CENTRE FOR EDUCATION POLICY
& EQUALISING OPPORTUNITIES

Working Paper No. 22-02

Effective teacher professional development: new theory and a meta-analytic test

Sam Sims
UCL

Harry Fletcher-Wood
Ambition Institute

Alison O'Mara-Eves
UCL

Sarah Cottingham
Ambition Institute

Claire Stansfield
UCL

Josh Goodrich
StepLab

Jo Van Herwegen
UCL

Jake Anders
UCL

Multiple meta-analyses have now documented small positive effects of teacher professional development (PD) on pupil test scores. However, the field lacks any validated explanatory account of what differentiates more from less effective in-service training. As a result, researchers have little in the way of advice for those tasked with designing or commissioning better PD. We set out to remedy this by developing a new theory of effective PD based on combinations of causally active components targeted at developing teachers' insights, goals, techniques, and practice. We test two important implications of the theory using a systematic review and meta-analysis of 104 randomised controlled trials, finding qualified support for our framework. While further research is required to test and refine the theory, we argue that it presents an important step forward in being able to offer actionable advice to those responsible for improving teacher PD.

Suggested citation: Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., Van Herwegen, J., Anders, J. (2022). *Effective teacher professional development: new theory and a meta-analytic test?* (CEPEO Working Paper No. 22-02). Centre for Education Policy and Equalising Opportunities, UCL. https://EconPapers.repec.org/RePEc:ucl:cepeow:22-02

Disclaimer

Any opinions expressed here are those of the author(s) and not those of the UCL Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

CEPEO Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Highlights

- We develop a new theory of effective teacher professional development (PD) based on combinations of 14 causally active components targeted at developing teachers' insights, goals, techniques, and practice.
- We test two important implications of the theory using a pre-registered systematic review and meta-analysis of 104 randomized controlled trials.
- We find that each of the 14 causally active components included in a PD programme is associated with a 0.01 SD increase in the effect of the PD on pupil test scores.
- In addition, we find that PD programmes containing at least one causally active component addressing each of insights, goals, techniques and practice have three time the effect of PD programmes that do not. However, this effect is imprecisely estimated.
- The paper represents a step forward in being able to offer actionable advice to those responsible for improving teacher PD.

Why does this matter?

Teachers spend 11 days per year engaged in PD of various forms. The findings presented here can help to ensure that this considerable investment pays dividends in terms of improved teaching and learning.

Effective teacher professional development: new theory and a meta-analytic test

Sam Sims[1], Harry Fletcher-Wood[2], Alison O’Mara-Eves[1], Sarah Cottingham[2], Claire Stansfield[1], Josh Goodrich[3], Jo Van Herwegen[1], Jake Anders[1]

[1] Institute of Education, UCL

[2] Ambition Institute

[3] StepLab

Author Note

Sam Sims https://orcid.org/0000-0002-5585-8202

Alison O’Mara-Eves https://orcid.org/0000-0002-0359-6423

Claire Stansfield https://orcid.org/0000-0002-0718-0409

Jo Van Herwegen https://orcid.org/0000-0001-5316-1818

Jake Anders https://orcid.org/0000-0003-0930-2884

Harry Fletcher-Wood and Sarah Cottingham declare that they work for a charity that provide PD to teachers and schools in return for fees. Alison O’Mara-Eves, Claire Stansfield, Jo Van Herwegen, Sam Sims, and Jake Anders declare that they work for a university that also provides PD to teachers in return for fees. Josh Goodrich works for a commercial organisation StepLab that provides PD software to teachers in return for fees. All authors declare no other conflicts of interest.

Correspondence concerning this article should be addressed to: Sam Sims (S.Sims@ucl.ac.uk), Centre for Education Policy and Equalising Opportunities, UCL Institute of Education, University College London, 20 Bedford Way, London, WC1H 0AL.

Abstract

Multiple meta-analyses have now documented small positive effects of teacher professional development (PD) on pupil test scores. However, the field lacks any validated explanatory account of what differentiates more from less effective in-service training. As a result, researchers have little in the way of advice for those tasked with designing or commissioning better PD. We set out to remedy this by developing a new theory of effective PD based on combinations of causally active components targeted at developing teachers' insights, goals, techniques, and practice. We test two important implications of the theory using a systematic review and meta-analysis of 104 randomised controlled trials, finding qualified support for our framework. While further research is required to test and refine the theory, we argue that it presents an important step forward in being able to offer actionable advice to those responsible for improving teacher PD.

Keywords: professional development, teachers, theory, meta-analysis

Introduction

Effective teachers improve pupil achievement, help close the gaps between rich and poor pupils, and increase pupil earnings in later life (Chetty et al., 2014; Hamre & Pianta, 2005; Slater et al., 2012). Policymakers and educators have therefore invested considerable time and money in trying to enhance the skills of the teaching workforce. As a result, teachers now spend an average of 10.5 days per year attending courses, workshops, conferences, seminars, observation visits, or other types of in-service training (Sellen, 2016). In parallel, governments worldwide have invested billions of dollars in research intended to find out how best to design this teacher professional development (PD; Boulay et al., 2018; Dawson et al., 2018).

This investment has resulted in a marked increase in the number of rigorous studies quantifying the impact of different approaches to teacher PD (Edovald & Nevill, 2021; Hedges & Schauer, 2018). In 2007, a review by Yoon *et al.* found just nine such studies, in 2016 a review by Kennedy found 28 such studies, and in 2019 Lynch *et al.* found 95 such studies focused on science and maths alone. Recent meta-analyses of this literature tend to find average effect sizes of teacher PD on standardized test scores of around 0.06 (Lynch et al. 2019). On average, PD has small positive effects on the quality of teaching, as reflected in pupil learning.

While much has been learned from this evaluation literature, fundamental questions remain. Most schools do not have access to the PD programmes that have so far been evaluated, either because they are not available on the open market, or are too geographically distant, or because of capacity constraints on providers. Moreover, meta-analysis suggests considerable variation in the impact of PD, depending on how the PD is designed (Basma & Savage; 2017; Didion et al., 2020; Kennedy, 2016; Kraft et al., 2018; Lynch et al. 2019). Policymakers and school leaders therefore need to know which characteristics of PD make it effective, so that they can design or commission the best PD available for their teachers (Hill et al., 2013).

Existing attempts to explain what differentiates more from less effective PD have not made much progress. One strand of the literature has employed narrative reviews and thematic analyses in an attempt to identify what differentiates more and less effective PD (Desimone, 2009; Timperley et al., 2007; Wei et al., 2009). Indeed, some of the researchers working in this tradition have even claimed that the field has reached a consensus on the characteristics of effective PD (e.g. Darling-Hammond et al., 2017). However, the narrative reviews on which this claim is based have two important methodological weaknesses. First, many of them include studies employing non-equivalent control groups. Second, they lack any method for differentiating the causally active from causally inactive components of the PD (Sims & Fletcher-Wood, 2020).

A second strand of research has used meta-regression to investigate the associations between different aspects of PD design and impact on pupil outcomes (Basma & Savage; 2017; Didion et al., 2020; Kraft et al., 2018; Lynch et al. 2019). However, there is presently little consensus on which specific characteristics of PD should be entered into such meta-regression models, with different papers testing different mediators (e.g., Kraft et al., 2018; Lynch et al. 2019). Previous research has provided rich ways of conceptualising and categorising PD (Boylan & Demack, 2018; Kennedy, 2016; Opfer & Pedder, 2011 & Sztjan et al. 2011). However, existing theory offers few testable hypotheses about what makes PD more or less effective, which leaves researchers guessing as to how their meta-regression models should be specified and thus how the coefficients should be interpreted.

In sum, we now know about the causal impact of a wide variety of PD programmes, but do not have much useful to say about what differentiates more from less effective PD. In Cummins' words "we are overwhelmed with things to explain, and somewhat underwhelmed by things to explain them with" (Cummins, 2000). In this paper, we set out to remedy this by proposing and empirically testing a new theory of effective teacher PD. Like all theorising, our goal is to *explain why* certain PD designs result in greater impact on teaching and learning. In doing so, we hope to provide a practical theory (Berkman & Wilson, 2021) that suggests actionable steps by which policymakers and school leaders can improve PD design. Our research team, which is composed of researchers and teacher educators, reflects this goal.

In the next section of the paper, we begin by theorising about the four things that PD needs to achieve in order to secure improvements in teaching. The subsequent section then synthesizes a set of mechanisms for achieving each of these four purposes of PD. We also derive some testable implications of this theory. Next, we set out the methods by which we conducted a systematic review and meta-analysis in which we code 104 experimentally evaluated PD programmes for the presence or absence of each of these mechanisms. The results section then presents the findings from a number of meta-analytic tests of our hypotheses.

Theorising how PD fails: Insights, Goals, Techniques, Practice

Practical theory building should begin with a review of research providing rich descriptions of the target problem (Berkman & Wilson, 2021; Scheel et al., 2021). This supports the identification of important concepts, which can then be used as the building blocks of a new theory (Hempel, 1966). We take as our central problem the difficulty of designing PD that results in sustained improvements in practice (Copur-Gencturk & Papakonstantinou, 2016; Hanno, 2021; Hobbiss et al., 2021). In this section, we review a range of descriptive research drawing on data from surveys, longitudinal classroom observations, interviews and diary studies which, taken together, suggest four important building blocks for our framework.

