Version 07.04.20

# Identifying the characteristics of effective teacher professional development: a critical review

Sam Sims

Harry Fletcher-Wood

Several influential reviews and two meta-reviews have converged on the position that teacher professional development (PD) is more effective when it is: sustained, collaborative, subject-specific, draws on external expertise, has buy-in from teachers and is practice-based. This consensus view has now been incorporated in government policy and official guidance in several countries. This paper reassesses the evidence underpinning the consensus, arguing that the reviews on which it is based have important methodological weaknesses, in that they employ inappropriate inclusion criteria and depend on an invalid inference method. The consensus view is therefore likely to be inaccurate. It is argued that researchers would make more progress identifying characteristics of effective professional development by looking for alignment between evidence from basic research on human skill acquisition and features of rigorously-evaluated PD interventions.

## 1. Introduction

International surveys suggest that teachers spend, on average, 10.5 days per year engaged in courses, workshops, conferences, seminars, observation visits or in-service training (Sellen, 2016). The motivation for this substantial investment in professional development (PD) is clear: improved pupil attainment is associated with improvements in income, happiness and health (Chetty, Friedman, & Rockoff, 2014; Hanushek, 2011; Lance, 2011). 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 attainment (e.g. Allen et al., 2011; Allen et al., 2015), most schools do not have access to these programmes, due to either cost or location. School leaders and teacher educators need instead to know which *characteristics* of professional development matter to help them design or commission effective PD (Hill, Beisiegel, & Jacob, 2013).

Scholarly attempts to identify the characteristics of professional development which improve pupil attainment stretch back to 1995 (Corcoran, 1995). Despite much research in the following years, there was little consensus among researchers (Guskey, 2003a). More recently, however, several reviews have converged on the position that PD is more likely to improve pupil attainment if it is sustained, collaborative, has teacher buy-in, is subject-specific, draws on external expertise and is practice-based (Desimone, 2009; Timperley, Wilson, Barrar, & Fung, 2007; Walter, 2012; Wei et al., 2009). These reviews have been summarised in two meta-reviews, which further endorse these principles (Cordingley et al., 2015; Dunst, Bruder, & Hamby, 2015).

Indeed, this convergence of opinion is marked enough that it is often explicitly referred to as a consensus (Desimone, 2009; Darling-Hammond, Hyler & Gardner, 2017; Hill et al., 2013; Van Driel, Meirink, van Veen, & Zwart, 2012; Wei et al., 2009). Moreover, it has now influenced policy in several countries (DfE, 2016; Caena, 2011; Desimone, 2009). This notable level of agreement motivates the research question addressed by this paper: is the consensus warranted by the existing evidence?

The paper begins in Section 2 by setting out the consensus view and the ways in which it is influencing research, policy and practice. Section 3 then describes our methods for scrutinising the relevant literature. In section 4, we set out the findings from our detailed investigation of the underpinning evidence. Having identified methodological weaknesses in existing research, we then move on to ask: how can we validly identify the characteristics of effective professional development? In section 5, we argue that this requires combining evidence that a PD programme has a causal impact on pupil attainment, with independent evidence of mechanism explaining how a characteristic of that programme has an impact. We illustrate our proposed approach with reference to the literature on instructional coaching. The paper then concludes in section 6 with a discussion of implications for policy, practice and research.

Of course, ours is not the first study to engage critically with this literature. Guskey (2003a) pointed out that many early review papers included poorly designed studies and failed to rigorously investigate the relationship between specific characteristics of PD and pupil learning. Kennedy (2016) then built on Guskey's criticisms by showing that excluding less rigorous studies from reviews leads to conclusions that diverge from the consensus view. In addition, Kennedy (2016), Opfer and Pedder (2011) and Sztjan, Campbell and Yoon (2011) have all called for better use of theory to help identify the characteristics of effective PD, though each in quite different ways.

The present study extends the literature on three fronts. First, we argue that recent studies using more rigorous inclusion criteria are still likely to lead to erroneous conclusions because they cannot distinguishing the active ingredients of rigorously evaluated interventions from the causally redundant components. Second, we respond to the calls for better use of theory by explicating precisely how theory combines with empirical evidence to help isolate characteristics of effective professional development. Third, and relatedly, this allows us to identify parts of the consensus view that are not supported by the existing evidence, as well as those which should be retained or adapted. The article therefore makes a number of novel contributions, as well as having implications for policy and practice.

## 2. Background and motivation

Several different literature reviews concur that PD is more effective if it incorporates six characteristics, which they conceptualise as necessary or sufficient conditions (e.g. Cordingley et al., 2015), critical features (e.g. Desimone, 2009), or simply as being important (e.g. Timperley et al., 2007). Despite disagreement at the margins and pervasive differences in terminology, the underlying claims are highly consistent, as illustrated by a recent meta-synthesis (Dunst et al., 2015). The six characteristics are discussed in turn below.

First, PD is claimed to be more effective if it is sustained over time (Blank & Alas, 2009; Cordingley et al., 2015; Desimone, 2009; Dunst et al., 2015; Timperley et al., 2007; Walter, 2012; Wei et al., 2009). Some of the reviews develop this point further by claiming that PD should be organised in a cycle or rhythm in which the content is revisited or iteratively developed. The justification for this is usually that it takes time for teachers to assimilate new knowledge. By contrast, single, one-day sessions are often cited as being particularly ineffective.

