention to the level of the IGTP framework (see Table 1). More precisely, each intervention will be coded as 1 in the I, G, T, or P columns if it includes at least one mechanism addressing that column. The exact mapping between the mechanisms and the IGTP columns will be determined in our review of theory and incorporated in our coding framework. In general, however, we will associate a mechanism with the furthest left column in the I-G-T-P framework for which we theorise it to prevent the 'failure' of the PD. For example, the mechanism 'prompt self-praise' would not (in theory) prevent PD failing to provide insight (the leftmost column). However, it would (in theory) help prevent PD failing to motivate new goal directed behaviour (the second leftmost column). Hence, 'prompt self-praise' would be mapped onto the G column. Prompting self-praise might also support the subsequent embedding of practice (the rightmost column) but would not (in theory) achieve this on its own. It would therefore not be entered in the P column. This is necessarily a theoretical exercise, in which we will be drawing on the existing literature. It should also be remembered that the framework will immediately be subjected to empirical tests.

The resulting aggregation will allow us to conduct a crisp-set QCA in order to test the necessity of PD interventions incorporating mechanisms that support all four of the changes set out in the IGTP framework. Our approach will follow standard crisp set QCA methods (Rihoux & De Meur, 2012). Any 'contradictory' configurations will be investigated by checking whether different mechanisms within a given I/G/T/P column can account for the contradictions. For example, we will check if the consistency in a row substantially increases when we exclude interventions that were only coded 1 for 'G' due to a single specific mechanism. Clearly, this particular part of the analysis will be exploratory in nature, aiming to develop, rather than test, theory.

Research question 3c) Which forms are associated with the largest effects on teacher practice and pupil achievement?

Our answer to question 3a will provide us with an empirical taxonomy of forms (defined as combinations of mechanisms). We will code each intervention based on its membership in this taxonomy. This will then allow us to test for variations in effect size estimates for pupil achievement and teacher practice across different forms. This will be done using sub-group meta-analysis with inverse variance weighting (see Harris et al. [2019] for a similar approach). We will also use the meta-analysis to look for interactions between forms and: 1) the area of focus for the PD, defined using Kennedy's (2016) four 'persistent problems' of teaching, 2) content of the PD e.g. formative assessment 3) teaching experience of the teachers and teacher trainers, 4) intervention "dosage", and 5) proportion of pupils eligible for free or reduced-price lunch.

Investigation of heterogeneity

We will calculate the three most popular homogeneity tests/statistics to determine the presence of statistical heterogeneity: Q-test, I^2 and tau-squared. If the Q-test result is statistically significant ($p < .05$), then this will be an indicator of significant heterogeneity (Lipsey & Wilson, 2001). If the I^2 exceeds 75%, then this will be considered 'considerable heterogeneity' (Section 10.10.2 in Deeks et al., 2020) and no overall mean effect size will be reported for the given analysis (although a forest plot of effect sizes will be provided). Irrespective of statistical heterogeneity, we anticipate that there will be some conceptual/characteristic heterogeneity that will be useful to explore. In particular, we will use sub-group meta-analysis to explore whether effect sizes vary based on whether the PD interventions also incorporate either curricular reform or the introduction of new educational technology.

Sensitivity analysis and publication bias (risk of bias checks)

We will conduct sensitivity analyses to check whether the results vary if we exclude effect sizes based on a number of variables that are relevant to whether RCTs are at risk of bias.

- Studies with high levels of non-compliance (top ranked 5% of interventions on teacher non-attendance)
- High levels of attrition / missing data (defined using the NCEE 'cautious' standards[ii])
- Number of units randomised (<50)

We will also test for publication bias using weight functions (Vevea & Hedges, 1995), trim-and-fill (Duval & Tweedie, 2002) and p-curve methods (Simonsohn, Nelson, & Simmons, 2014). The justification for using three methods is that they each rely on different assumptions.

