The effects of high-quality professional development on teachers and students

A rapid review and meta-analysis

Harry Fletcher-Wood and James Zuccollo

February 2020

Research Area:
Teaching and Leadership: Supply and Quality

About the Education Policy Institute

The Education Policy Institute is an independent, impartial, and evidence-based research institute that promotes high quality education outcomes, regardless of social background. We achieve this through data-led analysis, innovative research and high-profile events.

Education can have a transformative effect on the life chances of young people, enabling them to fulfil their potential, have successful careers, and grasp opportunities. As well as having a positive impact on the individual, good quality education and child well-being also promote economic productivity and a cohesive society.

Through our research, we provide insight, commentary, and a constructive critique of education policy in England – shedding light on what is working and where further progress needs to be made. Our research and analysis spans a young person's journey from the early years through to entry to the labour market.

Our core research areas include:

- Benchmarking English Education
- School Performance, Admissions, and Capacity
- Early Years Development
- Social Mobility and Vulnerable Learners
- Accountability, Assessment, and Inspection
- Curriculum and Qualifications
- Teacher Supply and Quality
- Education Funding
- Higher Education, Further Education, and Skills

Our experienced and dedicated team works closely with academics, think tanks, and other research foundations and charities to shape the policy agenda.

About Wellcome

Wellcome exists to improve health by helping great ideas to thrive. We support researchers, we take on big health challenges, we campaign for better science, and we help everyone get involved with science and health research. We are a politically and financially independent foundation.

We believe that science should be exciting for young people, giving them skills and opportunities to improve their futures. We want to make sure that all young people in the UK have access to a world-class science education.

Our work in education focuses on:

- Teacher expertise
- Primary science
- Science education research
- Informal science learning
- Practical science
- Young peoples' views on science

Research and evidence sits at the heart of our work. By providing compelling educational research, we hope to influence education policy and make sure that evidence is used to support change and improve teaching and learning.

About the Authors

Harry Fletcher-Wood has worked in schools in Japan, India and London, teaching history, organising university applications and leading teacher development. He now leads the Teacher Educator Fellows programme at Ambition Institute.

James Zuccollo is the Director for School Workforce at EPI where he leads the workforce research programme. His previous work as a professional economist has encompassed the measurement of graduates' wellbeing, development of value-added and teaching intensity measures in higher education, and evaluation of large public investments.

Acknowledgements

This report has been commissioned by the Wellcome Trust.

At EPI, Andy Schofield assisted with the meta-analysis while Sara Bonetti offered invaluable support.

At Ambition Institute, thanks to Peps McCrea, Nick Rose and Ed James for their help.

Sam Sims offered valuable advice at the beginning of the project.

We thank three anonymous reviewers for their comments and suggestions.

Contents

Executive summary

EPI and Ambition Institute were commissioned by the Wellcome Trust to examine the costs and benefits of a policy entitling all teachers to thirty-five hours of high-quality professional development every year. This rapid review of existing literature informs that analysis.

Teacher professional development is a promising approach to improving teaching quality and pupil outcomes. Evidence on effective teacher professional development is not yet conclusive, but the recent increase in rigorous evaluations of professional development interventions invites fresh analysis.

The impact of professional development on teachers and students

This review identified 53 randomised controlled trials of professional development interventions: interventions in which professional development played an important role in changing teachers' practices and improving student learning. We analysed data from these studies to identify the average impact of professional development.

We estimate an average effect size of professional development on student learning of 0.09, which compares favourably to other educational interventions (for example, comprehensive school reform models achieve effect sizes between 0.1 and 0.2). These trials do not provide sufficient evidence to reach firm conclusions about the effect of professional development on broader outcomes for students or teachers (such as student self-efficacy or teacher confidence), but we find indications that it can lead to increased student self-efficacy and confidence. Longitudinal studies of teachers' careers suggest professional development increases retention.

Offering all teachers high-quality professional development

Many teachers cannot currently access high-quality professional development. The literature we have reviewed suggests that teachers are more likely to experience high-quality professional development if designers of professional development:

- Anticipate and mitigate predictable problems, such as teacher turnover, leadership support and limited time.
- Harness predictable opportunities: teachers tend to be enthusiastic about professional development.

To improve the evidence base on professional development further, we suggest that future evaluations collect more follow-up data to show the lasting effects of the intervention and offer comprehensive descriptions of the theory of teacher change.

Introduction and purpose

The Education Policy Institute and Ambition Institute were commissioned to evaluate the costs and benefits of a policy entitling all teachers to thirty-five hours of high-quality professional development every year. This rapid review provides evidence to help model these costs and benefits. This goal informed the aims of the review, which were to:

- Identify rigorous evaluations of professional development interventions
- Estimate the impact of high-quality professional development on student learning
- Identify and, where possible, quantify other outcomes of high-quality professional development
- Outline barriers and opportunities to ensuring all teachers experience high-quality professional development

The quality of teaching that students receive is a crucial factor in improving their attainment (Jackson, Rockoff and Staiger, 2014). Teachers' qualifications, such as their degree classification and training route, play a very limited role in explaining differences in their effectiveness (Jackson, Rockoff and Staiger, 2014), while teachers' talents and motivation contribute to their effectiveness but are insufficient on their own for teachers to become experts (Berliner, 2001). It would be possible to increase teaching quality by replacing existing teachers with more effective ones, but this approach is unpromising (because it is difficult to identify effective teachers reliably; Staiger and Rockoff, 2010) impractical, given existing recruitment challenges, and would show meaningful effects very slowly (see Wiliam, 2016, Chapter 2, for discussion of this problem; Department for Education, 2019 for retention). Teachers improve gradually through experience, particularly if they teach the same course content for several years (Kini and Podolsky, 2016). This increase in effectiveness is considerably faster in a supportive professional environment: schools where students behave well, teachers collaborate, leadership is effective and professional development is strong (Kraft and Papay, 2014). Helping teachers improve appears to be a promising way to improve teaching quality.

Despite widespread support for the value of professional development in improving student learning, until recently, quantifiable evidence of its impact seemed limited. An early review of experimental studies identified 1,300 relevant studies but found only nine that met the evidence standards set by the US Institute of Education Sciences (i.e. using randomised controlled or quasi-experimental designs, which allow researchers to identify causation; Yoon et al., 2007). All nine were in elementary schools and their limitations made 'discerning any pattern in these characteristics and their effects on student achievement difficult' (Ibid., p. iv). Seven years later, a review of professional development in maths found five studies meeting the same evidence standards, of which only two showed a positive impact on student achievement. The authors noted that 'there is very limited causal evidence to guide districts and schools in selecting a math professional development approach or to support developers' claims about their approaches' (Gersten et al., 2014, p.1). While professional development is usually viewed favourably by teachers (Opfer and Pedder, 2010) and policy-makers (Department for Education, 2019), it has proved difficult to reach unambiguous conclusions about its impact on student learning or the features which matter most.

More recently however, there has been a considerable increase in the number of rigorous studies in education, in general, and of professional development in particular. Since 2006, the number of randomised controlled trials in education has increased dramatically (Connolly, Keenan and Urbanska, 2018). The Education Endowment Foundation (EEF) and its US equivalent, the National Center for Educational Evaluation and Regional Assistance (NCEE), registered 145 between them (Lortie-Forgues and Inglis, 2019). The 2007 review by Yoon et al. found only nine rigorous studies of professional development's impact on student achievement in compulsory education; however, more recently, Kennedy (2016) identified 28 and Kraft, Blazar and Hogan (2018) 22. In England, the EEF has published evaluations of 23 randomised controlled trials in professional development since 2014. Therefore, while many conclusions about professional development

made only a few years ago were, of necessity, tentative or speculative, it is now possible to review a far wider group of rigorous studies and reach more confident conclusions.

Alongside this proliferation of studies allowing causal claims, a burgeoning strand of literature has sought to identify the characteristics of effective professional development. Originating with the seminal review by Timperley et al. (2007), this literature has grown to encompass several cross-subject and single-subject reviews, culminating in a meta-review (Cordingley et al., 2015) and meta-synthesis (Dunst et al., 2015). Some reviews have focused upon the merits of specific forms of professional development (for example, teacher learning communities; Vescio, Ross and Adams, 2007), while others have sought to establish the research agenda that would permit the identification of characteristics of effective professional development (see, for example, Desimone, 2009; Sims and Fletcher-Wood, under review). However, the purpose of our review was quantifying the costs and benefits of professional development, so an extended treatment of this literature and debate is beyond the scope of our work. Instead, we have focused on those studies which have quantified the impact of that professional development on students and teachers.

Method

Our methodological choices were informed by our aims – conducting a rapid review to inform a cost-benefit analysis – and the constraints which these aims presented.

Focus on randomised controlled trials

We chose to focus on randomised controlled trials as rigorous evaluations of professional development interventions. Our decision was informed by the purposes of this study: while many forms of research permit better understanding of teacher change, randomised controlled trials offer robust causal evidence, with 'a closer approximation' to the 'attainable reality' for the effect of an intervention on student learning in ordinary circumstances (Cheung and Slavin, 2016, p.290).

We recognise limitations of randomised controlled trials, notably the problems of clustering effects, the difficulties of implementation and the possibility of attributing causal influence to redundant features of an intervention (Boylan and Demack, 2018; Wiliam, 2016; Sims and Fletcher-Wood, under review), and the range of suggestions for improvements (see, for example, Bonnell et al., 2012; Hill, Beisiegel and Jacob, 2013). However, randomised controlled trials provide the most robust causal evidence appropriate to the quantification required to conduct a cost-benefit analysis. Where conclusive evidence from randomised controlled trials was unavailable, we drew on other forms of study, for example, using longitudinal studies to examine teacher retention.