Second, PD is argued to be more effective if teachers take part as a group (Cordingley et al., 2015; Desimone, 2009; Dunst et al., 2015; Timperley et al., 2007; Walter, 2012; Wei et al., 2009). Most often the requirement for collaboration is formulated as the need to work with multiple peers or a ‘community of practice’. The justification for this is usually that it gives teachers the chance to challenge each other and clarify misunderstandings. The transfer of information directly from a course leader to an individual participant is often contrasted as being particularly ineffective.

Third, PD is said to be more effective if teachers identify with and endorse taking part in it (Cordingley et al., 2015; Timperley et al., 2007; Walter, 2012). This is often framed as the claim that voluntary PD is more effective than obligatory PD. However, some researchers make the more nuanced point that there can be strong buy-in for obligatory PD if the purpose and benefits of the PD are clearly explained to participants, so that they can see the value of taking part (Timperley et al., 2007; Dunst et al., 2015).

Fourth, PD is claimed to be more effective when it involves training in subject knowledge (Blank & Alas, 2009; Cordingley et al., 2015; Desimone, 2009; Dunst et al., 2015; Wei et al., 2009). This is often contrasted with PD that only involves training in general pedagogical techniques, divorced from the content that they would be used to deliver. Indeed, it is often argued that the two are complementary and PD is therefore most effective when *both* training on subject knowledge and general pedagogical techniques are delivered together.

Fifth, PD is said to be more effective when it involves outside expertise (Cordingley et al., 2015; Dunst et al., 2015; Timperley et al., 2007; Walter, 2012; Wei et al., 2009). In general, this means input from people that do not work in the same school. The justification for this is generally that this is needed to provide challenge or fresh input, as opposed to recycling existing expertise from inside the school, with which teachers may already be familiar.

Sixth, PD is argued to be more effective when it involves opportunities to use, practise or apply what has been learned (Blank & Alas, 2009; Cordingley et al., 2015; Desimone, 2009; Dunst et al., 2015; Timperley et al., 2007; Walter, 2012; Wei et al., 2009). Again, the justification for this is often that it helps teachers apply what they have learned in real classroom situations. This approach is often contrasted with lectures in which teachers receive new information passively but do not apply it.

Importantly, the consensus is now influencing research, policy and the design of professional development. Indeed, the consensus view has become embedded in official guidance in the UK (DfE, 2016), the EU (Caena, 2011) and the US (see Desimone, 2009). In the UK, it has directly informed the development of the Standards for Teacher Professional Development, which aim to provide guidance on effective PD based on the "best available research" (DfE, 2016, p. 4). The consensus has also influenced policy in the US through the Every Student Succeeds Act, which requires professional development to be sustained, collaborative and practice-based in order to attract federal funding (see Combs & Silverman, 2016). Furthermore, checklists have been created so that teacher educators can identify whether their PD sessions conform to the consensus view (Wei et al., 2009; Main & Pendergast, 2015) and questionnaire instruments have also begun to reflect it (e.g. Rutkowski et al., 2013). Finally, in influencing programme designs, some researchers explicitly refer to the consensus view and its characteristics in explaining their design choices, (e.g. Jacob, Hill, & Corey, 2017; Nugent et al., 2016), or map the approach their programmes take to these characteristics (e.g. Greenleaf et al., 2010; Penuel, Gallagher, & Moorthy, 2011). While we acknowledge that there are some who disagree with the consensus (e.g. Kennedy, 2016), it is clearly influencing practice – and thus warrants critical scrutiny.

Interestingly, several recent evaluations of PD interventions which include all of the consensus view characteristics have not found a positive a positive impact. For example, Garet et al. (2016) evaluated a programme that provided sustained, collaborative PD for volunteers focused on teachers' mathematical knowledge, led by outside experts, including active learning. The study was implemented as intended, but students in the treatment group showed *weaker* achievement on state tests than the control group. Similarly, Garet et al. (2011) evaluated a programme that offered two years' active, collaborative, PD focused on mathematics, with long-run follow-up, delivered by outside experts. The programme was implemented as intended, but led to no observable improvements in student achievement. Further, Jacob, Hill and Corey (2017) studied the Maths Solutions programme "because it meets the criteria articulated in Desimone's (2009) description of effective professional development program features" (p. 380), but the programme led to no improvement in student achievement. These findings further motivate this paper.

## 3. Methods

The present research examines the evidence supporting the consensus view of effective PD. It is therefore a methodological review, which aims to expose a strand of the literature to critical scrutiny (Grant & Booth, 2009). Consequently, we employed ancestry searching (Cooper, 2010; Conn et al., 2003) to trace backwards from policy documents, to the meta-reviews and reviews they cited, and then back a step further to the original studies that they cited. This allowed us to identify the research underpinning the consensus view, yielding a web of policy documents (DfE, 2016; Caena, 2011), supporting meta-reviews (Cordingley et al., 2015; Dunst et al., 2015), underpinning literature reviews (Desimone, 2009; Timperley, Wilson, Barrar, & Fung, 2007; Walter, 2012; Wei et al., 2009) and foundational original research articles, all of which have been cited in support of the consensus view. It is important to note here that our objectives stand in contrast to those of an aggregative review, which aims to identify and compile studies that provide a representative picture of the current evidence base (Gough, Thomas, & Oliver, 2012b). While the searches used have been extensive, methodological reviews like ours do not require exhaustive searches of the literature (Gough, Thomas, & Oliver, 2012a).