Reporting

Findings from the review will be presented in a report, to be hosted on the EEF website, summarising the results of all analysis described in this protocol. We will use the EEF reporting template structure, which will include presenting the implications of our findings for policy, practice and research. The findings will subsequently be used by the EEF to develop guidance on effective teacher PD.

Personnel

- Dr Sam Sims is a Lecturer at UCL Institute of Education. He has published on subjects including effective teacher PD and effective teaching. Sam has conducted experimental and non-experimental evaluations of teacher PD interventions, has published journal articles on the

conduct and interpretation of randomised controlled trials in education, and lectures on policy evaluation methods at UCL. Sam will lead overall management of the project and will work on analysis and write-up of Stage 1, 2 and 3.

- Dr Jo Van Herwegen is an Associate Professor in Developmental Psychology at UCL Institute of Education. Jo is an expert on systematic reviews, having recently published a completed systematic review (Van Herwegen & Simms, 2020), and currently working on two more. Jo is experienced in writing systematic review protocols and reporting systematic review findings in line with PRISMA standards. She will advise on systematic review methodology throughout the project.
- Dr Alison O'Mara Eves is an Associate Professor and Associate Director of the EPPI-Centre, specialising in systematic review methods and their application to diverse topics including education and public health. Alison is a frequent contributor to peer reviewed journals focused on systematic review methods, including *Research Synthesis Methods* and *Systematic Reviews* and teaches *Approaches to Systematic Review Synthesis* on the IOE's MSc Social Policy and Social Research. Alison has completed over 40 systematic reviews and evidence syntheses, including 10 in the field of education (e.g., Caird et al., 2014; Kneale, O'Mara-Eves, Rees, & Thomas, 2020). She has published original research on methods for study identification in systematic reviews (Miwa, Thomas, O'Mara-Eves, & Ananiadou, 2014; O'Mara-Eves et al., 2014; O'Mara-Eves et al., 2015; Shemilt et al., 2015) and published on methods of synthesis in education research (O'Mara-Eves & Thomas, 2016). Alison also has extensive experience using theory and conceptual framework to integrate large bodies of research (e.g., O'Mara-Eves et al., 2013).
- Dr Claire Stansfield is an Information Scientist and Senior Research Fellow of the EPPI-Centre. Claire has led the search phase of multiple systematic reviews for the UK policy-making context, for example systematic maps (Burchett et al., 2020; Kwan et al., 2020), and authored publications on systematic searching for broad-based topics (Stansfield 2019, Brunton et al., 2017), published research on methods for searching in systematic reviews (Stansfield et al., 2017; Stansfield et al., 2016), and collaborated with other information specialists internationally on systematic searching approaches (Rader et al., 2014), as well as supporting a range of systematic reviews on diverse topic areas.
- Harry Fletcher-Wood is an Associate Dean at Ambition Institute, where he leads a programme on how to design and lead PD. He is a PhD candidate at King's College London and author of two books about teaching. Harry is co-author of a recently published journal article on how to identify the characteristics of effective teacher PD and is also co-author of a recent meta-analysis of the overall effect of PD on pupil achievement (Fletcher-Wood & Zuccollo, 2020). He will work across Stage 1, 2 and 3 of the project.
- Sarah Cottingham is a Learning Design Fellow at Ambition Institute, where she specialises in designing PD that leads to sustained changes in teacher practice. Formerly, she worked as an

English teacher in London and then as a leader of PD at Teach First. Sarah is currently studying for a Master's in Educational Neuroscience at Birkbeck and the Institute of Education. She will work on Stage 2 and 3 of the project.

- John Higton is Director of Research at CFE Research. John has worked in social research for nineteen years, specialising in educational across all phases. Since joining CFE, John has led a number of relevant DfE studies, such as the Evaluation of T Level Professional Development programme, CFE's current evaluations on the Teacher Student Loan Reimbursement and Phased Maths Bursaries schemes, the Evaluation of STEM and MFL Teacher Recruitment and Retention, and our Evaluation of the Employer Ownership Pilot. John directs many of CFE's studies which involve complex evaluation of teaching and leadership programmes, and nearly all of CFE's behavioural research. John will lead on Stage 4 of the project, working with other members of the CfE research team.