Teachers' knowledge serves as the foundation on which they base decisions about their practice. Mixed methods studies illuminate various ways in which teachers' knowledge influences their practice (Carpenter et al., 1989; Franke et al., 2001; Hill et al., 2008) and measures of teachers' knowledge also correlate with estimates of teacher effectiveness (Hill & Chin, 2018). This suggests that one way in which PD might fail to improve teaching and learning is by failing to bring about changes in teachers' knowledge and understanding. This might happen because the knowledge provided by the PD is inaccurate or irrelevant, or - as longitudinal research with teachers has documented - because new learning tends to be forgotten over time (Arzi & White, 2008; Liu & Phelps, 2020). The first building block of our theory is therefore *insight*, which we define as teachers gaining an enhanced or expanded understanding of teaching and learning.

Knowledge alone is unlikely to bring about changes in practice (Lord et al., 2017; Kennedy, 2016). For example, diary studies have found that teachers report 50% of school-based learning experiences result in changes in their knowledge and beliefs, but in only a quarter of these cases do these changes in beliefs feed through into changes in their intended practice (Bakkenes et al., 2012). A systematic review of studies on formative assessment also found that PD is less likely to feed through into intentions to change practice in the absence of reinforcement from school leaders (Yan et al., 2021). Hence, PD might also fail to improve teaching if it does not motivate teachers to adopt goals around changing their practice. The second building block of our theory is therefore *goals,* which we define as motivating a teacher to consciously pursue a specific change in their practice.

Another point at which PD might fail is around teachers enacting what they have learned in the classroom. For example, a three-year study found that early-career science teachers espoused strong beliefs in the importance and value of student-centred teaching methods but that this was often not reflected in their classroom practice (Simmons et al., 1999). Tightly controlled laboratory studies show that knowledge of classroom management techniques and formative assessment practices is often insufficient to bring about changes in teachers' practice (Cohen & Wiseman, 2019; Cohen et al., 2020). However, when similar teachers are also given feedback on, and practice with, the target skill then this results in improvements in practice (Cohen et al., 2021). PD can therefore also fail when it neglects to provide teachers with the necessary skills. Our third building block is therefore developing *technique,* which we define as helping a teacher to utilize a new teaching practice.

Descriptive research also illuminates the difficulties in embedding change. For example, Copur-Gencturk & Papakonstantinou (2016) collected detailed observations on a group of teachers over a four-year period following a mathematics PD programme. The results show how PD can bring about initial changes in practice but this subsequently fades over time. Other studies have documented similar patterns of 'fade-out'. Boston & Smith (2011) report case studies illustrating how some teachers who implemented cognitively challenging maths instruction immediately after a PD programme no longer did so in a follow-up observation. Similarly, Hanno (2021) uses repeated classroom observation to how some improvements in practice dissipate quickly. The final building block for our theory is therefore embedding *practice,* which we define as supporting a teacher to consistently make use of some technique in the classroom.

In summary, we propose that PD needs to pay careful attention to four things if it is to bring about sustained improvements in teaching practice. First, it needs to provide *insight* (I) about teaching and learning. For example, a teacher might learn that working memory is composed of separate visual-spatial and phonological systems, each of which has limited capacity (Baddeley & Hitch, 1974). Second, PD should motivate teachers to adopt goal-directed (G) changes in practice. For example, a teacher might resolve to limit the cognitive load their exposition of a subject places on either the visual-spatial or the phonological system within working memory. Third, PD should provide *techniques* (T) for putting these insights to work. For example, a teacher might invite pupils to read text from the board in silence, rather than also reading out the text, in order to avoid overloading the phonological loop with both written and aural input. Fourth, PD must embed that change in *practice* (P). For example, a teacher might use the ‘read silently from the board’ technique multiple times, across different classes, until it becomes a routine part of their practice.

Table 1 summarizes our thinking about how PD can fail if (combinations of) these four purposes of PD are not addressed. If PD brings about the necessary changes to I and perhaps G, but not to T and P (row 2 and 3), then this is unlikely to change classroom practice - known in the teacher education literature as the ‘knowing-doing gap’ (Knight et al., 2013). If PD brings about the necessary changes to I, G, and T, but not P, then teachers will tend to revert to established routines (row 4). This reflects the extensive literature on the importance of automaticity and habits in teachers’ practice (Feldon, 2007; Hobbiss, 2021). Finally, if PD brings about the necessary changes to G, T and P, but not I (row 5) then 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 (Kennedy, 2016; Mokyr, 2002), sometimes referred to as a ‘lethal mutation’ in the education literature (Brown & Campione, 1996, p.259). By contrast, we theorize that when PD succeeds in addressing I, G, T and P, it is more likely to be effective.

Theorising how PD succeeds: mechanisms

Having theorized the different ways in which PD might fail, we now turn to consider how PD might successfully address all four of insights, goals, techniques, and practice. Which design features should PD incorporate in order to address all four of these purposes? As previously noted, an important challenge here is in differentiating the causally active from the causally inactive components of a PD design (Mackie, 1974). After all, associations between particular components of PD and the effects of that PD on pupil outcomes could be spurious. Yet a practical theory, capable of providing actionable advice for the design of better PD, requires the associations to reflect an underlying causal relationship. We refer to these causally active components of a PD programme as mechanisms, in that comprise the entities and activities (causally) responsible for bringing about the effects of that PD on teaching and learning (Illari & Williamson, 2012, p.14).

We theorize something to be a causally active component of PD only if we can find causal evidence that it helps achieve I, G, T or P from across multiple domains (Sims & Fletcher-Wood, 2021). Our reasoning here is simple: if a mechanism *x* helps to achieve I, G, T or P in multiple domains beyond teacher PD *and* we also observe an association between the presence of *x* in PD programmes and the impact of those PD programmes, then *x* is likely also a causally active component in PD. This type of reasoning - known as analogical abduction - is commonly used in developing explanatory theories: “if one finds a similar set of phenomena in another field that is better understood, then one can ‘borrow’ explanatory principles from that field to inform one’s own” (Borsboom et al., 2021, p.761). In developing our list of mechanisms, we draw heavily on empirical findings from cognitive science, behavioural science (Michie et al., 2013), and the literature on training medical doctors. We searched the literature for mechanisms that a) have sufficient empirical support across multiple domains and b) provide an explanatory account of *how* they affect I, G, T or P.

With respect to *insight* (I) - teachers gaining an enhanced understanding of teaching and learning - we found two such mechanisms. The first is to *manage the cognitive load* for the teachers taking part in the PD. This can be achieved by focusing on a single idea or task, removing redundant information, or by providing worked examples, all of which help to prevent working memory from becoming overloaded. For causal evidence that this helps with learning new material among school students and adult medical trainees, see the reviews by Sweller et al. (2019) and Fraser et al. (2015). The second mechanism is *revisit material,* which can be achieved by reteaching or prompting recall of important ideas on separate occasions, both of which help to strengthen memory. Causal evidence that this aids with learning in lab settings, as well as in history, maths and language learning at school, can be found in the reviews by Adesope et al. (2017), Rohrer (2015), and Yang (2021).

As regards *goals* (G) - motivating a teacher to pursue a specific change in their practice - we found three putative mechanisms, all of which were taken from Michie et al. (2013). The first is to go through an explicit *goal setting* process, in which teachers consciously agree on an objective around changing a specific part of their practice. This works by directing attention and energy toward the target change (Locke & Latham, 2002). Epton et al. (2017) provides a review of evidence that goal setting brings about change in sporting, health-related and educational settings. The second mechanism is to present evidence supporting the change from a *credible source,* by which we mean findings from empirical research. For reviews of evidence that statistical evidence or justified arguments help change people's minds and intentions in setting including health, crime and education, see O'Keefe (1998) and Hornikx (2005). The third mechanism is *reinforcement,* which can be achieved through praising or restating the value of a certain teaching practice. This has been shown to increase motivation in domains including arts, games and maths (Delin & Baumeister, 1998).

With respect to *technique* - helping a teacher to utilize a new teaching practice - we found five mechanisms that met our criteria: instruction, practical social support, modelling, feedback, and rehearsal (Michie et al., 2013). *Practical social support* involves arranging advice on how to implement a practice from a teacher's colleagues. Causal studies show that this supports practice change in medical training (Grierson et al., 2012) and in various health behaviour settings (Dale et al., 2012; Jolly et al., 2012; Ramchand et al., 2017). *Modelling* involves providing an observable example of the target teaching practice, which provides a visual guide for subsequent practice (Renkl, 2014). Many experimental studies in the medical education literature have found that modelling helps with acquisition of new clinical

(Cordovani & Cordovani, 2016) and surgical skills (Harris et al., 2018). The remaining three *techniques* mechanism are *instruction, feedback,* and *rehearsal* and (for space reasons) are discussed in full in Appendix A.

Finally, with respect to *embedding practice* - supporting a teacher to consistently make use of some technique - we found four potential mechanisms that met our criteria (Michie et al., 2013). *Action planning* involves specifying when and how a change in practice will be made in a future lesson. This creates situational cues that help trigger new practice (Webb & Sheeran, 2008) and has been shown to help change practice in health, education and lab settings (Gollwitzer & Sheeran, 2006). *Context specific repetition* refers to rehearsing the target practice in a realistic classroom setting. This helps overwrite existing cue-response relationships (habits) by re-associating the classroom setting with the new practice (Hobbiss et al., 2021). Experimental studies have shown that rehearsal in realistic simulators for surgical trainees (even without feedback from an observer) leads to improved practice on a delayed post-test (Andreatta et al., 2006; Van Sickle et al., 2008). Experimental studies have also found that interventions focused on overwriting old habits can help embed health behaviour change (Carels, 2011). The remaining two *practice* mechanisms - *prompts/cues* and *self-monitoring* - are discussed full in Appendix A.