Examining professional development interventions, rather than business as usual

We chose to examine trials of interventions that go beyond current practice in school, as best reflecting the impact of a policy entitling teachers to thirty-five hours high-quality professional development. This approach could be criticised for overlooking the many and varied ways in which teachers improve (and in which schools support their improvement) outside formal interventions through, for example, peer observation and support, joint planning and resource sharing: teachers who report receiving such support (and other indicators of supportive school culture) improve more rapidly (Kraft and Papay, 2014). Some reviews do include a wider variety of studies of teachers' learning, including 'naturalistic' studies of teachers' development (for example, Goldsmith et al., 2014).

Three factors guided our belief that the effects of interventions were a better way to quantify the effects of professional development: first, there is no obvious way to quantify the effects of business-as-usual support in schools (since there is no control condition or counter-factual); second, many professional development interventions are designed to catalyse such forms of support (for example, lesson observation and feedback, Worth et al., 2017 or improving mentoring and peer support, Glazerman et al., 2010). Most importantly, however, the cost-benefit analysis for which this literature review is designed tests the potential effects of an entitlement to thirty-five hours high-quality professional development for teachers, on the understanding that this goes beyond what most teachers receive currently (the effectiveness of current professional development is discussed further below). Our approach here is designed to reflect the provision of additional, structured professional development, rather than the continuation of existing practices. Trials of professional development interventions are therefore appropriate evidence for the impact this policy would have.

Study identification – international studies

We identified relevant studies using a strategy suited to a rapid review. Original literature searches were not viable in the time available, since recent systematic reviews have returned hundreds of potentially relevant studies (Yoon et al., 2007, identified 1300; Gersten, 2014, identified 900 in mathematics professional development alone). Instead, we found recent reviews whose inclusion criteria matched ours (randomised controlled trials of professional development interventions) through what Cordingley et al. (2015) described as '"connoisseurial accumulation"': 'using experts in the field to highlight known and relevant and valuable

reviews'. This method identified several recent reviews (Blank and de las Alas, 2009; Gersten et al., 2014; Kennedy, 2016; Kraft, Blazar and Hogan, 2018; Yoon et al., 2007) whose inclusion criteria provided appropriate studies.

Since the draft of this report was completed, a further review has been published (Lynch et al., 2019) and our attention has been drawn to one other review with similar inclusion criteria (Basma and Savage, 2017). Despite their omission, the final sample of studies eligible for our meta-analysis was 53 trials: this is substantially larger than any study prior to Lynch (2019): (Kennedy, 2016, identified 28 and Kraft, Blazar and Hogan, 2018, 22) and, when combined with the EEF trials described below, this offers a relevant and robust basis for a meta-analysis.

Existing reviews supply collective support for the impact of professional development on student learning across a range of subjects and ages. Most suggest only limited further conclusions based on the small sample of studies they include (Blank and de las Alas, 2009; Gersten et al., 2014; Kraft, Blazar and Hogan, 2018), but Kennedy (2016) highlights the importance of teacher motivation and autonomy.

Study identification – English studies

Given the focus of the cost-benefit analysis on English education policy, we also sought direct evidence about the impact of professional development in English schools by identifying randomised controlled trials of professional development interventions funded by the EEF. We included all completed projects classified by the EEF as 'Staff Development and Deployment' which focused on improving whole-class teaching (see Annex A for full inclusion criteria). Most of these projects evaluate interventions intended to change curriculum (what is taught) and/or pedagogy (how it is taught). Boylan and Demack (2018) have suggested distinguishing between professional development interventions and interventions which combine professional development with specific curricular resources: the effects of the latter may reflect the strengths of the professional development, the curricular resources, or both. However, reviewers conducting a recent meta-analysis argued that it is conceptually and practically difficult to separate the curricular/pedagogical elements of such interventions (such as changes in maths teaching) from the methods used (for example, coaching; Lynch et al., 2019). We consulted with the EEF on this question and were assured that the EEF only categorises a project as 'Staff Development and Deployment' if it deems the study to advance learning about staff development (it does not categorise a project as 'Staff Development' solely because it includes some professional development; Coleman, 2019). The implication is that – despite the potential overlap between curricular/pedagogical and professional development elements – these studies are appropriate for inclusion in this meta-analysis. Therefore, while we acknowledge the potential conflation of the impact of curricular materials and professional development, we have analysed them together as the most feasible strategy at this stage. This approach is consistent with that of comparable reviews. For example, Kraft, Blazar and Hogan's (2018) meta-analysis included studies which "provided teachers with instructional content materials such as curriculum, lesson plans, or guide books" (they made up 22 of the 60 included); they found that such studies were associated with effect sizes 0.21 standard deviations higher. More recently, Lynch et al. (2019) considered this question at length in framing their review. They acknowledged that:

> "The practice of reviewing professional development and curriculum studies separately creates conceptual and practical difficulties. On the conceptual side, most curriculum programs also include a professional development component for teachers; likewise, some professional development programs offer teachers materials to support the implementation of new practices within classrooms. On the practical side, the combination of professional development and materials together may be especially effective, as compared with either one alone or one with a minimal dose of the other (Cohen & Hill, 2001) (pp.264-5)."

They therefore argued that "Studying both within one review may enhance our understanding of how these instructional improvement efforts can complement one another (p.265)." Their results suggested that:

> "Most studies of curriculum materials in our data set included at least some component of professional development, and vice versa. However, examining studies that included both elements jointly leads us to see that, on average, programs that incorporated both professional development and new curriculum materials had larger impacts as compared with programs that included only one of these components (p.284)."

Including studies of professional development – including those with explicit curricular and pedagogical elements – is therefore aligned with both the policy approach of the EEF and the approach taken by other recent reviews. This assures us that the English studies included in the meta-analysis support evaluation of the impact of professional development on student learning.

Study inclusion

The cost-benefit analysis is structured around the effect of high-quality professional development on teaching and learning. The features of high-quality professional development have been articulated by the Department for Education's Standard for Teacher Professional Development (2016) and we sought to ensure that our studies met these criteria. Therefore, we retained only those studies that explicitly focused on improving classroom teaching in compulsory education and that measured student outcomes: this focus on student learning and adoption of an outcome measure (or measures) which the trial sought to influence meets the first item of the standard ("Professional development should have a focus on improving and evaluating pupil outcomes").

All studies included a rationale for their focus which referred to past studies and underpinning theories and had won academic, government or institutional support: while we were not able to evaluate the merits of the respective evidence bases reliably within the constraints of this review, this meets the second item of the standard ("Professional development should be underpinned by robust evidence and expertise").

All studies involved collaboration between teachers (whether as groups, pairs or otherwise) and all studies involved external experts working with schools, meeting the third item of the standard: "Professional development should include collaboration and expert challenge."

All studies tested interventions which were sustained: only three lasted less than a year (four weeks, but with daily activities during this period – Siegle and McCoach, 2007; two terms – Jay et al., 2017; six months – Wiggins et al., 2017); this meets the fourth item of the standard: "Professional development programmes should be sustained over time."

The fifth item requires that "Professional development must be prioritised by school leadership". Here the data is unclear: most studies in the USA do not provide any data on this question at all, while English studies with process evaluations often discuss it but do not offer explicit, comparable measures. We are therefore unable to judge reliably the support for the trials and the effect this may have had on outcomes; nonetheless, all trials had enough leadership support to gain initial authorisation.

These rules led us to analyse a total of 53 trials, of which 42 reported results in a manner that enabled us to extract an effect size and measure of uncertainty. Of those 42 studies, 23 are English studies commissioned by the EEF and 19 were conducted in the USA.

Data extraction

We read these studies and extracted data on their characteristics, including:

- Subject (generic vs subject-specific professional development)
- Location (e.g. US, England)
- Phase (primary/secondary)
- Outcome measure
- Sample size

- Duration of professional development
- Nature of professional development
- Outcome effect size and uncertainty
- Quality of professional development

Estimating the impact of professional development on student learning

We conducted a meta-analysis using 49 effect sizes, from 42 studies, to synthesise the impact of professional development on student learning shown in the individual studies and account for the differences between groups of studies. This allowed us to:

- Increase the chances of detecting an effect. Many individual studies are insufficiently powered to detect the small effects typical of interventions in education; by combining many studies we have a better chance.
- Improve the precision of the estimated effect. By combining studies, we can more precisely estimate the effect of professional development because we are drawing on a larger, pooled sample.
- Assess the degree of conflict between individual findings and, in many cases, to resolve it, by identifying studies with extreme results and showing how typical they are.

Our approach to the meta-analysis uses a multi-level regression model that weights the studies according to the accuracy of their estimates. That means the resulting estimate of the overall effect size will be more influenced by studies with a larger sample size.

The statistical model accounts for the possibility of differences between the effect of professional development for different subjects. It also allows us to explore the possibility that different measures of pupil attainment may produce systematically different results. We have reported only the main results here, but details of the model and the various scenarios are presented in Annex B.

Explaining effect sizes

We used effect sizes to quantify the impact of a professional development intervention on student learning. Since the interventions ranged across ages and jurisdictions, improvements in test results are not directly comparable. For example, it is not obvious how an intervention leading to a higher grade for a sixteen-year old at GCSE compares to an intervention leading to a higher grade for a ten-year old in a state test in the USA. Effect sizes overcome this problem by showing the difference in outcomes between two groups, in this case, between:

- Students taught by teachers receiving a specific professional development intervention (the treatment group), and
- Students taught by teachers who are not receiving the professional development intervention but receive the professional development normally offered by their school (the control group).

The effect size is the difference between the two groups divided by the overall spread of results (the standard deviation). Coe (2002) describes an effect size as a:

> 'Standardised, scale-free measure of the relative size of the effect of an intervention. It is particularly useful for quantifying effects measured on unfamiliar or arbitrary scales and for comparing the relative sizes of effects from different studies.