Review of practice

John Higton and the team from CfE Research will conduct primary research to explore how a range of schools, colleges and PD providers currently deliver PD, as well as any challenges or barriers they face in providing more effective PD (as identified in Stage 3). Data will be collected using in-depth interviews with those delivering and using PD across a range of education settings. In keeping with previous stages of the review, the collection and analysis of this data in this stage will be structured around the behaviour change framework set out in Michie et al. (2014).

It should be noted that we are deliberately focusing Stage 4 on investigating why schools might find it difficult to improve PD, on the basis that this is highly relevant to the development of subsequent EEF guidance. We have decided not to try to capture a representative picture of current practice on the grounds that it would be hard to acquire representative data within the available budget and during a pandemic. Furthermore, we believe this would be less useful for informing EEF guidance than a thorough exploration of the barriers to improved provision.

Sampling and recruitment

Since our aim is to explore the barriers to improved PD across a range of settings, a purposive sampling approach will be adopted. Table 1 below shows our planned purposive sample, reflecting the inclusion criteria of the systematic review as a whole.

Table 2: Purposive sampling criteria

Key stage	Pupils' ages	Educational phase	Interviews (Organisations)	Estimated participants	Providers interviews
Early years	<5	Primary	12 interviews (in 8 schools)	20	4
KS1	5-7				
KS2	7-11				
KS3	11-14	Secondary (some including sixth forms)	20 (in 10)	36	4
KS4	14-16				
Post-16	16-18	Sixth form / FE colleges	8 (in 4)	12	2
Total interviews / interviewees			40 (in 22)	68	10

Schools and colleges will be recruited by a specialist agency, using the publicly available sampling frame Get Information About Schools. This includes a school/college contact number, allowing telephone-based recruitment using a script designed by CFE. CFE and recruitment partners are compliant with ISO 27001 and will design all GDPR materials and processes to ensure compliance with the Data Protection Act. All recruitment support materials will be signed off by EEF before use. Within schools, we will establish whether PD differs across key stages or subjects and will arrange interviews to cover the full range of practice. Interviews with school staff will then be used to identify providers of PD, which we will then approach for further interviews.

Discussion guides and fieldwork

We propose to develop two discussion guides – one for schools and colleges, the other for PD providers. The schools guide would be tailored to suit particular phases or subjects, as appropriate. Both guides will cover how PD is currently designed and how this differs from the characteristics of effective PD identified in the systematic review. The schools guide will then explore a range of challenges that schools might face in achieving greater alignment with the characteristics of effective PD, structured around the capability-motivation-opportunity framework (Michie et al., 2014). For example, schools might not currently have the skills (capability), belief in the value of PD

(motivation), or time (opportunity) to deliver better PD. The PD provider discussion guide will use the same framework but tailored to commercial PD providers. Where schools or PD providers are already delivering PD that is well aligned with the findings from the systematic review, we will discuss the value of this form of PD and the challenges around good implementation.

All fieldwork will be conducted using phone/online conference call technology unless face-to-face is preferred by interviewees. Online call software will be the default option as it allows interviewers to present visual stimulus material. In the light of Covid-19, this will future-proof our design compared to offering face-to-face methods, should the pandemic continue into Spring 2021. We plan to conduct 50 separate interviews lasting an hour on average, as per Table 2.

Coding and analysis

With the permission of interviewees, all interviews will be recorded and transcribed (included in the quoted costs) to facilitate coding and improve analysis. All interviews will then be iteratively coded, which involves creating a first pass of initial codes based on the first few interviews, then reviewing the emerging code frame to ensure it is correctly classifying the views of representatives as more data becomes available. Initial code frames will be shared with EEF for sign-off and will be based on the findings from earlier stages. After this, all other transcripts will be coded, with the framework continuing to be refined as more interviews are processed. The resulting code book will be used to structure our analysis, drawing out the main themes to help inform the development of subsequent EEF guidance.