Table 2 summarizes the mechanisms across the four (IGTP) purposes of PD. Three clarificatory points are in order. First, while we have searched extensively, and have included every mechanism for which we could find sufficient supporting evidence, this list is unlikely to be complete. Indeed, even if we have identified every relevant mechanism documented in the existing literature, future research may identify additional relevant mechanisms. Second, mechanisms within each row of the table can be thought of as substitutes for each other, in that they achieve the same thing. However, they are also likely to have a cumulative effect. For example, incorporating managing cognitive load and revisiting prior learning in a single PD programme would likely contribute more to increased *insight* than only incorporating revisit material. Third, we make no assumptions about the size of the effects of the different mechanism. Our argument is only that improvements in e.g. *technique* are an increasing function of the number of technique mechanism incorporated in a given PD programme. This assumption is formalized in Appendix B.

Hypotheses

Having set out our theory, we now derive two hypotheses that will be tested in the remaining, empirical sections of the paper. Since we theorize that the fourteen mechanisms listed in Table 2 are all causally active components of PD with cumulative effects, we hypothesize that:

> H1: The number of mechanisms incorporated in PD programmes will be positively correlated with the impact of those PD programmes on pupil test scores.

In addition, since we theorized that PD is likely to be more effective if it addresses all four purposes of PD, we hypothesize that:

> H2: PD programmes that incorporate at least one mechanism in each of the four I/G/T/P categories (a 'balanced design') will have a larger impact on pupil test scores.

Our first hypothesis and the definition of a balanced design are formalized in Appendix B.

Methods

Systematic Review

We systematically searched the literature to identify primary research studies that could be used to test these hypotheses. We included studies in our meta-analysis if they met all of the following criteria: 1) they focused on qualified teachers working in formal education settings with children 3-18 years of age; 2) they evaluated a teacher PD programme, defined as structured, facilitated activity intended to improve their teaching ability; 3) the evaluation employed a randomized controlled trial (RCT) design, thus allowing clean causal inference; 4) the control group in the RCT received either business as usual or no PD; 5) the evaluation measured outcomes using a standardized (not researcher designed) test score outcome, to increase the comparability of effect size estimates (Cheung & Slavin, 2016); 6) the study was published during or after 2002; 7) the study was written in English and conducted in an OCED country.

We employed various combinations of search terms intended to capture three main concepts: (1) teachers (e.g. 'teachers', 'educators'); (2) professional development (e.g. 'in-service training', 'professional learning'); and (3) randomized controlled trials (e.g. 'RCT').[1] We used these terms to query eleven different databases and search engines during November 2020.[2] In addition, we searched the reference lists of eleven previous reviews,[3] employed reference-checking and forward citation searching of included studies,[4] and browsed eight websites containing education research repositories.[5] All records were uploaded into the *EPPI Reviewer* software, deduplicated and then screened on title and abstract using prioritized screening (O'Mara-Eves et al., 2015; Thomas et al., 2011).[6] All studies included at this stage were then reviewed in full. This process resulted in 121 eligible experimental studies (see the PRISMA flow diagram in Appendix C for further details).

We extracted Cohen's *d* effect sizes for each of the studies in our sample using the formulae from Lipsey & Wilson (2001). This was possible for 104 or the 121 studies. Cohen's *d* is known to display small bias in small studies and can be corrected using Hedges' *g* (Hedges, 1981). However, Hedges' *g* could not be calculated for two of the 104 studies due to missing data. We therefore present all our results using Cohen's *d*, on the basis that losing studies from the meta-analysis is highly undesirable.[7] In cases where eligible studies reported multiple standardized test score outcomes, we selected the primary tests core outcome (if specified), or else collected all standardized test score outcomes. Contour plots, trim-and-fill and p-curve analysis all suggested either zero or small publication bias.[8]

In addition to effect sizes, we coded the studies based on whether the PD incorporated each of our fourteen mechanisms. In cases where eligible studies reported evaluations of multiple versions of a PD programme, we focused on the most intensive version.[9]
Then two authors (SS and HFW) double coded 46 papers using this coding frame and achieved 82% agreement at the mechanism level. The two coders met to discuss discrepancies until consensus was reached. The coding frame was then revised to further eliminate ambiguity and to support consistent coding.[10] The remaining papers were then coded for mechanisms by a single author (HFW). In our empirical analysis, we test the sensitivity of our main results to the presence of measurement error using errors-in-variables regression, using the 82% figure as the best available assumption for the reliability with which our mechanisms are measured.

Figure 1 provides descriptive statistics about the mechanisms. The left-hand panel shows the number of mechanisms per PD programme. We observe a minimum of zero mechanisms (in just one PD programme) and a maximum of 13 (again in just one PD programme). The median number of mechanisms is five and there is a long right tail of mechanism-rich programmes. The right-hand panel shows the frequency with which each of the 14 mechanisms occur. All of our mechanisms occur at least once, with prompts/cues being the least common and instruction being the most common. The *techniques* (T) mechanisms are the most frequently occurring.

We collected three further types of information from each study. First, we coded for the geographic location, age group and subject focus for each experiment. Second, we coded for the 'broad area of focus' of the PD, based on whether the content of the PD was largely based on cognitive science, formative assessment, inquiry learning, or data-driven instruction.[11] Second, we coded for four important indicators of study quality: whether the experiment was pre-registered; whether the RCT met the *What Works Clearinghouse* ‘cautious’ standards for acceptable attrition; whether the study randomized more than 50 units to treatment and control; and whether the study employed a high-stakes test score outcome (Cheung & Slavin, 2016).[12] Table 3 summarizes the characteristics of the 104 studies included in our meta-analytic sample.

Meta-analytic tests

To calculate meta-analytic average effect sizes, we used robust variance estimation (RVE) random effect meta-analysis (Hedges, Tipton & Johnson, 2010; Tanner-Smith & Tipton, 2013). It was not possible to regress the effect sizes on all 14 mechanisms separately due to sample size constraints, further compounded by likely interactions between the various mechanisms. To test H1, we therefore plot the impact of all 104 PD programmes on test scores (expressed as an effect size) against the number of mechanisms per programme, and then add a meta-regression (precision weighted) line of best fit. These plots have the advantage of conveying more information about the underlying data than meta-regression tables. Since it is not possible to produce these plots using RVE, we use the primary outcome (if specified), or else one randomly chosen outcome per PD programme. Where the results of the RVE analysis are qualitatively different, we highlight this in the text. We then repeat this analysis a number of times, stratifying the data based on the broad content area of the PD and various indicators of study quality. One important caveat about these plots is that the experimental impact estimates on the Y axis all contain random (classical) measurement error. This artificially increases the variance on the y axis and, by extension, reduces the proportion of variance explained by the model. To test H2 we simply plot the interval estimates using RVE meta-analysis and all the standardized test score outcomes.

The mechanism incorporated in each PD programme in our sample are not themselves randomly assigned, meaning that our meta-analysis cannot estimate the causal effects of those mechanisms. So how can our study provide actionable advice to educators looking to improve the design of PD? When asked how observational studies can move beyond correlation to causation, Fisher advised researchers to "make their theories elaborate". The rationale for this is that empirical corroboration of a complicated pattern of predictions helps rule out alternative explanations (quoted in Rosenbaum, 2005, p. 8). With respect to H1, our theory synthesizes empirical evidence that our mechanisms are causally active in a range of other domains, which makes it less plausible that an association within the domain of PD programmes is spurious. H2 is also elaborate in the Fisherian sense. Why else would a PD programme with at least one mechanism in each of the I/G/T/P categories be more effective than e.g. a programme with at least one mechanism in three but not four of the categories? We also pre-registered both of our hypotheses prior to data collection, making the subsequent empirical analysis a genuinely risky test of our theory (Mayo, 2018).[13]

Results

Our first test of H1 can be found in Figure 2, which plots the number of mechanisms against the impact estimate (scaled as an effect size) for all 104 PD programmes in our sample. Larger circles representing studies with more precise estimates, with the size being proportional to the weight they are given in the analysis. The meta-regression line of best fit is upward sloping ($\beta = 0.01$, $p = .02$). PD interventions incorporating zero mechanisms have an expected effect size close to zero and PD mechanisms incorporating 13 mechanisms have an expected effect size close to .15. For context, the average effect size in our sample is 0.05 ($p < 0.001$), implying the number of mechanisms incorporated in PD can account for variation equivalent to three times the average effect.

Figure 2 suggest a considerable degree of unexplained variation. Figure 3 therefore stratifies the analysis based on the broad content area of the PD. The proportion of variance explained doubles from 16% to 36% - although both of these will be underestimates due to classical measurement error on the y axis. The gradient for formative assessment increases to

.02 but is no longer statistically significant at conventional levels (p = .09). The gradient for inquiry also increases to 0.03 (p=.046). However, the relationship among PD programmes focused on data-driven instruction breaks down entirely (β = -0.01, p = .64), albeit in a sample of just seven studies. For context, the average impact of PD focused on formative assessment (d = .04, p = .08) and data-driven instruction are not significantly different from zero in general (d = .04, p = .08). By contrast, the average impact of PD focused on inquiry is positive (d = .07, p = .01).

Our final analysis relating to H1 is to test the sensitivity of the hypothesized relationship to various indicators of study quality and treatment heterogeneity (Figure 4). We find a similar relationship among studies using high-stakes test scores (Panel 1, β = .01, p = .04), among large trials (Panel 3, β = .01, p = .04) and among PD programmes that do not include sets of new curriculum materials (Panel 5, β = .01, p = .03). We also find a very similar gradient in our errors in variables regression (Panel 6, β = .01, p = .12) and in studies with low attrition (Panel 2, β = .01, p = .06). These last two results are no longer statistically significant at conventional levels, however the result for attrition is significant when estimated using RVE (β = .02, p = .02), which suggests it is marginal.