This approach was combined with further searches using combinations of the terms: "teacher"; "professional development" or "continuing professional development" with "characteristics of" or "features of". References of the articles recovered were also searched. The articles identified using this search method provided important context and perspective on those identified through ancestry searching. In particular, this approach identified a number of reviews which did not endorse the consensus view (e.g. Kennedy, 2016; Lynch et al., 2019; Yoon, 2011; Kraft, Blazar, & Hogan, 2018), as well as a wide range of relevant original research papers.

An important part of our approach in this article is to focus exclusively on studies which use pupil achievement as an outcome measure. Our justification for this is based on Guskey's argument that to "gain authentic evidence and make serious improvements" research on PD must focus on "professional development's ultimate goal: improvements in student learning outcomes" (2003b, p.750). Of course, professional development programmes may achieve other desirable outcomes for teachers and schools. For example, studies have investigated intermediate outcomes including teacher self-efficacy (Nugent et al., 2016) and confidence (Kitmitto et al., 2018). However, the ultimate justification for professional development, and the time and taxpayer money invested in, is the impact on student learning.

## 4. Results

Our literature review revealed that the review articles underpinning the consensus view follow a set of common steps:

(1) Researchers have searched the literature to form a longlist of articles which have evaluated specific PD interventions.

(2) They have used inclusion criteria to remove articles deemed to be of limited relevance or quality.

(3) Researchers have sorted these articles into those that find the intervention they evaluate has had a positive impact, and those that did not.

(4) They have looked for characteristics of PD which are (in some way) related to the effectiveness of the evaluated PD interventions.

We structure our discussion of our findings around two important parts of this logical sequence. In section 4.1, we focus on the criteria used to include or exclude studies in step 2 above. In section 4.2, we discuss the inference process used in step 4.

### *4.1 Appropriateness of inclusion criteria*

The selection criteria employed by a literature review affect its conclusions (McDonagh et al., 2013) for at least two reasons. First, they determine the articles reviewed: missing important studies will give a partial and potentially inaccurate picture of the evidence. Second, the criteria must exclude studies that do not employ a research design capable of answering the research question posed by the review. In seeking to identify characteristics of effective PD, included studies must identify which PD interventions are effective in raising and attainment and which are not. The findings of the review will therefore be compromised if the studies included are incomplete or the methods employed in the studies are inappropriate for answering the questions. Hence, the PRISMA standards for reporting systematic reviews (Liberati et al., 2009, p. 5) states that “Knowledge of the eligibility criteria is essential in appraising the validity, applicability, and comprehensiveness of a review.”

We now consider the inclusion criteria in the meta-reviews and reviews on which the consensus view rest. Recall that ancestry searching involves beginning with a specific document and working back through the references cited to identify the underpinning evidence. We chose to begin with the meta-review by Cordingley et al. (2015), since it summarises several reviews and has directly influenced policy. This meta-review found 980 reviews which were rated on a three-point scale stretching from: 1 - methodology and weighting of evidence clear; 2 – methodology clear but no weighting of evidence; and 3 – methodology unclear. All level 1 and level 2 reviews were retained. No further details were given on how clarity of methodology or weighting were judged for each review. However, Cordingley et al. (2015) do rank the reviews that they use in their meta-review in terms of quality. The review which they give the highest score to is Timperley et al. (2007), which they describe as “the only fully consistent and rigorous review” which they emphasise as “a cornerstone for the umbrella review” that they conduct (Cordingley et al., 2015, p. 4).

Given the weight accorded to it, we now consider the inclusion criteria used by Timperley et al. (2007). This review judged quantitative studies on a three-point scale in three areas: sampling methods; control groups; and validity and reliability of test instruments. Qualitative studies were also judged on a three-point scale in three areas: depth of data collection and analysis; validity and reliability of assessment; and method of triangulation. Study inclusion was also based on impact: studies which demonstrated “medium to high impact” were designated core studies, while studies with “low, no, or negative impact” were designated supplementary studies, the results of which were used to support conclusions from the core studies (Timperley et al., 2007, p.23), a practice which is contrary to the norms of meta-analysis (Basma & Savage, 2017, p.5).

Table 10.2 in Timperley et al. (2007) lists eleven studies relevant to the characteristics of effective PD in secondary schools that were rated highly enough to be included (there is no equivalent section for primary schools). We identified these original studies and reviewed the research methods that they employed:

- Adey (1999) employed a simple research design in which participants were matched to controls based on age and ability.

- Anderson (1992) employed an experimental design but only had a sample size of 20, which dropped to 16 through attrition.
- Bishop, Berryman, Powell and Teddy (2005), Confrey, Castro-Filho and Wilhelm (2000) and D'Oria (2004) employed no control variables at all, relying instead on unadjusted comparisons of outcomes.
- Huffman, Goldberg and Michelin (2003) matched six comparison teachers to eight novices and seven experts, based on teaching experience and student demographics.
- Metcalf, Vontz, and Patrick (2000) employed ANOVA methods to compare group means.
- Moxon (2003), Ross (1994) and Ross, Roleiser and Hogaboam-Gray (1999) employ before and after designs but neither conduct any covariate adjustment.
- Schober (1984) does employ regression analysis but only adjusts for degree subject, urban location and average income.
- Tasker (2001) only reports qualitative findings.