Review of implementation

Effective implementation makes a substantial difference to the outcomes a programme achieves: one meta-analysis suggests that effective implementation at least doubles its effect (Durlak & DuPre, 2008). We will complement this primary research with the examination of the Implementation and Process Evaluations of a randomly chosen sample of studies included in Stage 3 of the review. This section of the review will seek to identify the most common barriers and supports to teacher change encountered in professional development programmes. While there are many aspects of implementation research which should be assessed through an Implementation and Process Evaluation our focus will be on fidelity (the extent to which teachers adhere to the intended intervention), responsiveness (the extent to which they respond to the intervention) and adaptation (the ways in which they adapt the intervention to suit their needs and preferences; Humphrey et al., 2016). (Other elements of implementation research, such as programme differentiation, will be considered only in as much as they influence fidelity, responsiveness, and adaptation). We adopt this focus to complement the work conducted in Stages 1-3: if we are able to say more about what

makes professional development effective, what would be needed to ensure that teachers benefit from these forms of professional development?

Sampling

We will randomly select twenty studies from the pool of studies included in Stage 3 review, which also have an Implementation and Process Evaluation (IPE). While we note the potential that these are not representative of the overall group studies in Stage 3, we are unable to avoid this situation given that we cannot conduct this analysis on studies which lack an IPE. We will report the proportion of studies in Stage 3 which do/do not include an IPE.

Coding and analysis

We will conduct qualitative content analysis to identify the critical barriers and supports for professional development (Graneheim and Lundman, 2004). We will code for:

- Factors supporting teacher responsiveness, fidelity and (or) positive adaptations
- Factors hindering teacher responsiveness and fidelity, or encouraging negative adaptations

And between features of the intervention, the support system, and the school context (Domitrovich et al., 2008). Since the process of coding and categorisation is intended to respond to the themes which emerge from the content (Cohen, Manion, & Morrison, 2018, p. 677), will not pre-specify our coding as part of this protocol. We will double-code the first five studies, reporting reliability statistics and resolving differences by discussion. We will then continue to double-code until acceptable reliability is reached.

Conflicts of interest

The work described in this protocol is being undertaken by researchers at the UCL Institute of Education and Ambition Institute and funded by the EEF. The views expressed are those of the authors and not necessarily those of the EEF.

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 and Sam Sims 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

Once finalised, this protocol will be registered on the Open Science Framework and on the EEF website.

Timeline

Dates	Activity	Staff responsible/ leading
September 2020	Research team appointed	EEF
November 2020	Protocol to be published	SS
December 2020	Evidence map complete	SS
June 2021	First draft of report to be completed	SS
November 2021	Systematic review to be published	EEF

References

Alexander, R. (2017). Developing dialogue: process, trial, outcomes. Paper presented at: 17th biennial EARLI conference. Tampere, Finland.

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

Alkire, S. (2005). Why the capability approach?. *Journal of Human Development*, *6*(1), 115-135.

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.

Baddeley AD, Hitch GJ. 1974. Working memory. In The Psychology of Learning and Motivation: Advances in Research and Theory, ed. GA Bower, pp. 47–89. Academic.

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J., & Macleod, M. R. (2019). Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews*, *8*(1), 1-12.

Basma, B., & Savage, R. (2018). Teacher Professional Development and Student Literacy Growth: A Systematic Review and Meta-analysis. *Educational Psychology Review*, *30*(2), 457-481.

Baumgartner, M. (2015). Parsimony and causality. *Quality & Quantity*, *49*(2), 839-856.

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 and Glaser R (eds.) Innovations in learning: New environments for education pp. 289-325. Lawrence Erlbaum Associates, Inc.

Brown P, Roediger H and McDaniel M (2014) *Make it Stick: The Science of Successful Learning*. Cambridge, MA: Belknap Press of Harvard University Press.