The real-world impact of an intervention's effect size can be explained in several ways. An effect size of 0.1 means that:

- 54 per cent of the treatment group will score higher than the average of the control group
- There is a 53 per cent chance that a person picked at random from the treatment group will have a higher score than a person picked at random from the control group
- 34.3 people would need to receive the intervention for one person to gain a better outcome.

In choosing between interventions, one with a higher effect size is likely to have greater impact.

However, an intervention may achieve a stronger impact, and show a higher effect size, because it is a more powerful intervention, or because of the research design. For example, an outcome measure aligned to the intervention (such as an algebra test after algebra professional development) is likely to produce a higher effect size than an unaligned outcome measure (such as a GCSE Maths exam; Cheung and Slavin, 2016). We return to this issue below ('Learning more about professional development').

Moreover, the benchmarks that have been used historically for assessing effect sizes are now considered unfeasibly high. Cohen's rules of thumb (>0.20 is small, > 0.50 is medium, >0.80 is large) lack context for the type of intervention or population (Coe, 2002; Hill et al., 2008). Hattie's 'hingepoint', which suggests educational interventions are worthwhile only if they achieve an effect size over 0.4, rests on meta-analyses which included many studies in laboratories, with small groups and narrow measures, which bias estimated effect sizes far above what can be achieved in schools (see Lortie-Forgues and Inglis, 2019; Simpson, 2017).

The effects of high-quality professional development

This section describes the effects of professional development on student achievement, teacher retention and other outcomes. It offers our best estimate of the benefits that professional development offers teachers and students, based on the studies in our sample.

Effects on student learning

Our review of the randomised controlled trials of professional development interventions conducted to date reveals a positive effect on student learning. The meta-analysis of 49 outcomes across 42 studies suggests an overall effect size of 0.09 on student learning.[1] This figure had a 95 per cent confidence interval of 0.06 to 0.13. This review therefore echoes other recent meta-analyses in identifying a positive effect for teacher professional development on student outcomes (Basma and Savage, 2017; Blank and de las Alas, 2009; Gersten et al., 2014; Kennedy, 2016; Kraft, Blazar and Hogan, 2018; Lynch et al., 2019; Yoon et al., 2007), while broadening its applicability by including a substantial number of studies conducted in England.

If the likely result of a policy mandating high-quality professional development is an increase in student learning with an overall effect size of 0.09, how meaningful is this impact for teachers and pupils? An obvious comparison is with other randomised controlled trials of education interventions: those funded by the Education Endowment Foundation (in England) and the NCEE (US) achieve a mean effect size of 0.06 (Lortie-Forgues and Inglis, 2019). However, Hill et al. (2008) suggest contextualising effect sizes, including against normal growth for the population (in this case, teachers) and against comparable interventions. We therefore offer two comparisons:

- Professional development has the potential to close most of the gap between the effectiveness of novice and experienced teaches. Our estimate of the effect of professional development on student learning is similar to estimates of the effect of having a more experienced teacher on student learning (Kraft, Blazar and Hogan, 2018). Kraft and Papay (2016), for example, find that experience leads to an increase in student learning of 0.11 after a decade.
- The effect sizes for professional development represent a greater improvement than estimates for the effect of other school-based interventions, including performance-related pay and lengthening the school day (Fryer, 2016).

Table 1 Comparison of effect size across interventions

Interventions	Mean effect size
EEF and NCEE interventions (Lortie-Forgues and Inglis, 2019)	0.06
Professional development	0.09
Teacher with a decade's experience (compared to a novice; Kraft and Papay, 2016)	0.11
One-to-one tutoring (Baye et al., 2018)	0.28

Simply comparing effect sizes also risks understating the contribution professional development can make to student learning for four reasons:

- **Robustness** – The effect sizes included in our meta-analysis all derive from randomised controlled trials; the average effect size for such trials in education is 0.16 (Cheung and Slavin, 2016).
- **Scale** – 31 of the 49 studies we included have samples of more than 2,000 students, which Cheung and Slavin (2016) consider 'large'. Cheung and Slavin (2016) found that studies with over 250 students achieved average effect sizes of 0.16, which decreased to 0.11 for studies with more than 2,000

[1] See Annex B for a discussion of how this effect size was calculated and for more detailed results.

students (we return to the issue of scale below). However, we note that the concept of a sample size is not easily applied to studies that treat teachers, measure the outcomes in students and randomise at school level.

- **Cost** – While one-to-one tutoring tends to achieve more dramatic effects for pupils, the programmes available in England cost far more than professional development (and do not always achieve these promising effects). For example, the Embedded Formative Assessment programme achieved an effect size of 0.1 for all pupils at a cost of £1.20 per pupil per year (Speckesser et al., 2017); dialogic teaching achieved an effect size of 0.09 in maths and 0.15 in English at a cost of £54 per pupil per year (Jay et al., 2017). Comparing these professional development trials with specific tutoring programmes is revealing: CatchUp Literacy achieved an effect size of 0.01 at a cost of £53 per pupil per year (EEF, 2019b), Switch-on reading achieved an effect size of 0.00 at a cost of £546 per pupil per year (EEF, 2019c), while a small-scale trial of graduate coaching achieved an effect size of 0.36 at a cost of £1,400 per pupil per year (Lord et al., 2015).
- **Feasibility and acceptability** – Many alternative approaches to improving student learning demand fundamental change to school systems and structures (for example, comprehensive school reform models, which achieve effect sizes between 0.1 and 0.2; Kraft, Blazar and Hogan, 2018). These approaches can also incur substantial costs in terms of staff turnover and dissatisfaction. By contrast, most evaluations of teacher professional development report enthusiastic teacher responses.

Broader effects on students

This review found limited evidence of the effects of professional development on broader student outcomes. Only five of the studies we reviewed recorded outcomes that were not related to student learning and the results seemed inconclusive:

- The Good Behaviour Game had no impact on students' concentration, disruptive behaviour, or prosocial behaviour (Humphrey et al., 2018).
- Thinking, Doing, Talking Science improved students' attitudes towards science, their confidence and self-efficacy in the subject, in both the efficacy and the effectiveness trials (Hanley, Slavin and Elliott, 2015; Kitmitto et al., 2018).
- Distance instructional coaching for science teachers increased students' self-efficacy, practice skills and engagement (Nugent et al., 2016).
- A programme specifically designed to improve emotional self-regulation and social skills had no impact upon them (Sloan et al., 2018).

We do not seek to draw broad conclusions from these results. Our search strategy and inclusion criteria for studies, which required academic outcome measures, may have led us to exclude relevant trials. Since only around one third of educational randomised controlled trials since 1980 have specified academic outcome measures (Connolly, Keenan and Urbanska, 2018), a review seeking such evidence may come to broader conclusions. Equally, we note that there is some doubt about the creation of valid measures for some of these broader outcomes (see, for example, Sloan et al., 2018, which noted that there is no reliable measure for the emotional self-regulation of young pupils).

Effects on teachers

Professional development could increase teachers' confidence and self-efficacy, and therefore the likelihood they remain in their school and/or the profession (Coldwell, 2017). The randomised controlled trials in our sample, however, offered little data with which to assess this proposition (in this, we echo other recent reviews, which have not commented on this finding). Of 53 trials, only seven quantified effects on other outcomes for teachers, and most did not find statistically significant results. These studies found:

- **Teacher satisfaction:** no impact of participation (Gersten et al., 2010; Glazerman et al., 2010).

- **Teacher self-efficacy:** Humphrey et al. (2018) found no effects on self-efficacy or stress for a classroom management intervention; Glazerman (2010) found no impact on new teachers' feelings of preparedness. Conversely, Nugent et al. (2016) recorded an increase in teachers' self-efficacy for a science intervention with coaching, while Kitmitto et al. (2018) reported increased teacher confidence in delivering science lessons for Year 5 teachers.
- **Teacher trust:** no impact (Gersten et al., 2010).
- **Retention:** no impact (Humphrey et al., 2018; Glazerman et al., 2010).

All but two EEF trials offered qualitative process evaluations of teachers' experiences of the professional development intervention. These offer a richer but less representative data source, primarily because in many of these evaluations, responding to some or all of the evaluation activities was voluntary for participants (e.g. Boylan et al., 2018; CfEE and IEE, 2015; Hanley et al., 2016). Teachers' responses varied enormously based on their perception of the intervention. For example, The Good Behaviour Game received very mixed reports, with passionate advocates and strong sceptics, and a high dropout rate (Humphrey et al., 2018), while Rosendale School's metacognition toolkit gained positive responses (Motteram et al., 2016) as did Grammar for Writing (Tracey et al., 2019). These process evaluations offer indications of the effects of participation on teachers, but do not allow robust conclusions about their lasting effects.

Since this review informs a cost-benefit analysis of the effects of high-quality professional development, we sought broader evidence of the impact of professional development on teacher retention. The link between professional development and retention is plausible, but the randomised controlled trials we examined do not prove whether this is the case, because most neither measured teacher retention nor collected data once the study was over (most often a single year). We therefore examined the impact on retention through studies that addressed this explicitly, exploiting longitudinal datasets which are better able to show teachers' decisions over time. Two studies offered compelling data:

- For early career teachers, a study compared teachers to demographically similar teachers who had experienced different support (using propensity score matching to find teachers with similar characteristics). This found that induction support significantly increased retention and that these effects persisted over the first five years of their career. The analysis was unable to identify which specific forms of support were most effective: it concluded that the critical influence is a 'package of supports', such as induction training and mentoring (Ronfeldt and McQueen, 2017).
- A study of the effect of taking part in National STEM Learning Network professional development showed a dramatic increase in retention in the profession with some indications of increased retention within the school. This study found a stronger impact among early career teachers; this analysis focused on science teachers, who are more likely to leave teaching than those teaching other subjects (Allen and Sims, 2017).