The most concerning part of Figure 4 is the panel for pre-registered studies, in which both the gradient and p value break down (Panel 4, β = .004, p = .32). In principle, the absence of a relationship among pre-registered studies could be explained by: p-hacking in trials that are not pre-registered; otherwise higher methodological standards in pre-registered trials; or inferior selection of PD programmes by the types of funders that require pre-registration. We probe these potential explanations further in Appendix F. Our p-curve analysis does not indicate motivated p hacking in our sample (Simonsohn et al., 2014a; Simonsohn et al., 2014b). Our comparison of methodological standards shows that pre-registered trial are indeed more likely to use high-stakes test score outcomes and have lower attrition (indicators of higher methodological standards). We also find some evidence that pre-registered PD is slightly less well designed, as indicated by the number of mechanisms that they incorporate.

We now turn to our empirical tests of H2. Figure 5 shows the average impact of five separate groups of PD. On the left are PD programmes that incorporate mechanisms addressing all four I/G/T/P purposes of PD ('balanced designs'). To the right of that are PD programmes that incorporate mechanism(s) addressing three or fewer purposes of PD ('imbalanced designs'). PD with an imbalanced design has an average impact of .05, regardless of whether it addresses 1, 2, or 3 purposes of PD. By contrast, PD with a balanced design has an average impact of .15 ($p = .03$). However, the 95% confidence interval for balanced PD programmes is wide and overlaps with the confidence interval for all imbalanced designs ($p = .22$). There are two reasons that the confidence interval is much wider on the balanced PD plot. First, there are fewer studies in the balanced (n=9) versus imbalanced plots (n=95). Second, there is more heterogeneity – captured by the standard deviation of the effect sizes (τ) among the effect sizes in the balanced ($\tau = .1$) versus imbalanced plots ($\tau = .05$). We return to this point in the discussion section.

Our theoretical framework encodes a set of assumptions about which mechanisms address which of the four purposes of PD. While we pre-registered the overall I/G/T/P framework, and we believe the match between mechanism and purposes to be well-grounded in theory, we did not pre-register our list of mechanisms. Figure 6 therefore checks the sensitivity of our findings to reallocating mechanisms in four cases where our assumptions might be arguable. First, we reallocated the feedback mechanism to the insight (I) purpose. Second, we reallocated the credible source mechanisms to the insight (I) purpose. Third, we reallocated the praise/reinforce mechanism to the embed practice (P) purpose. Fourth, we reallocated the context-specific repetition mechanism to the techniques (T) purpose. The results are qualitatively similar to those in Figure 5, suggesting that our findings are not particularly sensitive to these assumptions.

Discussion

We set out to develop and test a practically useful theory of effective teacher PD. To do so, we developed an account of four ways in which PD might fail to bring about sustained improvements in teaching practice. Against this, we synthesized a set of mechanisms hypothesized to be causally active in addressing each of these four purposes of PD. How successful has our research been in achieving this goal? Various frameworks have been suggested for evaluating theories (Kuhn, 1977; Gawronski & Bodenhausen, 2015; Van Lange, 2015). While there are differences in emphasis and language, all of these frameworks emphasize the importance of: abstraction/parsimony; plausibility/coherence; explanatory power; usefulness/applicability; and progress/fruitfulness. In the remainder of the paper, we assess the strength and limitations of our theoretical framework against these criteria.

Parsimony requires that theories abstract away from empirical detail, using the fewest assumptions or components necessary to explain the target phenomenon (Gawronski & Bodenhausen, 2015; Eronen & Bringmann, 2021). Our top-level framework involves just four components: insights, goals, techniques, practice. As set out in Table 1, this simple set-up allows us to account for a range of phenomena documented in the PD literature, including the knowing/doing gap, the importance of habits, and lethal mutations. Within each of the I/G/T/P categories, our framework includes between two and five mechanisms, which collectively allowed us to characterize and capture considerable variation in our set of 104 experimentally evaluated PD interventions. We found some instances of all fourteen mechanisms and one intervention containing 13 of these 14 mechanisms, suggesting the framework is not overly elaborate relative to current PD design. We were also able to formalize several aspects of our framework (Appendix B).

The coherence and plausibility of a theory depends on its degree of fit with existing knowledge or other, well-corroborated, theory (Scheel et al., 2021). An important and distinctive feature of our framework is the requirement that each mechanism be supported by empirical causal evidence from multiple domains. Our theory is therefore closely integrated with empirical findings from the health psychology, cognitive science and medical education literatures. It also builds on cognate theoretical frameworks (e.g. Michie et al., 2013). While we were careful to only include mechanisms for which we found sufficient supporting empirical evidence, we acknowledge that the strength of the evidence varies across our mechanisms. Some mechanisms have strong, direct supporting evidence from very many settings (e.g. goal setting), while others have evidence from fewer domains (e.g. rehearsal) or from fewer good studies (e.g. credible source). In future, mechanism should be added if basic research discovers sufficient evidence for new mechanisms; or removed if new research brings into doubt the evidence on which they are currently included.

A theory has explanatory power if it can provide an accurate account of how and why something occurred by citing earlier events (Cummins, 2000; Elster, 2015). Our theory achieves this in two senses. In a qualitative sense, it provides an account of how PD succeeds or fails to improve teaching practice via changes in I/G/T/P, brought about by the mechanisms incorporated in the PD. We were careful to provide such an account both for how different combinations of I/G/T/P affect teaching practice and for how each individual mechanism affects I/G/T/P. In a quantitative or statistical sense, our meta-analysis showed that the number and combination of mechanisms incorporated in the PD can explain variation in effects between 0 and 0.15 standard deviations – a range equivalent to three times the average impact of PD. Crucially, we argue that the persuasiveness of our account derives from the *combination* of these two types of evidence: independent causal evidence that each of our mechanisms is causally active in various other domains, plus evidence of an association between PD incorporating those mechanisms and the impact on test scores.

Having said that, we acknowledge that our empirical findings come with two important caveats. The first relates to the wide confidence intervals on the estimate for balanced designs. This is likely largely due to the smaller number of evaluations (n=9) for PD with a balanced design. Further experimental evaluations of PD are therefore needed in order to provide a more precise test of this hypothesis. The second caveat relates to the absence of a statistically significant relationship between the number of mechanism and the impact of the PD among the subset of pre-registered studies. Our additional analysis in Appendix F suggests that this likely reflects greater use of high-stakes test scores, lower attrition, and slightly weaker PD designs among pre-registered evaluations, but not *p*-hacking. Assuming this is correct, we would expect this to reduce the gradient we observe by shrinking variation on the y axis. This is broadly consistent with our finding that PD evaluations that are pre-registered (d = .01) have lower effect sizes than PD evaluations in general (d =.05; Appendix E) and with similar findings from the broader education literature (Kraft, 2020).[14] Our theory could be further stress-tested here by conducting pre-registered A/B tests in which the same PD content is delivered using low-mechanism and high-mechanism designs.

For a theory to be practical it should point towards actionable steps for solving a real-world problem (Berkman & Wilson, 2021). The weight of the evidence presented here suggests that PD incorporating more mechanisms should be favoured over PD incorporating fewer mechanisms on the grounds that it is more likely to be effective, other things equal. Likewise, it seems hard to explain the pattern of results found in Figure 5, other than by the importance of ensuring that PD addresses all four of insights, goals, techniques and practice. However, the imprecision of this finding means this latter recommendation should be kept under close review as further evaluations of PD using a balanced design are conducted.

The theory may also be useful for researchers in helping to express with clarity what is involved in the PD programmes that they evaluate. An important limitation of our analysis is that we did not achieve agreement on 18% of the mechanism codes for the studies in our sample. Looking beyond our study, this is clearly problematic, since attempts to scale-up or imitate successful interventions requires clarity on the causally active components of the PD programme. While tools have been developed to aid more precise descriptions of interventions in the medical literature (Hoffman et al., 2014) our review highlights that the education literature still has some way to go in this respect. Using our framework to describe the design of PD in evaluation reports would be a step forward in helping researchers increase the precision with which they report the likely causally active components – thus reducing ambiguity where ambiguity matters most.

For a theory to be fruitful – our final evaluative criteria – it should suggest avenues and hypotheses for future research (Ivani, 2018; Van Lange, 2015). We have tested two such hypotheses here and these should of course be tested further as new experimental evaluations using standardized test scores are published. As an auxiliary hypothesis, we suggest that researchers aim to achieve at least 82% item-level agreement when coding new PD evaluations prior to using this data in future tests. We look exclusively at test score outcomes in this analysis, however our theory also makes a number of predictions about intermediate outcomes (see Appendix B). Future tests could therefore also address whether the number of mechanisms addressing how e.g. Insight or Techniques predict measured changes in teacher knowledge or practice.

In conclusion, we submit that the I/G/T/P theory represents an important advance over existing theories of effective PD and, notwithstanding some important caveats in our empirical results, provides what is now the best-corroborated, genuinely *explanatory* account of what differentiates more and less effective teacher PD (Haig, 2009).

Acknowledgements: This research was supported by the Education Endowment Foundation.

References

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

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

Andreatta, P. B., Woodrum, D. T., Birkmeyer, J. D., Yellamanchilli, R. K., Doherty, G. M., Gauger, P. G., & Minter, R. M. (2006). Laparoscopic skills are improved with LapMentor™ training: results of a randomized, double-blinded study. *Annals of Surgery*, *243*(6), 854.

Arzi, H. J., & White, R. T. (2008). Change in teachers' knowledge of subject matter: A 17-year longitudinal study. *Science Education*, *92*(2), 221-251.

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

Bakkenes, I., Vermunt, J. D., & Wubbels, T. (2010). Teacher learning in the context of educational innovation: Learning activities and learning outcomes of experienced teachers. *Learning and instruction*, *20*(6), 533-548.