Brunton G, O'Mara-Eves A, Thomas J. (2014). The 'active ingredients' for successful community engagement with disadvantaged expectant and new mothers: a qualitative comparative analysis. Journal of Advanced Nursing, 70(12), 2847-2860. DOI: 10.1111/jan.12441

Brunton J, Graziosi S, Thomas J (2017) Tools and technologies for information management. In: Gough D, Oliver S, Thomas J (eds) An Introduction to Systematic Reviews. London: Sage, pages 145-180.

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

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.

Coe, R. (2002). *It's the Effect Size, Stupid. What effect size is and why it is important.* Retrieved from https://www.leeds.ac.uk/educol/documents/00002182.htm on 30 Oct 2020.

Cohen, L., Manion, L., & Morrison, K. (2018). Research methods in education (eight edition). Abingdon: Routledge.

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.

Deeks, JJ, Higgins, JPT, Altman, DG. (2020). Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). Cochrane, 2020. Available from www.training.cochrane.org/handbook.

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). Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/843068/2020_EYFS_assessment_and_reporting_arrangements_.pdf

DfE [Department for Education] (2019). Schools, pupils and their characteristics: January 2019. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/812539/Schools_Pupils_and_their_Characteristics_2019_Main_Text.pdf

Domitrovich, C.E., Bradshaw, C.P., Poduska, J.M., Hoagwood, K., Buckley, J.A., Olin, S., Romanelli, L.H., Leaf, P.J., Greenberg, M.T. and Ialongo, N.S., (2008). Maximizing the implementation quality of evidence-based preventive interventions in schools: A conceptual framework. Advances in School Mental Health Promotion, 1(3), 6-28.

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.

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.

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

Fiss, P. C. (2011). Building better causal theories: A fuzzy set approach to typologies in organization research. *Academy of Management Journal*, *54*(2), 393-420.

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. Available at: stem. org. uk/resources/elibrary/resource/465564/effects-high-quality-professional-development-teachers-and (accessed 26 March 2020)*.

Furnari, S., Crilly, D., Misangyi, V. F., Greckhamer, T., Fiss, P. C., & Aguilera, R. (2020). Capturing Causal Complexity: Heuristics for Configurational Theorizing. *Academy of Management Review*.

Goodson, I., & Hargreaves, A. (1996). *Teachers' professional lives*. London: Falmer Press.

Goodrich, J. (2021). *As yet untitled.* John Catt Educational.

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), 105-112.

Harris K, Kneale D, Lasserson TJ, McDonald VM, Grigg J, Thomas J. (2019). School-based self-management interventions for asthma in children and adolescents: a mixed methods systematic review. Cochrane Database of Systematic Reviews, Issue 1. Art. No.: CD011651. DOI: 10.1002/14651858.CD011651.pub2.

Hedges LV (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics,* 6(2), 107-128.

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

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

Higgins J., Savovic, J., Page M., Sterne J. (2019). Revised Cochrane risk-of-bias tool for randomized trials (RoB 2). Accessed on 9 October 2020 at https://www.riskofbias.info/welcome/rob-2-0-tool/current-version-of-rob-2.

Higgins JPT, Savović J, Page MJ, Elbers RG, Sterne JAC. Chapter 8: Assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). Cochrane, 2020. Available from www.training.cochrane.org/handbook.

Hobbiss, M., Sims, S., & Allen, B. (2020). Habit formation limits growth in teacher effectiveness: A review of converging evidence from neuroscience and social science. *Review of Education.*

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.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*(4), 509-539.

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

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.

Kraft, M. (2019). Interpreting effect sizes of education interventions. Annenberg EdWorkingPaper No. 19-10.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the IES website at http://ies.ed.gov/ncser/.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE publications, Inc.

Liu, S., & Phelps, G. (2020). Does Teacher Learning Last? Understanding How Much Teachers Retain Their Knowledge After Professional Development. *Journal of Teacher Education*, *71*(5), 537-550.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260-293.