By What Works Clearinghouse (WWC) standards, nine of the ten studies mentioned above would be graded 'Does Not Meet Evidence Screens' because they do not establish baseline equivalence of treatment and control groups. This is essential to establish the impact of a PD programme, because without baseline equivalence, any differences in post-participation outcomes between treatment and control groups may just reflected unmeasured differences in pre-treatment characteristics of these two groups (Mill, 1884). The randomised study may qualify for WWC 'Meets Evidence Standards Without Reservation' but the high rate of attrition (missing follow up data) means it would likely be disqualified altogether, because where attrition is correlated with treatment assignment, this undermines the baseline equivalence originally established by the randomisation (Shadish, Cook & Campbell, 2002).

Thus, the most highly-rated review in Cordingley et al. (2015) uses weak inclusion criteria which admit studies employing designs unable to establish whether the PD interventions were effective or not. Crucially, the validity of step 4 of (set out above) depends on identifying interventions which are, and are not, effective (step 3). Since step 3 uses studies which do not establish equivalent control groups, this casts doubt on the validity of the conclusions reached in step 4.

How do the inclusion criteria in other reviews in the literature compare? For space reasons, we limit ourselves here to cross-subject reviews that look across different types of PD. Wei et al. (2009, p. 3) explicitly allow studies using any methodologies including qualitative and case study methods, though they note that “the inferences that can be drawn from such research should be treated as suggestive rather than conclusive”. Desimone (2009) and Walter and Briggs (2012) do not employ any explicit inclusion criteria but both include case study research. Yoon et al. (2007) use the more rigorous What Works Clearing House standards to screen the papers in their review but conclude that “Because of the lack of variability in form and the great variability in duration and intensity across the nine studies, discerning any pattern in these characteristics and their effects on student achievement is difficult” (p. 3). Kennedy (2016) allows only experimental studies but finds no clear patterns between programme design features and pupil outcomes. In summary, many of the reviews which espoused the consensus view did not employ appropriate inclusion criteria; while those that did employ appropriate inclusion criteria tended not to endorse the consensus view.

### *4.2 Validity of inference methods*

Even if it were the case these reviews had employed appropriate inclusion criteria, it is unclear that the inferences in step 4 of the process would yield accurate conclusions about the characteristics of effective PD.

All four of the consensus-view, cross-subject reviews that conducted step 4 of the review process (Desimone 2009; Timperley et al., 2007; Walter & Briggs, 2012; Wei et al., 2009) used a thematic approach to identify the characteristics of effective PD, seeking to identify features that recurred among interventions that were found to be effective. For example, Timperley et al. (2007) note that all of their ‘core studies’ involve teachers working in structured professional groups. The authors interpreted any counterexamples as evidence that collaboration is necessary but not sufficient for effective PD. Desimone (2009) also looks for recurring features of successful interventions, adding that such regularities are more persuasive when they come from studies using a range of different research designs. Walter & Briggs (2012) and Wei et al. (2009) also look for recurring themes among effective interventions. The meta-reviews by Cordingley et al. (2015) and Dunst et al. (2015) then analysed the claims made across the various reviews and looked for agreement among them. For example, Cordingley et al. (2015) highlights agreement among the reviews that prolonged CPD is more effective than short CPD. This is accompanied by a caveat, citing Timperley et al. (2007), that not all prolonged PD programmes are effective and the claim that what distinguishes effective prolonged PD is what the additional time is used for.

The overall inference method described in the preceding paragraph is logically flawed. The regular occurrence of specific features of PD in effective interventions does not, in itself, warrant any inference about the effect of that feature of the intervention. The risk is that, in the terminology of Mackie (1974), effective interventions include causally redundant components. Put another way, consensus view characteristics could occur frequently in effective PD interventions for reasons other than their contribution to the effectiveness of that PD. Take collaboration: schools have limited budgets and collective PD will be cheaper to provide than one-to-one PD. Collaboration is therefore likely to occur in PD, even if it is causally redundant. Alternatively, consider buy-in: teachers may be enthusiastic about an effective PD programme because they notice its impact, rather than the programme being effective because teachers have bought into it (Guskey, 2002). Teacher enthusiasm may therefore occur in effective PD, even if it does not causally contribute to it effectiveness. We do not claim here that collaboration or buy-in are undesirable features of PD; only that - contra to the consensus view - compelling evidence of their contribution is currently lacking.

## 5. Alternative methods for identifying the characteristics of effective PD

### *5.1 Philosophical basis*

We began by arguing that school leaders need to be able to identify characteristics of effective PD if they are to design or commission such interventions. Since these characteristics will always come as part of a package, a different research approach may be needed. Russo and Williamson (2007) and Clarke et al. (2014) have revived the arguments of Bradford Hill (1965) and Mackie (1974) to show how this can be done. Their approach to identifying causally non-redundant characteristics involves combining two types of evidence.