When asked, teachers often describe factors other than professional development as greater influences on their decisions to leave or stay (such as support, leadership, and external pressure; Coldwell, 2017; Menzies et al., 2015). Nonetheless, these studies show that professional development can be the difference between remaining in teaching and leaving the profession for some teachers.

Retention is a major concern due to the increasing demand for teachers caused by a demographic bulge in secondary schools, increasing teacher turnover and low unemployment, which depresses teacher recruitment. The retention of new teachers is a particularly acute concern: over 20 per cent of new teachers leave within two years and 40 per cent leave within five years (Department for Education, 2019). Rates are higher for priority subjects like physics and maths, and shortages are particularly acute in schools in disadvantaged areas. Around 10 per cent of teachers leave teaching each year, while another eight per cent move schools each year. Moreover, teachers are more likely to move away from the schools where effective teaching is needed most (Sibieta, 2018; Watlington et al., 2010).

Increasing teacher retention should have substantial benefits. Calculating precise costs of teacher turnover is challenging (Watlington et al., 2010): they are spread between many budgets (for example, recruitment, supply, training) and organisations (schools, central governments, trainees). Cost may also be hard to quantify (for example, leadership time supporting a trainee) and calculations may be confounded if those leaving teaching later return. The figures offered here should therefore be considered illustrative rather than conclusive. If a teacher permanently leaves the profession, the cost of training a new teacher is at least £19,000 (for the cheapest training route, on 2013/14 figures). This does not include costs to schools or depressed achievement in classrooms with less experienced teachers (NAO, 2016). If a teacher shifts schools, costs are incurred for:

- Separation (for example, exit interviews)
- Recruitment (for example, advertising)
- Training for the incoming teacher (for example, the induction programme)
- Lost productivity (for example, learning new procedures, curricula and forming new relationships; Watlington et al., 2010).

Training costs and lost productivity are harder to quantify but more substantial in their impacts. One study found that they represented over 80 per cent of the cost of turnovers (Synar and Maiden, 2012). Another review discusses a range of costs, with the lowest estimate at $4,631 (in 2005 US dollars) and the highest, for an urban area, at $25,000 (in 2007 dollars; Watlington et al., 2010). Another suggests an average cost over several years of $14,500 per teacher (in 2012 dollars; Synar and Maiden, 2012). Even this may be an underestimate, since this study assumes the incoming teacher to be as productive as the outgoing teacher within five months of arrival: this seems unduly optimistic, particularly since the incoming teacher is likely to be less experienced than the outgoing one. Accepting Synar and Maiden's (2012) figure, adjusting for inflation and converting currencies offers a conservative estimate of a cost of £12,500 for each teacher moving schools. The school district they studied had 40,000 students and spent up to $5 million a year on turnover: this district is a comparable size to United Learning Trust, an English Multi-Academy Trust which runs 51 schools (Staufenberg, 2019).

Ensuring all teachers can access high-quality professional development

Showing the benefits of high-quality professional development achieved through specific interventions does not prove that these benefits can be achieved at a national scale. This section discusses potential barriers and possible solutions to offering all teachers high-quality professional development.

Barriers to spreading effective practices

The preceding sections have highlighted the current benefits of professional development to students, teachers, and society. It remains to be seen, however, whether the same programmes can achieve similar benefits when they are applied to a much larger number of schools. Reviewing literature about improving education in the developing world, Piper et al. (2018) note that '[t]he scale-up literature is pessimistic about both the initial take-up of educational actors and the long-term impact on learning outcomes, to say the least' (p. 294). Small-scale changes, which prove effective for a few determined teachers, often have limited effects at a larger scale. Elmore (1996) argues that the nearer a change is to the interaction between teacher and student, the less likely it is to achieve widespread change at scale. This problem affects all programmes: the assumption is often that they will grow to scale with the same cost-benefit ratio, but if there is a limited supply of any one input this will not be possible. For example, a programme that initially hires the best facilitators available will have to pay more to keep doing so – or accept less effective facilitators – as it grows (Davis et al., 2017).

These challenges are found in professional development interventions too. A particularly important distinction lies when developers seek to grow interventions and move from:

- An efficacy trial, which takes place 'under conditions that are conducive to obtaining an effect', such as close involvement of the developer and a small number of schools or teachers (Wayne et al., 2008, p.470), to,
- An effectiveness trial, in which an intervention is 'tested in the full range of settings in which it is designed to work' (Wayne et al., 2008, p. 470).

This distinction can be illustrated with the example of Thinking, Doing, Talking Science. An efficacy trial in 21 schools achieved an effect size of 0.22 (Hanley, Slavin and Elliott, 2015). When the same programme was shared across 102 schools, however, the effect size was 0.01, even though the only substantial change between the two stages was the introduction of facilitators to train this larger group (Kitmitto et al., 2018). This was despite the facilitators receiving training from the developers, the facilitators being assessed positively, and the teachers involved believing they had benefitted from the training. This trend was also found in Kraft, Blazar and Hogan's (2018) meta-analysis contrasting the effect of 'small' coaching programmes (fewer than 100 teachers) to 'large' coaching programmes (more than 100 teachers). They found that small programmes had a greater effect on both teaching (1.5 times larger effects) and student achievement (double the effect).

Nonetheless, the problem of scaling up impact does not seem insurmountable. Some effective trials were at a large scale. The two largest included around 25,000 students (Campbell and Malkus, 2011; Speckesser et al., 2018) and, in a developing country context, it has proved possible to scale up programmes nationally while replicating the effects of pilot programmes, through careful attention and testing of the role of each ingredient of the programme (Piper et al., 2018). Moreover, the impact on teacher retention may be easier to scale up. Neither of the longitudinal studies of teacher retention on which we base our conclusions (Allen and Sims, 2017 and Ronfeldt and McQueen, 2017) controlled for the quality of professional development or its effect on student learning in their analyses: it may be that the positive effects of professional development on teacher

retention derive more from teachers' participation than from the increased student learning and improved teaching that may result from it.

This section is not a specialist discussion of scale-up in general. It notes common themes of challenges and opportunities in the studies we reviewed – combined, where appropriate, with the wider literature on professional development – to highlight supports needed to spread this impact to other schools; this question will also be addressed in the companion cost-benefit analysis.

The effectiveness of current professional development

We first consider the effects of current 'business as usual' professional development on teachers. The data available is limited. The last substantial review of professional development in England concluded over a decade ago, expressing concerns about the quality, access, and impact of professional development (Pedder, Storey and Opfer, 2008). A snapshot of professional development on offer in 2010-11 suggested that only a fraction of it was designed to 'transform' teachers' practice (CUREE, n.d.). While there has been an increase in enthusiasm and expertise in professional learning and development through the activities of organisations such as the Teacher Development Trust and ResearchED, it is hard to demonstrate a substantial improvement in quality and impact of professional development on a national scale. A more recent review of subject-specific professional development highlights several barriers to the development of high-quality subject-specific professional development, including budgets, workload, competing priorities and a lack of high-quality provision: while the review noted that some schools are able to overcome these barriers, it called for 'an increase in effective CPD in the UK, and for building awareness of effective practice (Cordingley et al., 2018, p.6).' The disappointing results of many EEF trials may also reflect the current impact of professional development in schools, since many trials evaluate professional development interventions currently being used in schools (and most are evaluated on an 'Intention to Treat' basis: they examine the actual effects of the policy as implemented in schools, not the ideal impact in those schools which are able to prioritise professional development).

More recent data on teachers' experience of professional development can be gained from two complementary sources. TALIS 2018 (Jerrim and Sims, 2019), a survey of a nationally-representative sample of 4385 English teachers, asked: 'Thinking of all of your professional development activities during the last 12 months, did any of these have a positive impact on your teaching practice?' This sets a low bar for the overall impact of professional development (the TALIS survey asks teachers to consider nine specific forms of CPD, and of which almost all teachers have taken part in at least one, and most more than one). Among primary teachers, 91% believed professional development had a positive impact upon their teaching practices; among lower-secondary teachers, the figure was only 82%. The survey also asked teachers about the characteristics of the professional development activity with the greatest positive impact on them in the preceding year. Their responses suggest some limitations to existing professional development – for example, in whether professional development provided follow-up activity (51% agreed in lower-secondary; 61% in primary) and whether it took place over an extended period of time (41% agreed in lower-secondary; 43% in primary).

Surveys of English teachers run by Teacher Tapp offer corroborating evidence of teachers' doubts about the effectiveness of the professional development they receive. Teacher Tapp reaches a self-selecting, and slightly smaller sample of teachers than TALIS (the questions discussed below were answered by 3009 teachers). However, it avoids the clustering of teachers found in TALIS (which surveyed more teachers, but in fewer schools) and so offers an impression of national practice drawn from a wider pool of schools (Teacher, Tapp, 2019). Teachers want to improve: 100 per cent agreed that they had more to learn (Teacher Tapp 2018a). However, a substantial minority of teachers were unconvinced that professional development would help them to do so: only 38 per cent agreed that 'time and resource allocated to professional development are used in ways that enhance teachers' instructional capability (34 per cent disagreed while 28 per cent were neutral; Teacher Tapp 2018b). In addition, 31 per cent believed that if there were no In-Service Training days and no funds for professional development for the next five years there would be no impact on their teaching

(20 per cent said they would get better anyway; 11 per cent said they would not change anyway); only 29 per cent believed this would have a significantly negative impact on their practice. A question framed differently revealed that 74 per cent of respondents believed professional development was having at least a 'moderate' impact on their teaching, but this figure was lower for classroom teachers than school leaders. Teachers also question to what extent the professional development they receive is evidence-based: 70 per cent of secondary school classroom teachers believed half or less of their professional development was 'evidence-based' (Teacher Tapp, 2018b).