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

Miake-Lye, I. M., Hempel, S., Shanman, R., & Shekelle, P. G. (2016). What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. *Systematic Reviews*, *5*(1), 28.

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

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097.

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

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. *National Bureau of Economic Research Working Paper Series*, (w27476).

Opfer, V. D., & Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of educational research*, *81*(3), 376-407.

O'Mara-Eves A, Brunton G, McDaid D, Oliver S, Kavanagh J, Jamal F, Matosevic T, Harden A, Thomas J (2013). Community engagement to reduce inequalities in health: a systematic review, meta-analysis and economic analysis. Public Health Research, 1(4). DOI: 10.3310/phr01040.

O'Mara-Eves, A., Thomas, J., McNaught, J. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews, 4*(1), 5.

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. (2016). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data (No. w21986). National Bureau of Economic Research.

Ragin C. (2008). *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. London: University of Chicago Press.

Rihoux, B., & De Meur, G. (2009). Crisp-set qualitative comparative analysis (csQCA). *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*, *51*, 33-68.

Rogers, S., Brown, C., & Poblete, X. (2020). A systematic review of the evidence base for professional learning in early years education (The PLEYE Review). *Review of Education*, *8*(1), 156-188.

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 *Perspectives on scientific argumentation* (pp. 235-264). Springer, Dordrecht.

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

Senn, S. (2013). Seven myths of randomisation in clinical trials. *Statistics in Medicine*, *32*(9), 1439-1450.

Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, Kelly MP, Thomas J (2014) Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Research Synthesis Methods 5: 31-49.

Sims, S., & Fletcher-Wood, H. (2019). Identifying evidence-based professional development: programmes, forms and mechanisms. In Scutt, C., & Harrison, S., (Eds.) Teacher CPD: International trends, opportunities and challenges. John Catt Educational.

Sims, S., & Fletcher-Wood, H. (2020). Identifying the characteristics of effective teacher professional development: a critical review. *School Effectiveness and School Improvement* https://doi.org/10.1080/09243453.2020.1772841

Tanner-Smith, E. 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.

Thiem, A. (2017). Conducting configurational comparative research with qualitative comparative analysis: a hands-on tutorial for applied evaluation scholars and practitioners. *American Journal of Evaluation*, *38*(3), 420-433.

Thomann, E., & Maggetti, M. (2020). Designing research with qualitative comparative analysis (QCA): Approaches, challenges, and tools. *Sociological Methods & Research*, *49*(2), 356-386.

Thomas J, Graziosi S, Brunton J, Ghouze Z, O'Driscoll P, Bond M . *EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis*. EPPI-Centre Software. London: UCL Social Research Institute; 2020.

Thomas J, McNaught J, Ananiadou S (2011) Applications of text mining within systematic reviews. Research Synthesis Methods 2: 1-14.

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

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

Tsou AY, Treadwell JR, Erinoff E, Schoelles K (2020). Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Syst Rev 9, 73 https://doi.org/10.1186/s13643-020-01324-7

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

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.

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.

White IR, Thomas J. Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. Clinical Trials 2005; 2: 141–151.

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* (Issues & Answers Report, REL 2007– No. 033). Institute of Education Sciences.

Appendix A: Example database search

Example database search for ERIC (EBSCO interface)

No. of results run on 2/11/2020: 3,763
Key: N = within *n words* proximity to; *=truncation

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"
S2	TI (teacher OR teachers) OR AB (teacher OR teachers) OR TI (educators OR educator) OR AB (educator OR educators)
S3	S1 OR S2
S4	S3 - Limiters - Education Level: Adult Basic Education, Adult Education, Higher Education
S5	S3 - Limiters - Education Level: Early Childhood Education, Elementary Education, Elementary Secondary Education, Grade 1, Grade 2, Grade 3, Grade 4, Grade 5,
S6	(S3 NOT S4)
S7	S5 OR S6
S8	TI (professional N3 (development OR training OR learning OR education)) OR AB (professional N3 (development OR training OR learning OR education))
S9	TI CPD OR AB CPD
S10	TI PD
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*))
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*))
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