The first is evidence of correlation, which they define as probabilistic dependence between two phenomena. For example, certain types of PD might be highly correlated with pupil learning gains in rigorous evaluations. It is important to note that, although Clarke et al. (2014) use the term evidence-of-correlation to refer to this broad category of evidence, they still privilege the sub-set of correlational evidence which is causal in nature e.g. from randomised trials. The second type is evidence of mechanism, which is defined as activities organised in such a way that they are responsible for the phenomenon. In social science, evidence of mechanism might come from basic research describing fundamental characteristics of human motivation or learning, which hold across diverse contexts (see section 5.2 for an example).

Clarke et al. (2014, p. 19) argue that these two types of evidence "integrate in a special way" to become more than the sum of their parts. Evidence of (causal) correlation between a PD intervention and pupil outcomes provides evidence that an intervention as a whole is effective. However, correlational evidence alone cannot distinguish causally redundant from causally non-redundant characteristics of interventions. For example, If a PD programme with a collaborative component affects pupil attainment, this does not demonstrate that collaboration played a causal role; an adapted version of the intervention might have been just as effective if teachers had worked individually. Conversely, evidence of mechanism can help identify non-redundant components of a cause, but cannot determine whether a component will have a causal effect when implemented as part of an intervention. For example, knowing that collaboration is useful in several settings does not guarantee that any PD intervention incorporating collaboration will improve pupil attainment. When both types of evidence converge however, can we be more confident that a non-redundant characteristic of a collectively sufficient causal condition has been identified. That is, if we found a PD intervention incorporating characteristic X, which had been shown to be effective *and* there was evidence that X is effective in bringing about learning or behaviour change in a range of settings, there is far stronger evidential warrant that X is genuinely characteristic of effective PD.

### *5.2 Illustrative example: Instructional coaching*

This section provides an example of this approach using instructional coaching. Our aim is not to provide an exhaustive account or definitive argument but to illustrate how this approach could be employed in the social sciences.

PD interventions based on instructional coaching – an observation, feedback, practice cycle in which an individual teachers received guidance from an expert mentor

- show consistently positive correlations with pupil achievement.[1] Indeed, a recent meta-analysis identified 44 evaluations of instructional coaching programmes, with an average impact on pupil learning of 0.15 standard deviations (Kraft et al., 2018).[2] The inclusion criteria for the meta-analysis required studies to employ either difference-in-difference, regression discontinuity or randomised controlled trial designs. Such designs are able to account for differences between the treatment and control groups that are not recorded in the data, which enables them to credibly identify the causal impact of the coaching programmes on pupil learning (Rosenbaum, 2017). The empirical evidence for instructional coaching also includes replicated randomised controlled trials (Allen et al., 2011; Allen et al., 2015) and evidence from AB testing (Albornoz et al., 2017). An important limitation of the existing evidence is that many of the studies focus on literacy outcomes and are targeted at younger pupils. Nevertheless, taken together, this research provides good evidence of (causal) evidence-of-correlation between instructional coaching and pupil achievement.

As set out above however, it is necessary to combine this with evidence of specific mechanisms in order to persuasively identify the causal characteristics of instructional coaching. One example of evidence of mechanism for coaching comes from research on how and when people change their practice. PD programmes often fail to bring about intended changes in teacher practice (Copur-Gencturk & Papakonstantinou, 2016) and meta-analysis of causal studies in a range of settings suggests that habits - behaviours cued automatically by environmental stimuli - are the most important reason that people fail to change their actions in this way (Webb & Sheeran, 2006). This is because repetition in the presence of specific environmental cues causes behaviour to become automatic (Lally, van Jaarsveld, Potts, & Wardle, 2009). Indeed, research has shown that certain teaching practices becomes more habitual over the early years of teachers careers (Sims, Hobbiss, & Allen, In Press). Research in a very wide range of settings – car use, recycling, blood donation, voting – has shown that people maintain these habitual behaviours, even if their goals change (Wood & Neal, 2007). Neuroscientists

[1] It should be noted that, in line with the argument in sections 4.2 and 5.1, any apparent similarity between the features of instructional coaching and the consensus view is not, in and of itself, evidence that consensus view is correct.

[2] We note that this meta-analysis was published after the meta-reviews bv Cordindley et al. (2015) and Dunst et al. (2015). However, many of the original studies cited in Kraft et al. (2018) were published by 2015.

have also shown how behaviours which are repeated many times become governed by different regions of the brain, making them more resistant to change (Seger & Spiering, 2011). The evidence-of-mechanism reviewed here also has limitations. For example, there are a lack of studies in similar public service settings, such as among doctors or police. Despite this, the range of methods by which, and contexts in which, habits have been shown to influence behaviour suggests that they constitute an important potential mechanism in changing teachers practice.

Coaching incorporates characteristics which are known to promote habit change. Most notably, coaching programmes require teachers to repeatedly practice new skills in their own classrooms. For example, teachers enrolled in the My Teaching Partner programme submit fortnightly videos of themselves practising specific skills in their own lessons, which they then review along with their coach (Allen et al., 2011). Experimental and observational research in a range of contexts, as well as evidence from neuroscientific research, shows that it is necessary to repeatedly practice new behaviours before they become automatic (see Wood and Neal, 2007). Moreover, meta-analysis suggests that repeatedly practicing the new techniques in the environment where you aim to reproduce them in future (i.e. the classroom) helps replace old habits by overwriting the established cue-response relationships (Webb & Sheeran, 2006). The repeated review and feedback incorporated in coaching models helps strengthen these new cue-response relationships even further. This evidence of mechanism for repeating a new technique in the target environment to help ingrain new practices - combined with evidence of correlation between coaching and pupil attainment - suggests that this type of practice is a characteristic of effective PD.