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.

Berkman, E. T., & Wilson, S. M. (2021). So useful as a good theory? The practicality crisis in (social) psychological theory. *Perspectives on Psychological Science*, 16(4) 864–874.

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

Boston, M. D., & Smith, M. S. (2011). A 'task-centric approach'to professional development: Enhancing and sustaining mathematics teachers' ability to implement cognitively challenging mathematical tasks. *ZDM*, *43*(6-7), 965-977.

Boulay, B., Goodson, B., Olsen, R., McCormick, R., Darrow, C., Frye, M., ... & Sarna, M. (2018). The Investing in Innovation Fund: Summary of 67 Evaluations. Final Report. NCEE 2018-4013. *National Center for Education Evaluation and Regional Assistance*.

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

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

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

Carels, R. A., Young, K. M., Koball, A., Gumble, A., Darby, L. A., Wagner Oehlhof, M., ... & Hinman, N. (2011). Transforming your life: An environmental modification approach to weight loss. *Journal of Health Psychology*, *16*(3), 430-438.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C. P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal*, *26*(4), 499-531.

Charalambous, C., & Hill, H. (Eds.). (2012). Teacher knowledge, curriculum materials, and quality of instruction: Unpacking a complex relationship [Special section]. *Journal of Curriculum Studies*, *44*(4), 443–576.

Chetty, R., Friedman, J., & Rockoff, J. (2014). Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633–79.

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

Cohen, J., & Wiseman, E. (2019). *Approximating complex practice: Teacher simulation of text-based discussion.* Paper presented at the annual meeting of the Association for Public Policy Analysis and Management, Denver, CO.

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

Cohen, J., Krishnamachari, A., & Wong, V. C. (2021). Experimental Evidence on the Robustness of Coaching Supports in Teacher Education.

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

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.

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

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

Cummins, R. (2000). "How does it work?" versus "What are the laws?": Two conceptions of psychological explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 117–144). The MIT Press

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

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute.

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

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

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

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

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.

Edovald, T., & Nevill, C. (2021). Working out what works: The case of the Education Endowment Foundation in England. *ECNU Review of Education*, *4*(1), 46-64.

Elster, J. (2015). *Explaining social behavior: More nuts and bolts for the social sciences*. Cambridge University Press.

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

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, 1745691620970586.

Franke, M. L., Carpenter, T. P., Levi, L., & Fennema, E. (2001). Capturing teachers' generative change: A follow-up study of professional development in mathematics. *American educational research journal*, *38*(3), 653-689.

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

Gawronski, B., & Bodenhausen, G. V. (2015). *Theory and explanation in social psychology*. Guilford Publications.

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

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

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

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

Haig B. D. (2009) Inference to the best explanation: A neglected approach to theory appraisal for psychology. *American Journal of Psychology, 122*, 219-234.

Hamre, B., & Pianta, R. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure?. *Child Development, 76*(5), 949–67.

Hanno, E. C. (2021). Immediate changes, trade-offs, and fade-out in high-quality teacher practices during coaching. *Educational Researcher*, 0013189X211062896.

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

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

Hedges, L. V. (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.

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

Hempel, C. G. (1966). *Philosophy of natural science*. Prentice-Hall.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, *26*(4), 430-511.

Hill, H. C., & Chin, M. (2018). Connections between teachers' knowledge of students, instruction, and achievement outcomes. *American Educational Research Journal*, *55*(5), 1076-1112.

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

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., ... & Michie, S. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*, *348*, 348:g1687.

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

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.

Ivani, S. (2019). What we (should) talk about when we talk about fruitfulness. *European Journal for Philosophy of Science*, *9*(1), 1-18.

Ivers, N., Jamtvedt, G., Flottorp, S., Young, J. M., Odgaard-Jensen, J., French, S. D., ... & Oxman, A. D. (2012). Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews*, (6).

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

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

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

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.

Kuhn, T. S. (1977). *The essential tension*. University of Chicago Press.

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.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705.

Lord, P., Rabiasz, A., Roy, P., Harland, J., Styles, B., & Fowler, K. (2017). Evidence-Based Literacy Support: The" Literacy Octopus" Trial.. *Education Endowment Foundation*.

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.

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

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

Mayo, D. G. (2018). *Statistical inference as severe testing: how to get beyond the statistics wars*. Cambridge University Press.

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

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

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

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), 1-22.

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

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

Feldon, D. F. (2007). Cognitive load and classroom teaching: The double-edged sword of automaticity. *Educational Psychologist*, *42*(3), 123-137.

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

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

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.

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744-755.

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

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

Simmons, P. E., Emory, A., Carter, T., Coker, T., Finnegan, B., Crockett, D., ... & Labuda, K. (1999). Beginning teachers: Beliefs and classroom actions. *Journal of Research in Science Teaching*, *36*(8), 930-954.

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

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

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

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Van Herwegen, J., & Anders, J. (2021). *What are the characteristics of effective teacher professional development? A systematic review and meta-analysis*. Education Endowment Foundation.

Slater, H., Davies, N., & Burgess, S. (2012). Do teachers matter? Measuring the variation in teacher effectiveness in England. *Oxford Bulletin of Economics and Statistics, 74*, 629–45.

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

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

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

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

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

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

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, *85*(4), 475-511.

Van Lange, P. A. (2013). What we should expect from theories in social psychology: Truth, abstraction, progress, and applicability as standards (TAPAS). *Personality and Social Psychology Review*, *17*(1), 40-55.

Van Sickle, K. R., Ritter, E. M., Baghai, M., Goldenberg, A. E., Huang, I. P., Gallagher, A. G., & Smith, C. D. (2008). Prospective, randomized, double-blind trial of curriculum-based training for intracorporeal suturing and knot tying. *Journal of the American College of Surgeons*, *207*(4), 560-568.

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

Webb, T. L., & Sheeran, P. (2008). Mechanisms of implementation intention effects: The role of goal intentions, self-efficacy, and accessibility of plan components. *British Journal of Social Psychology*, *47*(3), 373-395.

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

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice*, *28*(3), 228-260.

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

Tables

Table 1

Summary of how PD can fail to bring about sustained improvements in teaching and learning

Instil Insight (I)	Motivate Goals (G)	Develop Techniques (T)	Embed Practice (P)	Consequences
✓	✓			Knowing-doing gap
✓				Knowing-doing gap
✓	✓	✓		Revert to established habits
	✓	✓	✓	Misapplication
✓	✓	✓	✓	More likely to be effective

Table 2

Combining the mechanism and IGTP

Purpose	Mechanism
Instil insight (I)	1. Manage cognitive load 2. Revisit prior learning
Motivate goals (G)	3. Goal setting 4. Credible source 5. Praise/reinforce
Teach techniques (T)	6. Instruction 7. Practical social support 8. Modelling 9. Feedback 10. Rehearsal
Embed practice (P)	11. Prompts/cues 12. Action planning 13. Self-monitoring 14. Context-specific repetition

Table 3

Descriptive statistics for the meta-analytic sample

Characteristics	Count	Proportion
Location		
USA	73	70.2%
UK	25	24.0%
Other	6	5.8%
Age group		
Early years/Pre-kindergarten	29	27.9%
Primary/Elementary	52	50.0%
Middle/Secondary/High	28	26.9%
Subject targeted		
Literacy/first language	52	50.0%
Maths	30	28.9%
Science	12	11.5%
Other subjects	6	5.8%
Cross-curricular	17	16.4%
Broad area of focus		
Cognitive science	1	1.0%
Inquiry	16	15.4%
Formative assessment	14	13.5%
Data-driven instruction	7	6.73%
Pre-registered		
Yes	26	25.0%
No	78	75.0%
What Works Clearinghouse Attrition		
Acceptable	36	34.6%
Unacceptable/Unclear	68	65.4%
Number of units randomized		
>50	64	61.5%
≤50	40	38.5%
Test type		
High-stakes standardized	29	27.9%
Low-stakes standardized	75	72.1%
Total:	104	100%

Note. Percentages may not sum to 100 within cells, and counts may not sum to 104 within cells, due to rounding or due to sub-categories not being exhaustive.