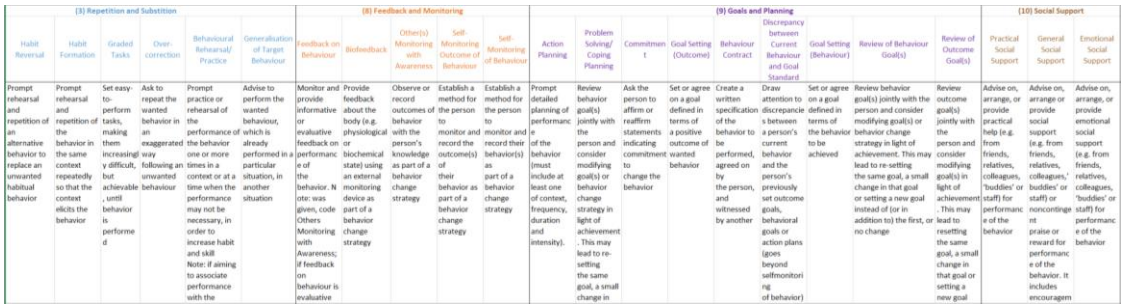
	program*") OR TI ((coaching OR mentoring) N2 program*) OR AB ((coaching OR mentoring) N2 program*)
S14	DE "Professional Education" OR DE "Teacher Improvement" OR DE "Continuing Education" OR DE "Improvement Programs" OR DE "Mandatory Continuing Education" OR DE "Teacher Workshops" OR DE "Professional 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"
S15	S8 OR S9 OR S10 OR S11 OR S12 OR S13 OR S14
S16	TI ("pre service" OR "pre-service" OR preservice OR "prospective teacher*") NOT TI("in service" OR inservice OR "in-service")
S17	S7 NOT S16
S18	S17 AND S15
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*))
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*))
S21	TI ((teacher OR teachers OR educator OR educators) AND "continuing learning") OR AB ((teacher OR teachers OR educator OR educators) N5 "continuing learning")
S22	TI ((teacher OR teachers OR educator OR educators) AND "continuing training") OR AB ((teacher OR teachers OR educator OR educators) N5 "continuing training")
S23	TI ((teacher OR teachers OR educator OR educators) AND "skills program*") OR AB ((teacher OR teachers OR educator OR educators) N5 "skills program*")
S24	TI ((teacher OR teachers OR educator OR educators) AND "continuing development") OR AB ((teacher OR teachers OR educator OR educators) N5 "continuing development")
S25	TI ((teacher OR teachers OR educator OR educators) AND "continuing education") OR AB ((teacher OR teachers OR educator OR educators) N5 "continuing education")
S26	S19 OR S20 OR S21 OR S22 OR S23 OR S24 OR S25
S27	S18 OR S26
S28	DE "Randomized Controlled Trials" OR DE "Outcome Measures" OR DE "Program Evaluation" OR DE "control groups" OR DE "experimental groups"
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")
S30	TI RCT OR AB RCT
S31	TI random* OR "cluster random*" OR trial
S32	AB "trial registration"
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))
S34	AB Random* AND (controlled N2 (trial OR trials OR study OR studies OR experiment OR design OR evaluat*))
S35	AB Random* AND (Comparison N2 (trial OR trials OR study OR studies OR experiment OR design OR evaluat*))
S36	AB (Control OR Comparison OR Intervention OR Experiment*) AND ("Treatment as Usual" OR TAU OR "business as usual" OR BAU)
S37	AB "Treatment group*" OR "Treatment arm" OR "Treatment arms" OR (Treatment N2 teachers) OR (Treatment N2 educators) OR "Treatment participants" Or "treatment 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"
S39	AB "Experimental group*" OR "Experimental arm" OR "Experimental arms" OR "Experimental teachers" OR "Experimental educators" OR "Experimental participants" OR "Experimental condition"
S40	AB "Comparison group*" OR "Comparison arm" OR "Comparison arms" OR "Comparison teachers" OR "Comparison educators" OR "Comparison participants" OR "Comparison condition"
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"
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*))

S43	AB (prospective* N2 control*) OR (prospective* N2 comparison*) OR (prospective* N2 trial)
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)))
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))
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
S47	S27 AND S46 - Limiters - Date Published: 20020101-20201231
S48	S47 - Limiters - Language: English

Appendix B: Sample of the Michie framework

The Michie framework is too large to include in full but can be found online using the reference provided herein.