While the evidence is limited, it is hard to show that existing professional development in schools is having a powerful effect on student learning. It is beyond the scope of this paper to set out policy steps to move from the current situation to one in which all teachers have access to high-quality professional development, or to discuss the important contribution of Initial Teacher Education, professional development providers, universities, schools and the government in full. Nonetheless, we can sketch out some promising avenues based on the studies reviewed for this report. Most EEF studies include implementation and process evaluations, which allow us to highlight likely barriers to professional development interventions and possible ways to overcome them.

Predictable barriers to ensuring all teachers receive high-quality professional development

Reading many evaluations of professional development interventions highlights the frequent recurrence of specific challenges. Since the organisational challenges of making professional development work have been "neglected in most studies on PD (Van Driel et al., 2012, p.134) it seems appropriate to make brief mention of challenges which emerge repeatedly in EEF trials:

Leadership support

In many cases, interventions at school level cease because they lose support from school leaders. This is particularly the case where a school gains a new leader or receives a disappointing Ofsted grade. It is understandable that leaders focus their efforts on other matters at such times and that a programme they are not committed to, and may not have signed up to, may be an easy target to be cut. Writing of the challenges facing teachers in general, Kennedy (2010) notes that both 'reform clutter' and researchers' demands on teachers' time and classrooms may act as barriers to teachers' improvement – and therefore appropriate targets for leaders looking to help their teachers improve. Nonetheless, leadership support is an important barrier to ensuring professional development has sustained effects.

Teacher turnover

Many programmes report high teacher turnover as a barrier to achieving their goals. For example, the RISE programme saw 40 per cent turnover among the research leads taking part in their 30-month programme (Wiggins et al., 2019). Given teacher turnover of around 18 per cent per year (Sibieta, 2018), and higher turnover in schools serving disadvantaged areas (Allen, Mian and Sims, 2016), this is a barrier which may be anticipated in any intervention at school level. Teachers asked to take part in a programme part-way through may find themselves at a loss regarding what they have missed. Latecomers in the RISE programme found they 'lacked the context of the project and the shared camaraderie' earlier groups had achieved (Wiggins et al., 2019, p.41). This problem is magnified if a programme relies on teachers handing over an intervention to their colleagues. For example, one programme developed a computing intervention designed to be used in lessons in Year 5 and Year 6. However, since most teachers in the second year of the programme (teaching Year 6) had not attended training or taught the programme in Year 5, they knew little about it: their attendance at the training and use of the techniques were therefore considerably lower than in Year 5 (Boylan et al., 2018). These problems are most acute where a programme focuses on a single individual as the point of change. Moreover, these programmes may be good for the individual (who gains expertise and experience and may use this to gain a new job) but not necessarily for the school (Wiggins et al., 2019).

Demands on staff, school, and systems

An obvious, but critical, issue is the need for any professional development intervention to establish its importance and feasibility among teachers whose workload is already considerable (Department for Education, 2019). Many interventions struggled where teachers were unable to set aside the time intended by the developers, where schools found the demands of the programme too extreme, or where teachers concluded that the interventions were not pitched appropriately for the school or their pupils (for example, CfEE and IEE, 2015; Humphrey et al., 2018).

Planning mitigations

The weight of evidence provided by the EEF makes these barriers predictable and, therefore, at least partially surmountable. While some of the barriers described above may reflect the specific demands upon schools of taking part in externally-designed trials of professional development; programme designers may anticipate these barriers and plan to mitigate them. For example:

- **Leadership support:** the point at which leaders are most likely to withdraw support for a programme is when a new head takes over, a school becomes an academy, or a school receives a negative Ofsted rating. Designers may plan to contact a new head early in their tenure to highlight the benefits of their programme and discuss adaptations the school may need.
- **Teacher turnover:** programmes may offer supplemental support for teachers joining a programme during the year. Only one evaluation in our sample described this taking place, Parkinson et al. (2015) who noted that they 'provided supplemental coaching to new teachers and long-term substitute teachers in an effort to catch them up with the other teachers in the school' (p. 2).
- **Staff time:** designers can strive to minimise the demands they place on teachers, and to make change easy, attractive, social, and timely (Service et al., 2014).

Predictable opportunities in encouraging professional development

Despite the scepticism many teachers feel about the professional development they experience ordinarily, the evaluations of most professional development interventions find that teachers overwhelmingly welcome their programmes and support the opportunity to learn, develop, find new resources and better serve their students. This is particularly the case where interventions are seen as well-designed and appropriate to the needs of the school (for example, Tracey et al., 2019).

A pertinent example comes from Motteram et al. (2016) which (unusually) trialled the dissemination of an approach developed by one primary school (rather than a researcher-led trial) and found that teachers particularly welcomed the materials as addressing their needs. This enthusiasm is also evident from reports of schools' keenness to sign up for programmes (for example, Rose, 2017) and the disappointment reported among schools assigned to the control group (who do not receive an intervention but act as a comparison group for those receiving the intervention; for example, Wiggins et al., 2019). Teachers also welcome the chance to learn from peers and opportunities to talk to one another more. Since motivation has been posited as a crucial factor in professional development (Kennedy, 2016), designers of professional development may wish to find ways to harness teachers' enthusiasm and interest as rapidly as possible.

Sustaining improvement: how the effect of professional development accumulates over time

In estimating the effect of high-quality professional development, we must also identify how long it takes for a professional development intervention to have an impact, and whether that impact continues, increases or diminishes. This incorporates three, related dimensions:

How intense should professional development be?

Early analyses suggested that the intensity of professional development affected its impact. Yoon et al. (2007) found that six studies with more than 14 hours of professional development had an impact and three with fewer than 14 hours did not. However, more recent analyses of a larger sample of experiments call this into question. For example, Kraft, Blazar and Hogan (2018) did not find that programme duration had a significant impact on outcomes. Extreme cases demonstrate the unpredictable effects of intensity on outcomes: two two-hour meetings (Supovitz, 2013), and two days of training and 5.5 hours of coaching across the year (Sailors and Price, 2010) had a positive impact while programmes offering 114 hours (Garet et al., 2011) or 19 days of training did not (Rimm-Kaufman et al., 2011). It is not clear that more professional development automatically leads to further improvement and the design of the professional development is likely to affect the intensity needed: thirty-five hours seems a reasonable amount of time to have a meaningful impact.

How long should professional development last?

It is also difficult to show how long a programme should continue (in terms of total length of the programme, rather than intensity) in order to have an impact. There are two-year programmes which estimate negative effect sizes (eg Thurston et al., 2018; Hanley et al., 2016) and shorter programmes that have greater impact (eg Jay et al., 2017, two terms; numerous one-year programmes including Parkinson et al., 2015; Penuel, Gallagher and Moorthy, 2011; Papay et al., 2016). While sustained professional development is supported by most reviews (Cordingley et al., 2015), three recent meta-analyses have found no link between longer professional development and impact (Basma and Savage, 2017; Kraft, Blazar and Hogan, 2018; Lynch et al., 2019).

Most studies stop collecting data when the professional development ends, with the majority lasting only one year. This makes it difficult to ascertain the lasting impact of an intervention. Twelve studies in our sample offered some guidance on this question:

- For two studies, the effects did not change over multiple years (Garet et al., 2011; Wiggins et al., 2019). In both cases, the effects were indistinguishable from zero.
- In several cases, effects grew from one year to another; in some cases, there was no impact in the first year and a strong impact subsequently (Campbell and Malkus, 2011; Glazerman et al., 2010; Miller et al., 2017; Olson et al., 2017; Papay et al., 2016; Parkinson et al., 2015).
- In two cases, there was no significant impact during the programme year but an impact in year two (Allen et al., 2011; Greenleaf et al., 2011).
- Additionally, the sub-sample of a larger programme reported on by Matsumara et al. (2010) benefited from joining an established programme in its second year.

On balance therefore, the studies suggest that the impact of professional development grows the longer that the programme continues, and that this is an appropriate assumption for the cost-benefit analysis.

Learning more about professional development

The outcomes of these interventions in our sample varied widely. Partly, this reflects the design of the evaluation, not the design of the intervention. For example, tests of student learning developed by researchers to align with professional development interventions produce higher effect sizes than standardised tests (such as GCSEs), which are not aligned to specific interventions (Cheung and Slavin, 2016); we find similar effects in our meta-analysis. Even restricting the discussion to interventions measuring outcomes via standardised tests, however, we find estimates of effect size ranging from -0.09 (Garet et al., 2008) to 0.1 (Speckesser, et al., 2018) and as high as 0.29 (Matsumara, Garnier and Spybrook, 2013).

However, it is possible that these outcomes may also vary for several reasons associated with design and delivery of the programme (see Wayne et al., 2008). These include the potential of the approach to teaching being promoted: some programmes promoted approaches such as reduced teacher input and student problem solving (for example: Hanley et al., 2016; Humphrey et al., 2018; IEE, 2016), which often proves unpromising; others are designed around robust evidence, such as formative assessment (Speckesser et al., 2018). Another relevant aspect is the alignment of the intervention with principles known to promote behaviour change, such as making change easy and attractive (Service et al., 2014). For example, Jacobs et al. (2007) prioritised fundamental algebraic ideas teachers could use across their maths lessons and sought to encourage teachers to see themselves as successful, while Parkinson et al. (2015) carefully and gradually combined training, resources and individualised support. Finally, the fidelity of implementation will play a significant role in its outcome, with diminishing teacher attendance (Wiggins et al., 2017) or teachers' feeling that a programme was an unrealistic imposition (Santagata et al., 2011) potentially limiting its effects. An approach which differentiates between programmes based on these factors might prove valuable in deepening our understanding of how professional development increases student learning.