Figures

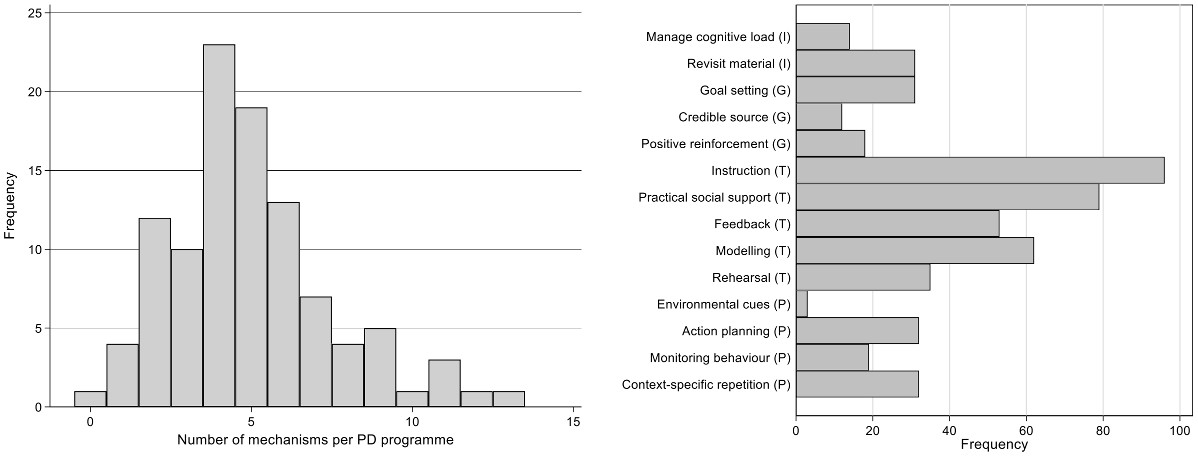


FIGURE 1. *Mechanisms descriptive statistics. N=104 PD programmes*

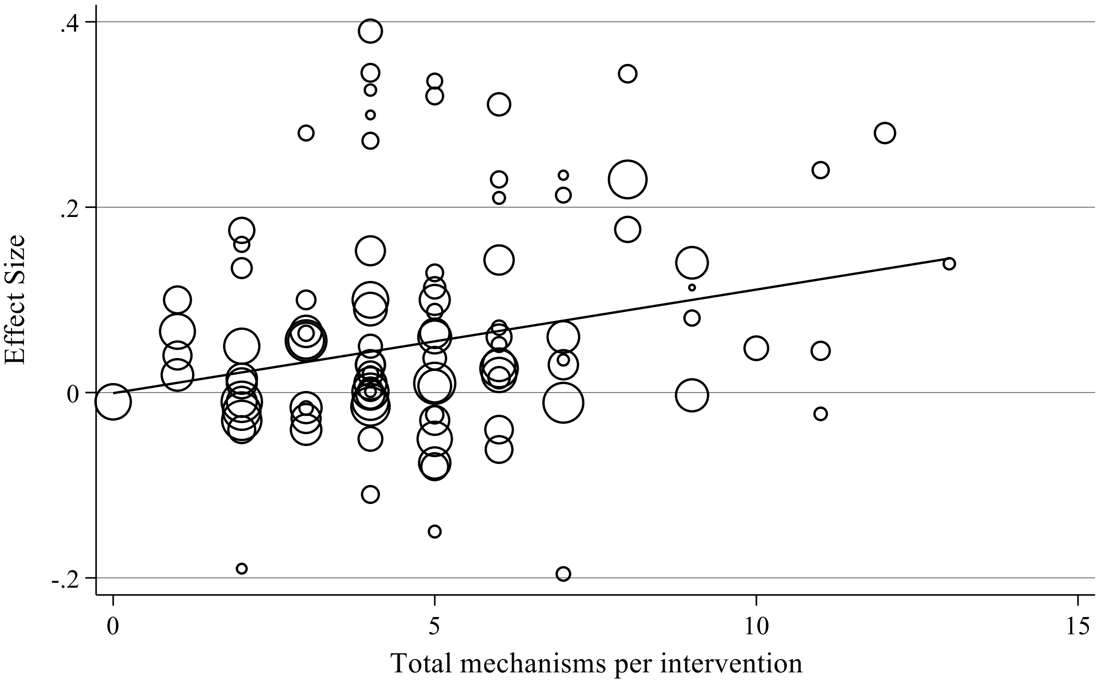


FIGURE 2. *Relationship between the number of mechanisms in a PD programme and impact on pupil test scores*

Note. n = 104 studies. Uses the primary outcome as specified in the study or else one randomly selected outcome per study. Effect sizes >.5 or <-0.2 are used in the underlying meta-regression but are not shown in the figure to aid visual clarity.

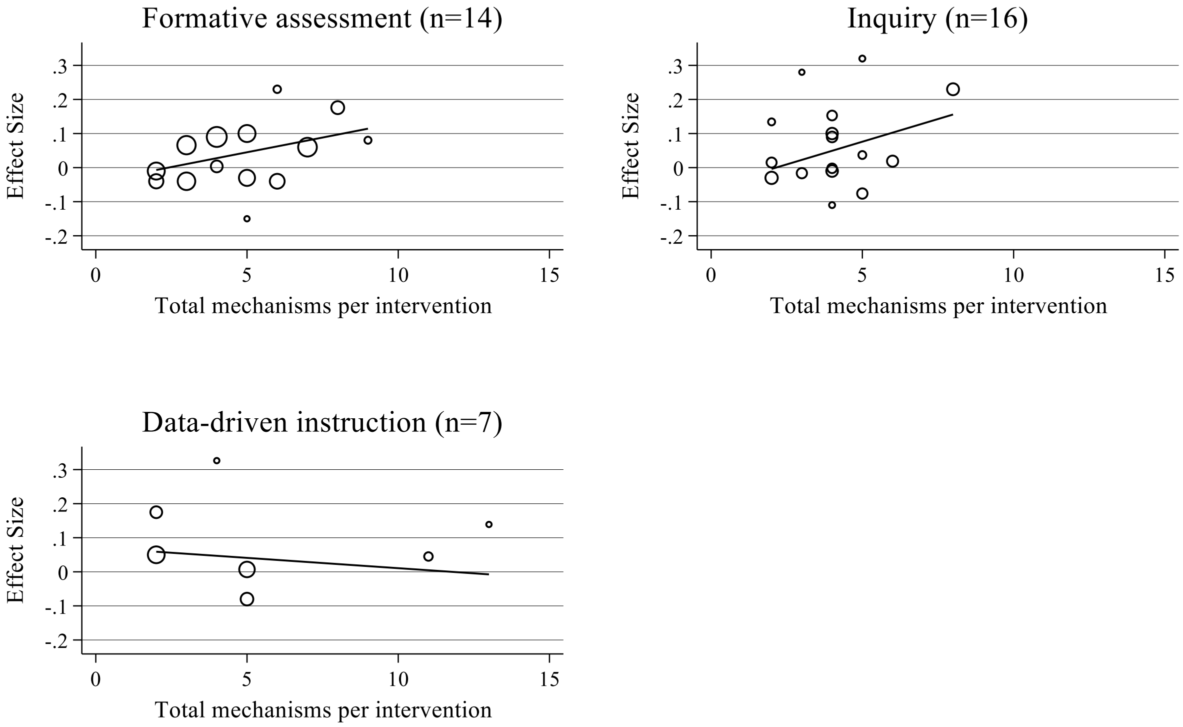


FIGURE 3. *Relationship between the number of mechanisms in a PD programme and impact on pupil test scores, by content area of the PD*

Note. N = number of separate experimental studies. Uses the primary outcome as specified in the study or else one randomly selected outcome per study. Effect sizes >.5 or <-0.2 are used in the underlying meta-regression but are not shown in the figure to aid visual clarity.

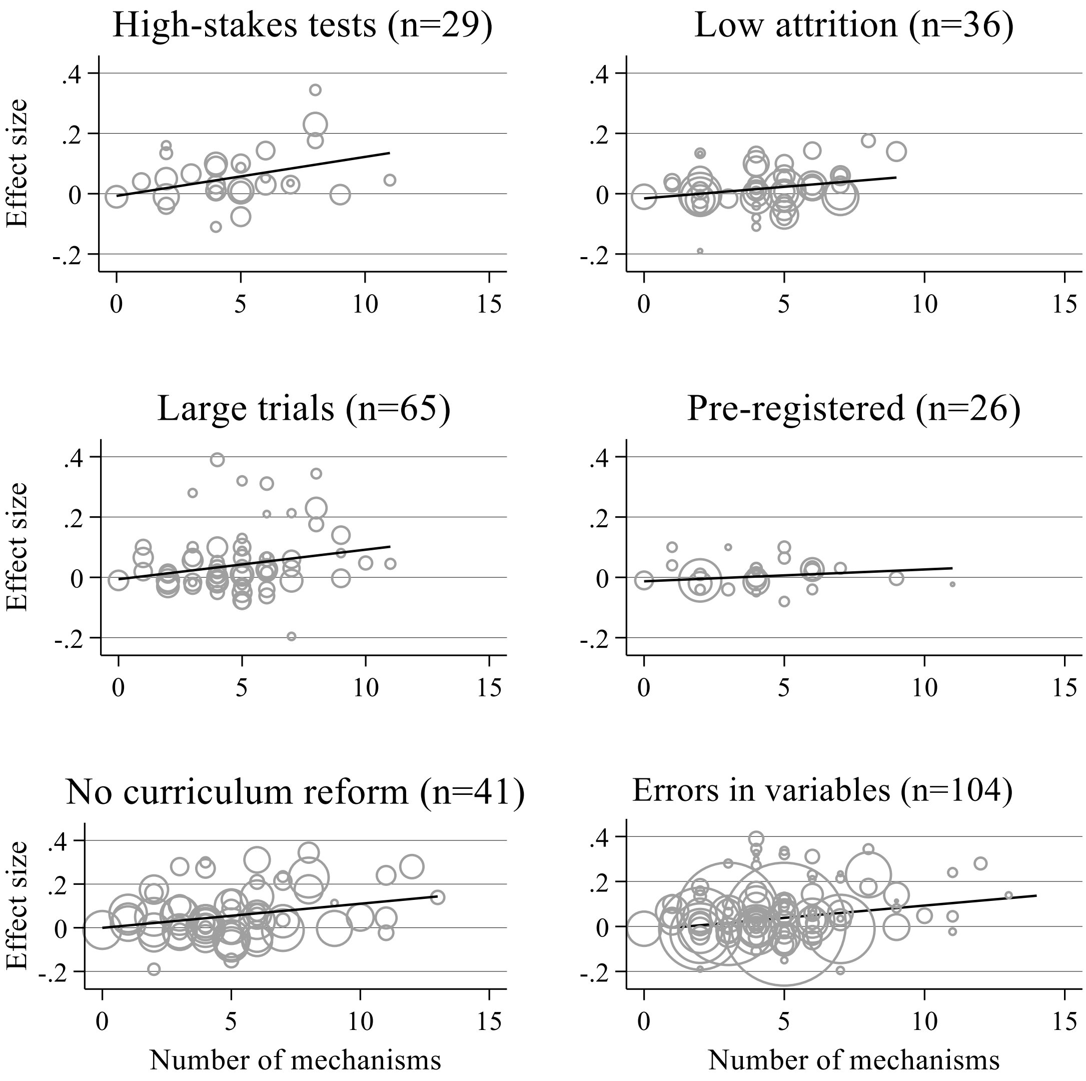


FIGURE 4. *Relationship between the number of mechanisms in a PD programme and impact on pupil test scores, by indicators of study quality*

Note. N = number of separate experimental studies. 'Large trials' involve more than 50 units randomized to treatment or control. Uses the primary outcome as specified in the study or else one randomly selected outcome per study. Effect sizes > .5 or < -.2 are used in the underlying meta-regression but not shown in the chart in order to aid visual clarity.