(3) Repetition and Substitution						(8) Feedback and Monitoring					(9) Goals and Planning								(10) Social Support		
Habit Reversal	Habit Formation	Graded Tasks	Over-correction	Behavioural Rehearsal/ Practice	Generalisation of Target Behaviour	Feedback on Behaviour	Biofeedback	Other(s) Monitoring with Awareness	Self-Monitoring Outcome of Behaviour	Self-Monitoring of Behaviour	Action Planning	Problem Solving/ Coping Planning	Commitment	Goal Setting (Outcome)	Behaviour Contract	Discrepancy between Current Behaviour and Goal Standard	Goal Setting (Behaviour)	Review of Behaviour Goal(s)	Review of Outcome Goal(s)	Practical Social Support	General Social Support	Emotional Social Support
Prompt rehearsal and repetition of an alternative behavior to replace an unwanted habitual behavior	Prompt rehearsal and repetition of the behavior in the same context repeatedly so that the context elicits the behavior	Set easy-to-perform tasks, making them increasingly difficult, but achievable, until behavior is performed	Ask to repeat the wanted behavior in an exaggerated way following an unwanted behaviour	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	Advise to perform the wanted behaviour, which is already performed in a particular situation, in another situation	Monitor and provide informative or evaluative feedback on performance of the behavior. Note: was given, code Others Monitoring with Awareness; if feedback on behaviour is evaluative	Provide feedback about the body (e.g. physiological or biochemical state) using an external monitoring device as part of a behavior change strategy	Observe or record outcomes of behavior with the person's knowledge as part of a behavior change strategy	Establish a method for the person to monitor and record the outcome(s) of their behavior as part of a behavior change strategy	Establish a method for the person to monitor and record their behavior(s) as as part of a behavior change strategy	Prompt detailed planning of performance of the behavior (must include at least one of context, frequency, duration and intensity).	Review behavior goal(s) jointly with the person and consider modifying goal(s) or behavior change strategy in light of achievement. This may lead to re-setting the same goal, a small change in	Ask the person to affirm or reaffirm statements indicating commitment to change the behavior	Set or agree on a goal defined in terms of a positive outcome of wanted behavior	Create a written specification of the behavior to be performed, agreed on by the person, and witnessed by another	Draw attention to discrepancies between a person's current behavior and the person's previously set outcome goals, behavioral goals or action plans (goes beyond selfmonitoring of behavior)	Set or agree on a goal defined in terms of the behavior to be achieved	Review behavior goal(s) jointly with the person and consider modifying goal(s) or behavior change strategy in light of achievement. This may lead to re-setting the same goal, a small change in that goal or setting a new goal instead of (or in addition to) the first, or no change	Review outcome goal(s) jointly with the person and consider modifying goal(s) in light of achievement. This may lead to resetting the same goal, a small change in that goal or setting a new goal	Advise on, arrange, or provide practical help (e.g. from friends, relatives, colleagues, 'buddies' or staff) for performance of the behavior	Advise on, arrange or provide social support (e.g. from friends, relatives, colleagues,' buddies' or staff) or nonconting nt praise or reward for performance of the behavior. It includes encouragem	Advise on, arrange, or provide emotional social support (e.g. from friends, relatives, colleagues, 'buddies' or staff) for performance of the behavior

[i] https://soltreemrls3.s3-us-west-2.amazonaws.com/marzanoresearch.com/media/documents/pdf/AppendixB_DTLGO.pdf

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