The opportunity to compare studies conducted in England and the USA illustrates strengths of each and areas where further learning might be possible:

- The EEF's consistent reporting standards and rigorous analytical approach makes comparison between studies easier and results more credible. For example, all EEF studies specify a 'primary outcome' and pre-register trials and statistical analysis plans. Some US studies report a range of outcomes without specifying one as a primary outcome (for example, Gersten et al., 2010) and it is not always clear which analyses have been pre-planned. Adopting similarly exacting standards to those used by the EEF may encourage confidence and improve the replicability of studies.
- The EEF data offers implementation data about what happened. Some American studies describe the treatment as intended, but not the results (for example, if teachers were due to receive ten days of coaching, how many did they attend?). This is crucial to understanding the impact of an intervention.
- Some studies allow researchers to trace the effect they have on teachers and students. Most are in the USA: for example, Greenleaf et al. (2011) show the increase in teacher knowledge because of training, the changes in their practice (both through self-report and external evaluation) and the improved results this causes for students. In some cases, researchers can use this information to conduct moderator analyses, which allow researchers to find the impact of a programme feature on the overall results. For example, Sailors and Price (2015) show that the amount of contact teachers have with coaches increases the amount they change. In England, Allen et al. (2011) show that it is changes in teacher-student interactions which lead to improved student outcomes. Such information allows us to trace the impact of specific aspects of a professional development intervention on teachers, and the effects each aspect has on students; doing so would allow us to refine theories of teacher change more easily.

In the future, it will be easier for researchers to learn more about the effect of professional development on teaching and student learning if research includes:

- Routine collection of light-touch follow-up data for one or more years after the programme has completed: in England, this could be sourced from the National Pupil Database, for example, to minimise the impact on schools.
- More detailed and clearer reporting of the theory of teacher change being adopted. Past reviews have noted the limited articulation of the theory of teacher change (Mandaag et al., 2017) and this issue is prevalent outside education (for example, Powell, Proctor and Glass, 2014). Future evaluations should articulate both the theory of change and the link between specific aspects of the theory of change and specific changes they hope teachers will make.
- The aggregation of comparable samples. The sample sizes needed to demonstrate an effect are often unrealistically large: in most cases, the effect size which can be detected is substantially above the effect size reported. In an extreme case, a study was powered (had enough participants) to detect an effect of 0.355 (Wiggins et al., 2019). For promising programmes, comparable randomised controlled trials would offer us greater confidence in our estimates of their effects (Hill, Beisiegel and Jacob, 2013).

Finally, we note one other barrier to estimating the effect of professional development programmes. Some evaluation reports found rapid changes in 'business as usual' control group schools, which were adopting aspects of the intervention being trialled:

- Sloan et al. (2018) found that control group schools had increased the time spent on social and emotional learning too, and concluded that their control group was 'no longer "business as usual" or "no treatment"; instead, it seems likely that the counterfactual represents the delivery of other [Social and Emotional Learning] programmes, in which case we cannot confidently conclude that Zippy's Friends is not effective—only that it is not more effective than other programmes' (p.41).
- Stokes et al. (2017) reviewed an intervention seeking to improve students' mathematical reasoning but found that most control group schools (56 of 57) were also using some form of external resources or support to improve students' mathematical reasoning.
- Speckesser et al. (2017) found more powerful effects for schools trialling Embedded Formative Assessment which had not previously introduced a similar programme (TEEP – Teacher Effectiveness Programme).

Improvements in control schools may diminish the measured effects of improvements in specific teaching practices on student learning. Researchers may have to accept that an intervention is not being tested against the absence of similar measures.

Annex A: Literature review method

We excluded studies from our analysis if they:

- Did not focus on compulsory education (Reception – Year 13/K-12)
- Targeted small intervention groups outside normal class teaching (eg literacy interventions for struggling readers who missed their normal English lessons; for example, Vernon-Feagans et al., 2013)
- Did not measure student achievement
- Were published before 2000
- Were interim reports (where final reports were available; for example, Isenberg et al., 2009).
- Were conference abstracts, papers, or unpublished dissertations (for example, Supovitz, 2013)
- Did not use randomised controlled designs with clear explanations of the method used (for example, Freiberg, Connell and Lorentz, 2001 – no explanation of randomisation; Saxe, Gearhart and Nasir, 2001; Roth et al., 2011 – self-selection into condition; West et al., 2017 – matched comparison design)
- Could not be obtained online (Niess et al., 2005)
- Analysed a subsample of data included elsewhere (Matsumara et al., 2010 – subsample of teachers included in Matsumara, Garnier and Spybrook).

All but one EEF evaluation specified a primary outcome measure, but many American studies did not. When faced with multiple outcome measures and no explicit indication of which one represented a primary outcome in the paper, we adopted the following rules to select an effect size:

- If an evaluation tested multiple treatments, we chose the most intensive (for example, for Glazerman et al. (2010) we used the results from the two-year programme rather than the one-year programme; for Penuel, Gallagher and Moorthy (2011) we used results from the 'hybrid' programme which combined both treatments)
- If a study used both a researcher-developed and a standardised test, we chose the latter.
- Where multiple cohorts benefitted from the trial, we chose the cohort which had been exposed to intervention the longest (for example, Cohort 1 in Campbell and Malkus, 2011)
- Where data was collected in multiple years, we took the final year of data (irrespective of whether this was the final year of the programme or follow-up data collection after the programme had concluded; for example, Allen et al., 2011)
- Where a study was conducted with multiple classes and effects were not pooled, for example because measures were not comparable, we took the results for the oldest age group (for example, Nugent et al., 2016)
- Where multiple outcome measures were used, we chose the measure which related most closely to growth in student knowledge (as opposed to process skills; for example, Cotabish et al., 2013)
- Where multiple outcome measures were supplied and all seemed to reflect similar dimensions, we chose the first listed (Gersten et al., 2010).

All effect sizes we collect are standardised differences in means between a treatment and control group. There are several such measures available but the Cohen's-d, Hedges-g and -g* are similar for samples as large as those in the studies we survey. In addition, the studies often do not make clear which they are reporting. That means it is both infeasible and unnecessary to convert between them. To illustrate the size of these differences, a Cohen's-d of 0.07, with sample sizes of 500 for both the treatment and control, differs from the corresponding Hedges-g by only 0.075 per cent. That difference is far smaller than the rounding error in the effect sizes we have collected. Consequently, we treat all effect sizes as comparable.

Annex B: Meta-analysis method

Aim

This meta-analysis quantitatively synthesises the effects of professional development on pupil attainment found in the studies in this rapid review. There are several reasons for conducting a quantitative meta-analysis:

- It increases the chance of detecting an effect. Individual studies are often insufficiently powered to detect the small effects characteristic of interventions in education. By combining many studies, we have a better chance.
- It improves the precision of the estimated effect. By combining studies, we can more precisely estimate the effect of CPD because we are drawing on a larger, pooled sample.
- It allows the degree of conflict between individual findings to be assessed and, in many cases, resolved. Some studies, through simple chance, will have extreme results and a quantitative assessment can identify those and help understand how unusual the results are.

The main purpose for this report is to generate an overall effect size and associated measure of uncertainty that can be used in the later cost-benefit analysis.

The meta-analysis in this report, in keeping with the nature of a rapid review, is less detailed than would be the case in a more systematic review. For example, we do not deal with the issue of publication bias, nor do we report details of the analysis such as the between-study heterogeneity.

Data

The data for this meta-analysis includes all measures of pupil attainment in the surveyed literature for which we were able to calculate both an effect size and a standard error. There are 49 such outcomes across 42 studies, with some studies reporting outcomes for multiple subjects (eg science and English).