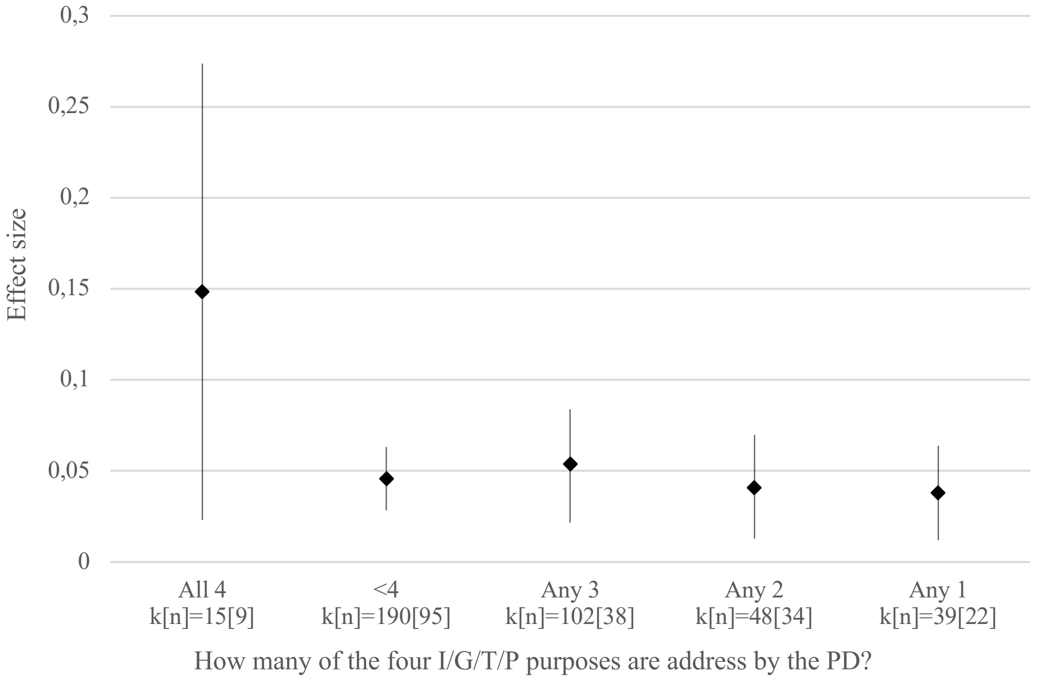


FIGURE 5. *Average impact of PD on test scores, by how many of the four 'purposes of PD' are addressed by the PD*

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

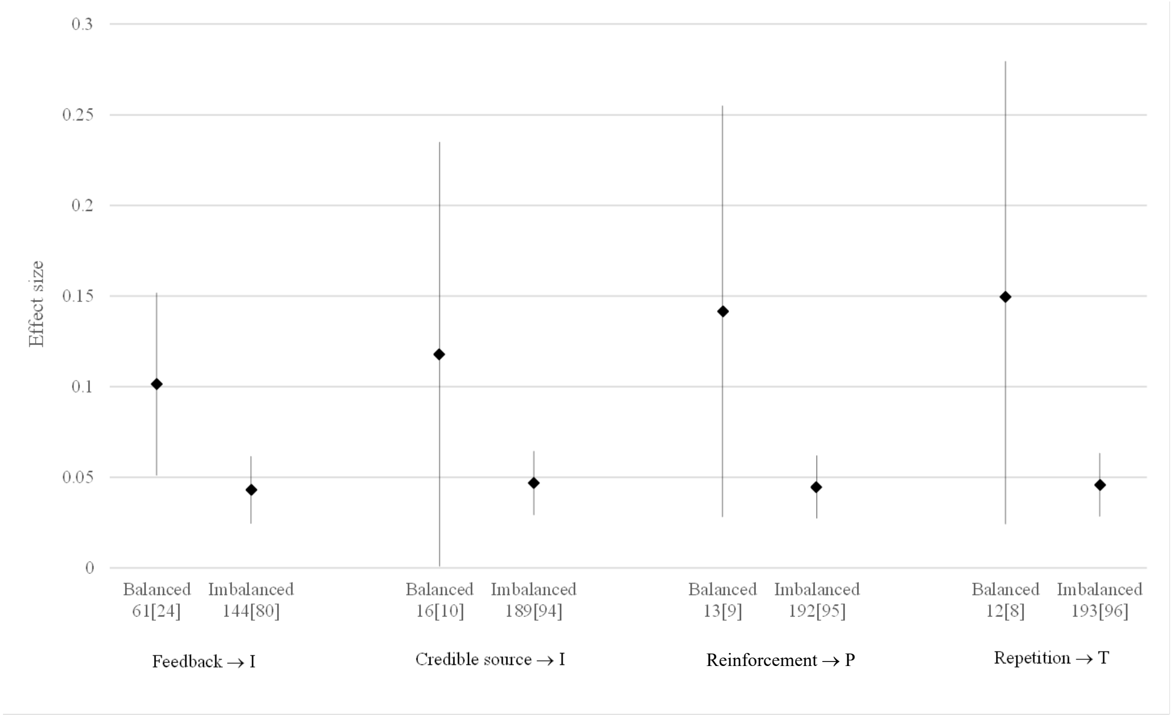


FIGURE 6. *Sensitivity analysis for meta-analytic average impact of PD on test scores, by how many of the four 'purposes of PD' are addressed by the PD design*

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

Appendix A: Description of additional mechanisms

The three *insight* mechanisms not discussed in the body of the text are *instruction, feedback,* and *rehearsal. Instruction* is the provision of directive advice on how to implement some practice. *Instruction* works by eliminating ambiguity about what is required to successfully use a procedure and has been shown to be beneficial in science education and medical training contexts (Kirschner et al., 2006; Sweller et al., 2019). *Feedback* is the provision of evaluative guidance based on prior observation of the target practice. It works by identifying and then advising on areas for improvement and has been shown to improve learning among pupils and motor-cognitive skills among dental and medical trainees (Al-Saud et al., 2017; Hatala et al., 2014; Ivers et al., 2012; Van Der Kleij et al., 2015). Finally, *rehearsal* refers to stuctured practice outside of a real classroom setting. This improves accuracy and speed of future performance. There is considerable correlational evidence for the importance of rehearsal across various domains (Macnamara et al., 2016) with causal evidence from medical education (McGaghie et al., 2011).

The two *embed practice* mechanisms not discussed in the body of the text are *prompts/cues* and *self-monitoring. Prompts/cues* involves introducing environmental stimuli with the purposes of prompting the desired practice. Prompts/cues have been shown to trigger increased goal-directed behaviour in experimental research on gym attendance (Calzolari & Nardotto, 2017), changing doctors' clinical practice (Shojania et al., 2010), and in increasing appointment attendance by patients (Guy et al., 2012). Finally, *self-monitoring* involves establishing a method for somebody to record and then review their own practice. Causal research shows that self-monitoring helps to embed health behaviour changes around weight loss, sleep hygiene and physical activity (Burke et al., 2011; Compernolle et al., 2019; Todd & Mullan, 2014).

Appendix B: Formalising the theory and hypotheses

Below is a formal statement of the hypothesized relationship between the fourteen mechanisms $x_1, x_2, \dots x_{14}$ and the four I/G/T/P purposes of PD. The subscripts on the x's correspond to the numbers in Table 2.