## 6. Discussion

Several reviews and two meta-reviews have established a consensus around the characteristics of effective PD. In this article, we have argued that the underpinning research does not support this consensus because it employs inappropriate inclusion criteria and a flawed inference method.

Some parts of the consensus view, such as collaboration, currently lack evidential warrant. Our argument here is primarily negative, highlighting an absence of evidence that this is characteristic of effective PD. Certainly, the (meta) reviews on which this claim is often based have not established this. Inappropriate inclusion criteria mean these reviews likely do not identify effective PD interventions, let alone their common characteristics. Even if they had identified effective interventions, collaboration may be causally redundant, rather than an active ingredient. Similar arguments apply to the claim that effective PD should be subject-specific. Again, existing research does not warrant these claims: there is currently an absence of evidence.

The claim that PD should be sustained may require revision. In this case, in addition to this absence of evidence, there is also evidence of absence. Moderator analysis from two meta-analyses show that, among interventions which include repeated practice of specific skills, the overall duration (length of time) of the PD programme shows no relationship with the impact on pupil attainment (Basma & Savage, 2017; Kraft et al., 2018). In line with this, evidence reviewed in section 5 suggests that it may be repeated practice that matters, rather than PD being sustained. The difference between these two points is substantively significant. For example, a sustained PD programme might provide fortnightly sessions for two years, but if each part of the curriculum is covered only once, then the intervention does not incorporate repeated practice and is less likely to change teachers' practise. Crucially, it is the combination of evidence of correlation and evidence of mechanism which makes the evidential warrant for repeated practice more compelling than that for PD being sustained.

More generally, we conclude that there are reasons to be sceptical about the methods employed by researchers in developing the consensus view. In particular, our research highlights the dangers involved in meta-reviews (or reviews of reviews) which are often employed to summarise the evidence from a field in a short space of time, in order to inform policy (Thomas, Newman, & Oliver, 2013). As we have seen however, this approach is problematic because the quality of the preceding reviews cannot be fully assessed without a detailed investigating of the underpinning primary research - of the sort we reported in section 4.1. Ironically, this obviates the time/cost benefits of conducting rapid meta-reviews in the first place (Caird et al., 2015; Whitlock et al., 2008). Our study illustrates how, absent this level of scrutiny, reviews of reviews can lead to the propagation of weakly warranted findings through the hierarchy of reviews and onward into public policy, practice and research.

### *6.1 Limitations*

These findings should, of course, be interpreted with regard to the limitations of this study. One such limitation is that this paper has focused solely on studies using student attainment as a criterion for effective PD. While we believe this to be justified, we cannot rule out that an alternative approach that used intermediate, non-attainment outcomes could have come to differing conclusions. Relatedly, most of the studies we have discussed focus on English and maths attainment. This reflects the current state of the literature but a broader evidence base on effective PD in the humanities and creative subjects might allow a more nuanced set of conclusions to be reached. In addition, our discussion has focused exclusively on investigating the evidence base underpinning the consensus view. Readers looking to gain a representative picture of the literature on PD should therefore consult alternative studies such as those by Basma & Savage (2017), Kraft et al., (2018) and (Lynch et al., 2019). Finally, we would like to emphasise again that the example of our propose methods in section 5.1 is intended to be illustrative, rather than definitive.

### *6.2 Implications*

Despite these limitations, our paper has implications for the field. In particular, we believe that researchers looking to identify the characteristics of effective PD should seek alignment between evidence-of-mechanism and evaluations of specific PD interventions which include these mechanisms. For example, a careful consideration of the literature on near- and far-transfer of skills may provide relevant evidence-of-mechanism to support the claim that subject-specific professional development is more effective. In combination with e.g. meta-analytic evidence on the effect of subject-specific professional development, this would provide stronger warrant for the claim that subject-specificity is characteristic of effective PD. This may require inter-disciplinary collaboration between psychologists engaged in basic research about how people learn and acquire skills with applied researchers evaluating PD programmes. Recently developed taxonomies of mechanisms provide a good starting point for researchers looking to pursue this approach (Michie et al., 2013).

Our findings also have direct implications for policy and practice. In the US, the Every Student Succeeds Act currently requires PD to be both sustained and collaborative in order to qualify for federal funding. Policymakers should consider dropping the collaborative criteria and revising the sustained criteria, as discussed above. In England, the Standards for Teachers' Professional Development also recommend that PD should be collaborative. Policymakers should also consider revising this guidance. This is necessary in order to avoid spending scarce resources on programmes that may not be effective and to avoid teacher educators designing existing programmes in line with the consensus view. Policymakers, school leaders and teacher educators should focus instead on commissioning and designing PD with characteristics for which there is strong evidence of both (causal) correlation and mechanism. Funders should also resist calls to organise research on teacher PD around the consensus view (Desimone, 2009).

## Funding

This research did not receive grants from any funding agencies in the public, commercial or not-for-profit sectors.

## References

Adey, P. (1999). *The science of thinking, and science for thinking: A description of Cognitive Acceleration through Science Education (CASE)*. Geneva, Switzerland: International Bureau of Education.

Albornoz, F., Anauati, M. V., Furman, M., Luzuriaga, M., Podesta, M. E., & Tayor, I. (2017). *Training to teach science: Experimental evidence from Argentina* (CREDIT Research Paper No. 17/08). Retrieved from: https://www.nottingham.ac.uk/credit/documents/papers/2017/17-08.pdf

Allen, J., Hafen, C., Gregory, A., Mikami, A., & 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., Pianta, R., Gregory, A., Mikami, A., & Lun, J. (2011). An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement. *Science*, *333*(6045) 1034-1037.

Anderson, V. (1992). A teacher development project in transactional strategy instruction for teachers of severely reading-disabled adolescents. *Teaching and Teacher Education*, *8*(4), 391-403.

Basma, B., & Savage, R. (2017). Teacher professional development and student literacy growth: a systematic review and meta-analysis. *Educational Psychology Review*, 30, 457–481.

Bishop, R., Berryman, M., Powell, A., & Teddy, L. (2005). *Te Kotahitanga: Improving the educational achievement of Maori students in mainstream education*. Wellington, New Zealand: Ministry of Education.

Blank, R. K., & De Las Alas, N. (2009). *The Effects of Teacher Professional Development on Gains in Student Achievement: How Meta Analysis Provides Scientific Evidence Useful to Education Leaders*. Washington DC: Council of Chief State School Officers.

Bradford Hill, A. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, *58*, 295–300.

Caena, F. (2011). *Literature review: Quality in teachers' continuing professional development*. Brussels, Belgium: European Commission. Retrieved from: http://ec.europa.eu/dgs/education_culture/repository/education/policy/strategic-framework/doc/teacher-development_en.pdf

Caird, J., Sutcliffe, K., Kwan, I., Dickson, K., & Thomas, J. (2015). Mediating policy-relevant evidence at speed: are systematic reviews of systematic reviews a useful approach?. *Evidence & Policy: A Journal of Research, Debate and Practice*, *11*(1), 81-97.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593-2632.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi*, *33*(2), 339-360.

Combs, E. & Silverman, S. (2016) *Bridging the Gap: Paving the pathway from current practice to exemplary professional learning*. Frontline Research & Learning Institute. Retrieved from https://www.frontlineeducation.com/uploads/2018/01/ESSA_Bridging_the_Gap.pdf

Confrey, J., Castro-Filho, J., & Wilhelm, J. (2000). Implementation research as a means of linking systemic reform and applied psychology in mathematics education. *Educational Psychologist, 35*(3), 179-191.

Conn, V. S., Isaramalai, S. A., Rath, S., Jantarakupt, P., Wadhawan, R., & Dash, Y. (2003). Beyond MEDLINE for literature searches. *Journal of Nursing Scholarship*, *35*(2), 177-182.

Cooper, H. (2010). Research synthesis and meta-analysis: A step-by-step approach (4th ed.). Los Angeles: Sage

Copur-Gencturk, Y., & Papakonstantinou, A. (2016). Sustainable changes in teacher practices: a longitudinal analysis of the classroom practices of high school mathematics teachers. *Journal of Mathematics Teacher Education*, *19*(6), 575-594.

Corcoran, T. C. (1995). *Transforming professional development for teachers: A guide for state policymakers*. Washington, DC: National Governors' Association.

Cordingley, P., Higgins, S., Greany, T., Buckler, N., Coles-Jordan, D., Crisp, B.,..., & Coe, R. (2015). *Developing great teaching*. London, UK: Teacher Development Trust. Retrieved from https://tdtrust.org/wp-content/uploads/2015/10/DGT-Full-report.pdf

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective Teacher Professional Development*. Palo Alto, CA: Learning Policy Institute.

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]. (2016). Standard for teachers' professional development. London, UK: Department for Education. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/537031/160712_-_PD_Expert_Group_Guidance.pdf

D'Oria, T. (2004). *How I improved my teaching practice in Grade 9 boys' physical education to increase students' participation and enjoyment*. Unpublished Masters thesis, Nipissing University, Canada.

Dunst, C.J., Bruder, M.B., and 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.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., … & Warner, E. (2011). *Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation. Institute of Education Sciences.* Washington, D.C.: Institute of Education Sciences.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., Garrett, R., Yang, R., & Borman, G. D. (2016). *Focusing on mathematical knowledge: The impact of content-intensive teacher professional development (NCEE 2016-4010).* Washington, DC: Institute of Education Sciences. Retrieved from: https://ies.ed.gov/ncee/pubs/20164010/pdf/20164010.pdf

Gibson. D. (2004). Role models in career development: New directions for theory and research. *Journal of Vocational Behaviour*. 65(1), pp.134-156.

Gough, D., Oliver, S., & Thomas, J. (2012). *Introducing systematic reviews*. Sage Publications.

Gough, D., Thomas, J., & Oliver, S. (2012b). Clarifying differences between review designs and methods. *Systematic Reviews*, *1*(1), 28.

Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*(2), 91-108.

Greenleaf, C. L., Litman, C., Hanson, T. L., Rosen, R., Boscardin, C. K., Herman, J., Jones, B. (2011). Integrating literacy and science in biology: Teaching and learning impacts of Reading Apprenticeship professional development. *American Educational Research Journal*, 48, 647–717.