$$I = f(\sum_{i=1}^{2} x_i) \qquad f'(x) > 1$$

$$G = f(\sum_{i=3}^{5} x_i) \qquad f'(x) > 1$$

$$T = f(\sum_{i=6}^{10} x_i) \qquad f'(x) > 1$$

$$P = f(\sum_{i=11}^{14} x_i) \qquad f'(x) > 1$$

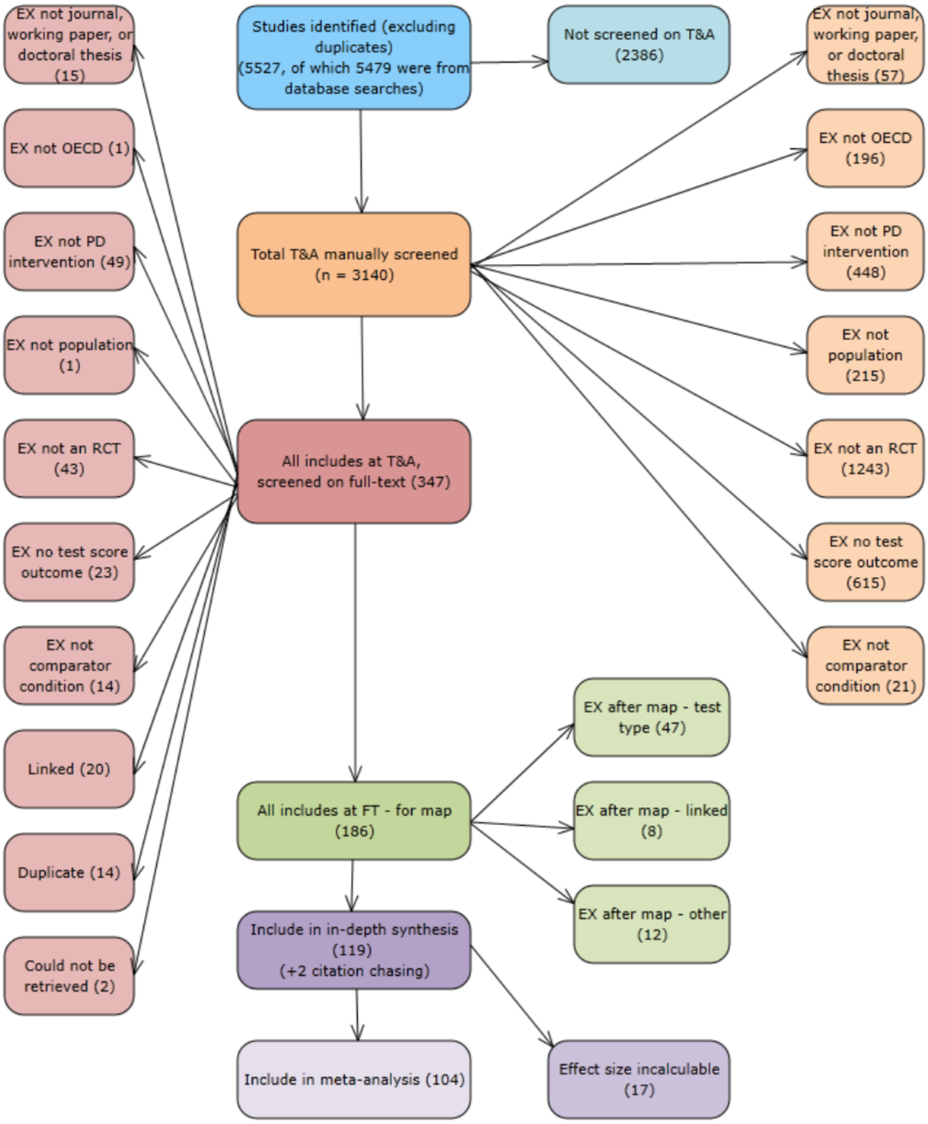
Formal statement of Hypothesis 1 (H1):

$$TestScores = f(\sum_{i=1}^{14} x_i) \qquad f'(x) > 1$$

Formal definition of a balanced design:

A 'balanced' PD design satisfies the following: $(x_1 \vee x_2) \wedge (x_3 \vee x_4 \vee x_5) \wedge (x_6 \vee x_7 \vee x_8 \vee x_9 \vee x_{10}) \wedge (x_{11} \vee x_{12} \vee x_{13} \vee x_{14})$

Appendix C: PRISMA



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

FIGURE A1. *PRISMA flow diagram*

Appendix D: References for the meta-analytic sample

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., Uekawa, K., Falk, A., Bloom, H. S., Doolittle, F., Zhu, P., & Sztejnberg, L. (2008). *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement* (NCEE 2008-4030). National Center for Education Evaluation and Regional Assistance.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., Garrett, R., Yang, R., & Borman, G. D. (2016). *Focusing on Mathematical Knowledge: The Impact of Content-Intensive Teacher Professional Development* (NCEE 2016-4010). National Center for Education Evaluation and Regional Assistance.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Landry, S. H., Zucker, T. A., Taylor, H. B., Swank, P. R., Williams, J. M., Assel, M., Crawford, A., Huang, W., Clancy-Menchetti, J., Lonigan, C. J., Phillips, B. M., Eisenberg, N., Spinrad, T.

L., de Villiers, P., Barnes, M., Starkey, P. & Klein, A. (2014). Enhancing early child care quality and learning for toddlers at risk: the responsive early childhood program. *Developmental psychology*, *50*(2), 526–541.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Savage, R., Abrami, P. C., Piquette, N., Wood, E., Deleveaux, G., Sanghera-Sidhu, S. & Burgos, G. (2013). A (Pan-Canadian) cluster randomized control effectiveness trial of the ABRACADABRA web-based literacy program. *Journal of Educational Psychology*, *105*(2), 310–328.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Wolf, B., Latham, G., Armstrong, C., Ross, S., Laurenzano, M., Daniels, C., Eisenger, J. & Reilly, J. (2018). *English Language and Literacy Acquisition-Validation i3 Evaluation (Valid 22) Final Report*. Center for Research and Reform in Education.

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

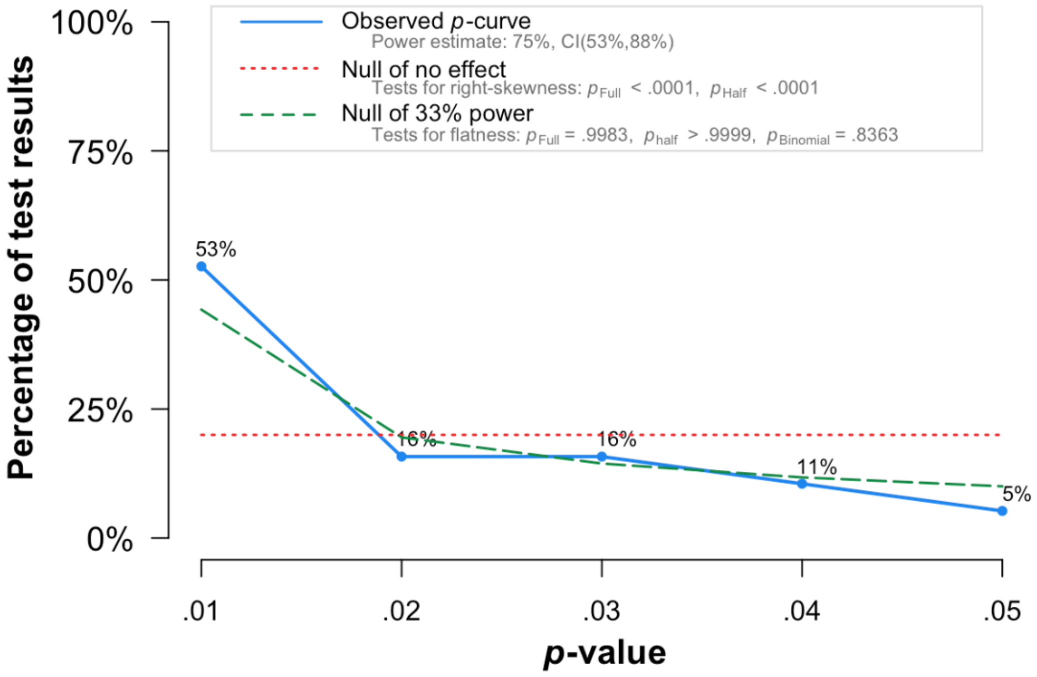
Wright, H., Carr, D., Wiese, J., Stokes, L., Runge, J., Dorsett, R., Heal, J. & Anders., J. (2020). *URLEY Evaluation Report*. Education Endowment Foundation.

Appendix E: Average impacts by indicators of study quality

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

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

Appendix F: Probing explanations for difference among pre-registered studies



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

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

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

[1] For an example database search see Appendix 2 of Sims et al. (2021). Further details about search terms are available on request from the authors.

[2] Australian Education Index (Proquest); British Education Index (BEI); EconLit (EBSCO); Education Resources Information Center (ERIC) (EBSCO); Education Abstracts (EBSCO); Educational Administration Abstracts (EBSCO); EPPI-Centre database of education research; ProQuest Dissertations & Theses; PsycINFO (OVID); Teacher Reference Center (EBSCO); Google Scholar.

[3] 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.

[4] Forward citation searching was done for all included studies that were available in Microsoft Academic.

[5] Center for Coordinated Education MRDC publications; CUREE—Centre for the use of evidence and research in education; Digital Education Resource Archive; Education Endowment Foundation (EEF); EIPEE search

portal; EPPICentre database of education research; Institute of Education Studies What Works Clearinghouse; Nuffield Foundation.

[6] See pages 80-81 of Sims et al. (2012) for further details.

[7] In Sims et al. (2021) where we show that the main results are no different when we use Hedges' *g* effect sizes among the 102 studies for which it was available

[8] See appendix 7 in Sims et al. (2021) for all three analyses.

[9] By most intensive, we mean that the other versions of the intervention include (1) some but not all of the same components, and (2) no additional components. Where it was not possible to clearly distinguish more and less intensive versions, we picked a version at random.

[10] See Appendix 5 in Sims et al. (2021) for the full coding frame.

[11] Cognitive science is PD focused on the use of findings from cognitive science relating to how memory works and how humans learn. Formative assessment is PD focused on how to elicit evidence of pupil understanding and then use this evidence to adapt the next steps in instruction. Inquiry is PD focused on pedagogy that encourages students to construct knowledge for themselves via solving problems and completing authentic tasks, working with autonomy. Data driven instruction is PD using cyclical class-wide testing to systematically collect data on pupil progress and then refocusing or differentiating instruction based on the findings. For more on this, see page 16 of Sims et al. (2021).

[12] We define a test as being high stakes if it's administration is a legal requirement by any level of government.

[13] https://d2tic4wvo1iusb.cloudfront.net/documents/guidance/EEF._Systematic_Review_of_Professional_Development._Dr_Sam_Sims._Protocol.pdf

[14] See the 'DoE' column of table 1, which finds that government funded trials (which are often pre-registered) have effect sizes around one third to one half the size of effects more generally.

@cepeo_ucl

ucl.ac.uk/ioe/cepeo