Guskey, T. (2002) Professional Development and Teacher Change. *Teachers and Teaching: theory and practice*, 8(3/4) 381-391

Guskey, T. R. (2003a). Analyzing lists of the characteristics of effective professional development to promote visionary leadership. *NASSP Bulletin*, *87*(637), 4-20.

Guskey, T. R. (2003b). What makes professional development effective? *Phi Delta Kappan*, *84*(10), 748-750.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466-479.

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

Huffman, D., Goldberg, F., & Michelin, M. (2003). Using computers to create constructivist learning environments: Impact on pedagogy and achievement. *Journal of Computers in Mathematics and Science Teaching*, *22*(2), 153-170.

Jacob, R., Hill, H., & Corey, D. (2017). The impact of a professional development program on teachers' mathematical knowledge for teaching, instruction, and student achievement. *Journal of Research on Educational Effectiveness*, *10*(2), 379-407.

Kennedy, M. (2016). How Does Professional Development Improve Teaching?. *Review of Educational Research, 86*(4)*,* 945-980.

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

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

Lally, P., van Jaarsveld, C., Potts, H., & Wardle, J. (2009). How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology, 40*(6), 998-1009.

Lance, L. (2011). Non-production benefits of education: crime, health, and good citizenship. In E.A. Hanushek, S. Machin, & L. Woessman (Eds.), Handbook of the economics of education (pp. 183–282). The Netherlands: Elsevier B.V.

Liberati, A., Moher, D., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*, *6*(7), e1000097.

Mackie, J. L. (1974). *The cement of the universe*. Oxford, UK: Oxford University Press.

Main, K., & Pendergast, D. (2015). Core Features of Effective Continuing Professional Development for the Middle Years: A Tool for Reflection. *RMLE Online*, *38*(10), 1–18.

McDonagh, M., Peterson, K., Raina, P., Chang, S., & Shekelle, P. (2013). Avoiding bias in selecting studies. Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet]. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK47095/

Metcalf, K. K., Vontz, T. S., & Patrick, J. J. (2000). *Effects of Project Citizen on the civic development of adolescent students in Indiana, Latvia, and Lithuania*. Bloomigton, IN: ERIC Clearinghouse for Social Studies. Retrieved from: https://files.eric.ed.gov/fulltext/ED447047.pdf

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

Mill, J. S. (1884). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. London, UK: Longmans, Green, and Co.

Moxon, J. (2003). *A study of the impact of the Restorative Thinking Programme within the context of a large multi-cultural New Zealand secondary school*. Unpublished MA thesis, University of Aukland, New Zealand.

Nugent, G., Kunz, G., Houston, J., Kalutskaya, I., Wu, C., Pedersen, J…, Lee, S., DeChenne, S., Luo, L. Berry, B. (2016). The effectiveness of technology-delivered science instructional coaching in middle and high school. National Center for Research on Rural Education, Institute of Educational Sciences, U.S. Department of Education.

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

Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth science: A comparison of three professional development programs. *American Educational Research Journal*, 48, 996–1025.

Rosenbaum, P. R. (2017). *Observation and experiment: an introduction to causal inference*. Cambridge, MA: Harvard University Press.

Ross, J. A. (1994). The impact of an inservice to promote cooperative learning on the stability of teacher efficacy. *Teaching and Teacher Education, 10*(4), 381-394.

Ross, J. A., Roleiser, C., & Hogaboam-Gray, A. (1999). Effects of collaborative action research on the knowledge of five Canadian teacher-researchers. *The Elementary School Journal, 9*(3), 255-274.

Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, *21*(2), 157–170.

Rutkowski, S., Rutkowski, L., Bélanger, J., Knoll, S… (2013) *Teaching and Learning International Survey 2013 Conceptual Framework*. Paris, France: OCED Publishing.

Schober, H. M. (1984). The effects of inservice training on participating teachers and students in their economics classes. *The Journal of Economic Education*, *15*(4), 282-295.

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

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

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

Sztjan, P., Campbell, M. P., & Yoon, K. S. (2011). Conceptualizing professional development in mathematics: Elements of a model. *PNA, 5*(3), 83–92.

Tasker, G. (2001). *Students' experience in an HIV/AIDS-sexuality education programme: What they learnt and the implications for teaching and learning in health education*. Unpublished doctoral thesis, Victoria University of Wellington, Wellington, New Zealand.

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

van Driel, J. H., Meirink, J. A., van Veen, K., & Zwart, R. C. (2012). Current trends and missing links in studies on teacher professional development in science education: a review of design features and quality of research. *Studies in Science Education*, *48*(2), 129-160.

Walter, C., & Briggs, J. (2012). *What professional development makes the most difference to teachers*. Oxford, UK: Oxford University Press. Retrieved from: http://www.education.ox.ac.uk/wordpress/wp-content/uploads/2010/07/WalterBriggs_2012_TeacherDevelopment_public_v2.pdf

Whitlock, E. P., Lin, J. S., Chou, R., Shekelle, P., & Robinson, K. A. (2008). Using existing systematic reviews in complex systematic reviews. *Annals of Internal Medicine*, *148*(10), 776-782.

Wei, R. C., Darling-Hammond, L., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession*. Washington, DC: National Staff Development Council.

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.* Washington, D.C.: Institute of Education Sciences. Retrieved October 23, 2018, from https://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/REL_2007033.pdf