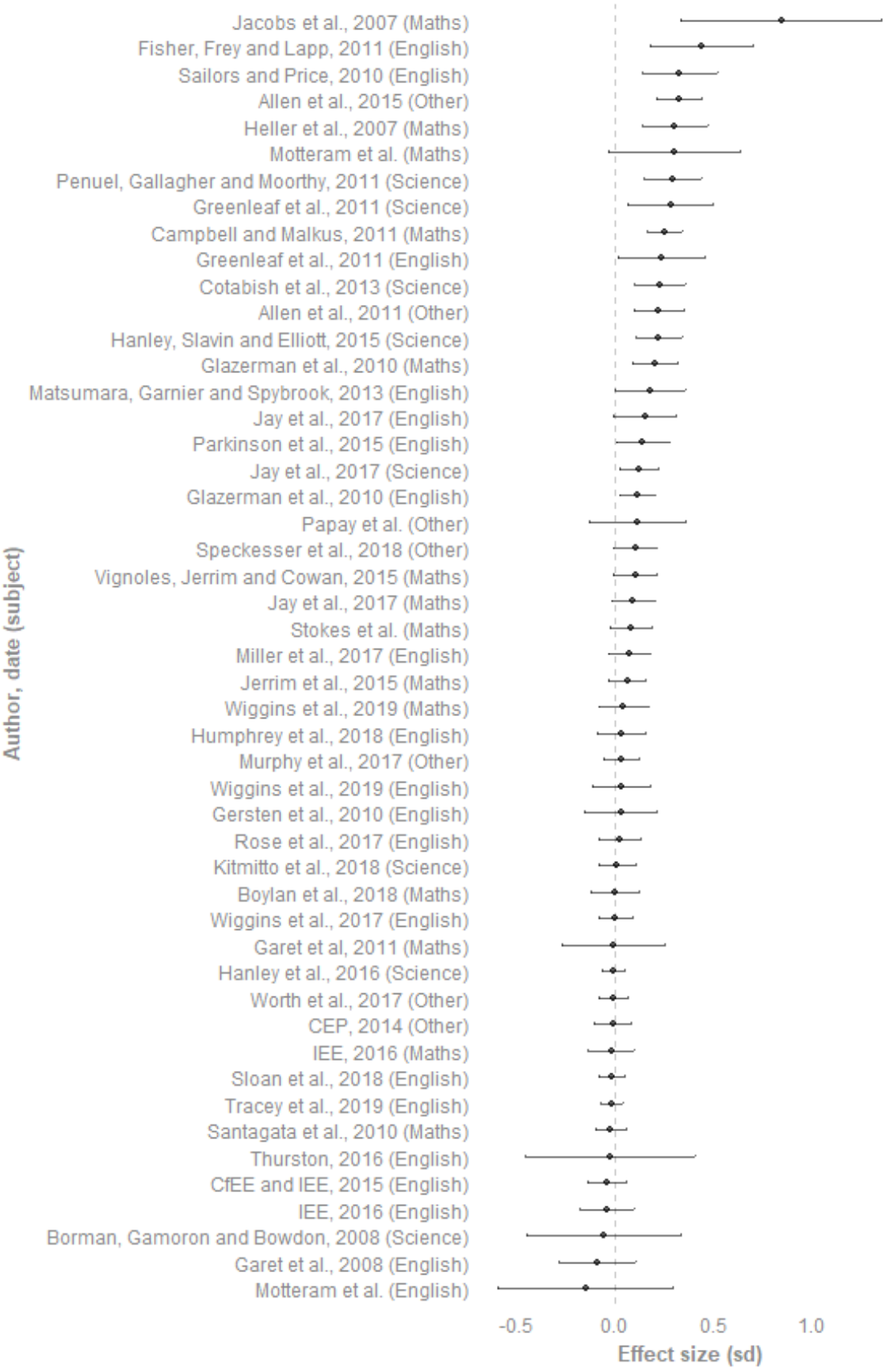
The studies include both RCTs randomised at pupil level and cluster RCTs randomised at school or teacher level. We have relied on the authors of the studies to appropriately account for the clustering in their study when reporting the uncertainty in their estimates of the mean improvement in pupil attainment.

For each attainment outcome reported in each study we recorded the average effect size and a measure of the uncertainty associated with it. Where studies reported the standard error of the effect size, we used that. Where they reported a p-value or confidence interval we calculated the standard error from it.

For seven studies that reported only an unstandardised regression coefficient we calculated the implied Hedges-g and standard error. In these cases, we were unable to account for the variance structure in the estimates, which means the standard errors may be incorrect, though the mean effect size should still be correct. Unfortunately, those studies have a mean effect greater than the average for all studies, so removing them from the estimates reduces the estimated mean effect size. We have reported results below both with and without those studies but have chosen to keep the studies in our central estimate because discarding 14% of the surveyed outcomes seems a disproportionate response to the potential bias in the standard errors.

Finally, many of the studies were insufficiently powered to detect the estimated overall effect. This is a common issue in education research and has been discussed with reference to NCEE and EEF trials in (Lortie-Forgues and Inglis 2019). They point to three possible reasons: under-powered trials, unreliable evidence on which to base trials, and poorly designed or implemented trials.

Figure 1. Effect sizes and 95% confidence intervals of the included studies



Approach

We employ a three-level, random-effects modelling framework for the meta-analysis. This approach weights studies by the inverse of the sampling variance in their reported effect sizes; however, it does not account for uncertainty in the estimates caused by other factors such as the design choices in the trial and in the trial's analysis. Those have been limited to some extent by our inclusion criteria and a more thorough investigation is outside the scope of this rapid review.

Dependent effect sizes

Our analysis faces the common issue of having multiple outcomes for some studies. The outcomes cannot be treated as independent and the possible correlation between them must be accounted for in the variance structure of the model. We use a three-level hierarchical model that decomposes the variance in observed effect sizes into sampling variance, between-study variance, and between-outcome variance (Moeyaert et al. 2017). This approach has been found to yield unbiased estimates of both the effect size and variance parameters (Van den Noortgate et al. 2013) and has previously been used in education meta-analysis (Konstantopoulos 2011).

The model we estimate is

$$g_{so} = \gamma + \mathrm{v_s} + \mathrm{u_{so}} + \varepsilon_{so}$$

where g_{so} is the observed effect size for outcome o in study s, γ is the overall effect size, v_s is the deviation from the mean effect attributable to the study, u_{so} is the deviation attributable to the outcome o in that study, and ε_{so} is the error term due to sampling variation. All variance parameters are assumed to be normally distributed, eg $v_s \sim \mathcal{N}\left(0, \sigma^2_{v_s}\right)$. The model was implemented in the R package, 'metafor', using a restricted maximum-likelihood estimator (Viechtbauer 2010).

Heterogeneity

A central concern for the meta-analysis is that the diverse studies in the review may lead to a synthesis that compares apples and oranges. For example, tests created by the researcher tend to show greater effects than external tests such as GCSEs. These differences between studies are often referred to as heterogeneity and can be usefully distinguished into two types:

- **Implementation heterogeneity**: This encompasses differences in the interventions, outcomes studied and other details of the implementation. These differences usually represent real differences in the effect we wish to estimate.
- **Methodological heterogeneity**: These are the differences caused by the study design. For example, whether the study had external or researcher-developed testing. We have limited this heterogeneity by narrowing our inclusion criteria to published RCTs but the difference in testing protocols remains. Within our sample, researcher-developed tests appear to yield greater effect sizes, as expected (Cheung and Slavin 2016).

Table 2. The frequency of test types is associated with quality

Test type	Unweighted mean effect size (n)
External	0.065 (20)
Researcher-developed	0.27 (7)
Standardised	0.11 (21)

Internal	-0.06 (1)

Note: All figures rounded to 2sf

We deal with this issue by including the type of test as a moderator in one specification and estimating a model with only standardised and externally-developed tests. That leaves four specifications of our meta-analytic model:

- **Pool all studies, no moderators**. This is our best estimate of the overall average effect of CPD.
- **Test type as a moderator**. This helps understand whether the test type is markedly influencing the effect size.
- **Only studies with external/standardised tests**. Researcher-developed tests are excluded because they may have incomparably high effect sizes as an artefact of their design.
- **Excluding studies that reported only coefficients**. This drops studies with a less reliably estimated standard error.

Differences also exist between English and American studies. However, US studies use researcher-developed tests far more often than the English studies in our sample, which confounds any reliable estimation of a difference. Moreover, we are unaware of any reason they would have systematically different results

Results

The headline results of the three scenarios are summarised by their average effect size (γ) in the table below. For model 2 there are multiple averages, one for each level of the moderating factor.

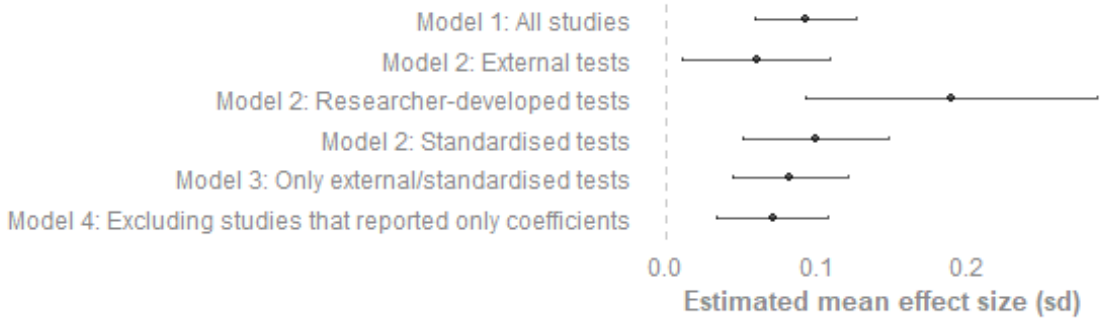
Table 3. Estimated overall effect sizes

Scenario	Effect size	Standard error	Confidence interval
Model 1: All studies	0.092	0.017	(0.058, 0.13)
Model 2: External tests	0.059	0.025	(0.01, 0.11)
Model 2: Researcher-developed tests	0.190	0.049	(0.092, 0.29)
Model 2: Standardised tests	0.099	0.025	(0.051, 0.15)
Model 2: Internal school tests	-0.060	0.220	(-0.49, 0.37)
Model 3: Only external/standardised tests	0.082	0.020	(0.043, 0.12)
Model 4: Excluding studies that reported only coefficients	0.070	0.019	(0.033, 0.11)

Note: All figures rounded to 2sf

Examining a plot of the results shows that researcher-developed tests do seem to lead to far greater effect sizes. However, there is a large degree of overlap between the results of the remaining scenarios. Consequently, we feel confident that including all studies in our estimate of the overall effect fairly represents the findings of the studies reviewed.

Figure 2. Comparison of overall effect sizes



Note: "Model 2: Internal school tests" not included because the category has only one study.

References

References marked * are included in the meta-analysis

* Allen, J., Hafen, C., Gregory, A., Mikami, A. and 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), pp.475-489.

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

Allen, R., Mian, E. and Sims, S. (2016) Social inequalities in access to teachers Social Market Foundation Commission on Inequality in Education: Briefing 2

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

Baye, A., Lake, C., Inns, A., Slavin, R., (2018). A Synthesis of Quantitative Research on Reading Programs for Secondary Students. Best Evidence Encyclopedia.

Berliner, D. (2001). Learning about and learning from expert teachers. International Journal of Educational Research 35, pp.463–482.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T. and Moore, L., 2012. Realist randomised controlled trials: a new approach to evaluating complex public health interventions. Social science & medicine, 75(12), pp.2299-2306.

* Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects. Journal of Research on Educational Effectiveness, 1, 237–264. doi:10.1080/19345740802328273

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

* Boylan, M., Demack, S., Wolstenholme, C., Reidy, J. and Reaney-Wood, S. (2018). ScratchMaths: Evaluation report and executive summary. Education Endowment Foundation.

* Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. Elementary School Journal, 111, 430–454. doi:10.1086/657654

* Centre for Economic Performance (2014). Hampshire Hundreds: Evaluation Report and Executive Summary. Education Endowment Foundation.

* Centre for Effective Education, & Institute for Effective Education. (2015). Quest Evaluation Report and Executive Summary. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Quest.pdf

Cheung, A. and Slavin, R. (2016). How Methodological Features Affect Effect Sizes in Education. Educational Researcher, 45(5), pp. 283–292.

Clarke, B., Gillies, D., Illari, P., Russo, F. and Williamson, J. (2013). Mechanisms and the Evidence Hierarchy. Topoi, 33(2), pp.339-360.

Coe, R. (2002). It's the Effect Size, Stupid: What effect size is and why it is important. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002.

Coldwell, M. (2017). Exploring the influence of professional development on teacher careers: developing a path model approach. Teaching and teacher education, 61, pp.189-198.

Coleman, R. (2019). Personal communication, 30th July. [Coleman is Head of Policy at the EEF].

Connolly, P., Keenan, C., and Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016, Educational Research, DOI: 10.1080/00131881.2018.1493353,

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.

* Cotabish, A., Robinson, A., Dailey, D., Hughes, G. (2013). The Effects of a STEM Intervention on Elementary Students' Science Knowledge and Skills. School Science and Mathematics, 113(5), pp.215-226.

CUREE (n.d.). Evaluation of CPD providers in England 2010-2011 Report for School Leaders

Davis, J. Guryan, J., Hallberg, K., Ludwig, J. (2017). The Economics of Scale Up. National Bureau of Economic Research Working Paper 23925.

Deans for Impact (2015) The Science of Learning. Austin, TX: Deans for Impact

Department for Education (2016). Standard for teachers' professional development.

Department for Education (2019). Teacher Recruitment and Retention Strategy.

Desimone, L. (2009). Improving Impact Studies of Teachers' Professional Development: Toward better Conceptualizations and Measures. Education Researcher 38(3), pp.181-199.

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

Education Endowment Foundation (2019a). Success For All [webpage, accessed 22nd May 2019] https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/success-for-all/

Education Endowment Fund (2019b). Catch-up Literacy (re-grant) [webpage, accessed 22nd May 2019] https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/catch-up-literacy-effectiveness-trial/

Education Endowment Fund (2019c). Switch-on Reading (re-grant) [webpage, accessed 22nd May 2019] https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/switch-on-effectiveness-trial/

Elmore, R. (1996). Getting to Scale with Good Educational Practice. Harvard Educational Review, 66(1), pp.1-26.

* Fisher, D., Frey, N. and Lapp, D. (2011). Coaching middle-level teachers to think aloud improves comprehension instruction and student reading achievement. The Teacher Educator, 46(3), pp.231-243.

Freiberg, H. J., Connell, M. L., & Lorentz, J. (2001). Effects of consistency management on student mathematics achievement in seven Chapter I elementary schools. Journal of Education for Students Placed at Risk, 6, 249–270. doi:10.1207/ S15327671ESPR0603_6

Fryer, R. (2016) The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments. NBER Working Paper No. 22130.

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

* Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., Doolittle, F., Warner, E., (2011) Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation. Institute of Education Sciences.

* 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, 694–739. doi:10.3102/0002831209361208

Gersten, R., Taylor, M. J., Keys, T.D., Rolfhus, E., and Newman-Gonchar, R. (2014). Summary of research on the effectiveness o math professional development approaches. (REL 2014–010). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs.

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

Goldsmith, L.T., Doerr, H.M. & Lewis, C.C. (2014). Mathematics teachers' learning: a conceptual framework and synthesis of research. Journal of Mathematics Teacher Education. 17(1), pp.5-36.

* Greenleaf, C. L., Litman, C., Hanson, T. L., Rosen, R., Boscardin, C. K., Herman, J., Jones, B. (2011). Integrating literacy and science in biology: Teaching and learn-ing impacts of Reading Apprenticeship professional development. American Educational Research Journal, 48, 647–717. doi:10.3102/0002831210384839

Gregory, A., Ruzek, E, Hafen, C., Mikami, A., Allen, J. and Pianta, R. (2017) My Teaching Partner-Secondary: A Video-Based Coaching Model, Theory Into Practice, 56:1, 38-45, DOI: 10.1080/00405841.2016.1260402

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

* Hanley, P., Böhnke, J. R., Slavin, B., Elliott, L., & Croudace, T. (2016). Let's Think Secondary Science Evaluation Report and Executive Summary. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Lets_Think_Secondary_Science.pdf

* Hanley, P., Slavin, R., & Elliott, L. (2015). Thinking, Doing, Talking Science Evaluation report and Executive Summary. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Oxford_ Science.pdf

* Heller, J., Curtis, D., Rabe-Hesketh, S. and Verboncoeur, C. (2007). The Effects of Math Pathways and Pitfalls on Students' Mathematics Achievement. National Science Foundation Final Report.

Heller, J., Daehler, K., Wong, N., Shinohara,M., Miratrix, L. (2012). Differential Effects of Three Professional Development Models on Teacher Knowledge and Student Achievement in Elementary Science. Journal of Research in Science Teaching, 49(3), pp.333–362.

Hill, C.J., Bloom, H.S., Black, A.R., and Lipsey, M.W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. Child Development Perspectives, 2(3), pp.172–177.

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

* Humphrey, N., Hennessey, A., Ashworth, E., Frearson, K., Black, L., Petersen, K., Wo, L., Panayiotou, M., Lendrum, A., Wigelsworth, M., Birchinall, L., Squires, G. and Pampaka, M. (2018). Good Behaviour Game: Evaluation report and executive summary

* Institute for Effective Education. (2016). Teacher Effectiveness Enhancement Programme Evaluation Report and Executive Summary. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/TEEP.pdf

Isenberg, E., Glazerman, S., Bleeker, M., Johnson, A., Lugo-Gil, J., Grider, M., & Dolfin, S. (2009, August). Impacts of comprehensive teacher induction: Results from the second year of a randomized controlled study. Washington, DC: Institute for Education Sciences

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher Effects and Teacher-Related Policies. Annual Review of Economics, 6(1), pp.801-825.

* Jacobs, V. R., Franke, M. L., Carpenter, T., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. Journal for Research in Mathematics Education, 38, 258–288

* Jay, T., Willis, B., Thomas, P., Taylor, R., Moore, N., Burnett, C., Merchant, G., & Stevens, A. (2017). Dialogic Teaching Evaluation Report and Executive Summary. London: Education Endowment Foundation.

* Jerrim, J., Austerberry, H., Crisan, C., Ingold, A., Morgan, C., Pratt, D., Smith, C. and Wiggins, M. (2015). Mathematics Mastery: Secondary Evaluation Report. Education Endowment Foundation.

Jerrim, J. and Sims, S. (2019). The Teaching and Learning International Survey (TALIS) 2018. Department for Education.

Kennedy, M. (2010). Attribution Error and the Quest for Teacher Quality. Educational Researcher, 39(8), pp.591-598.

Kennedy, M. (2015). Parsing the Practice of Teaching. Journal of Teacher Education, 67(1), pp.6-17.

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

Kini, T., Podolsky, A. (2016). Does Teaching Experience Increase Teacher Effectiveness? A Review of the Research. Learning Policy Institute

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

Konstantopoulos, S. (2011). Fixed Effects and Variance Components Estimation in Three-Level Meta-Analysis. Research Synthesis Methods, 2(1), pp.61–76.

Kraft, M., 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), pp.547-588.

Kraft, M., Papay, J. (2014) Can Professional Environments in Schools Promote Teacher Development? Explaining Heterogeneity in Returns to Teaching Experience Educational Evaluation and Policy Analysis 36(4) 476-50

Lord, P., Bradshaw, S., Stevens, E., and Styles, B. (2015). Perry Beeches Coaching Programme Evaluation report and Executive summary. Education Endowment Foundation.

Lortie-Forgues, H. and Inglis, M. (in 2019) Rigorous Large-Scale Educational RCTs are Often Uninformative: Should We Be Concerned? Educational Researcher. ISSN 0013-189X.

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. https://doi.org/10.3102/0162373719849044.

Mandaag, D., Helms-Lorenz, M., Lugthart, E., Verkade, A., and Van Veen, K. (2017). Features of effective professional development interventions in different stages of teacher's careers: A review of empirical evidence and underlying theory. University of Groningen.

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

Matsumura, L. C., Garnier, H. E., Correnti, R., Junker, B., & Bickel, D. D. (2010). Investigating the effectiveness of a comprehensive literacy coaching program in schools with high teacher mobility. Elementary School Journal, 111, 35–62. doi:10.1086/653469

Mayer, R., (2004) Should There Be a Three-Strikes Rule Against Pure Discovery Learning? The Case for Guided Methods of Instruction. American Psychologist 59(1), pp. 14–19.

* Miller, S., Biggart, A., Sloan, S. and O'Hare, L. (2017). Success for All: Evaluation report and executive summary. Education Endowment Foundation.

* Motteram, G., Choudry, S., Kalambouka, A., Hutcheson, G., and Barton, A. (2016). ReflectED: Evaluation report and executive summary. Education Endowment Foundation.

Moeyaert, M., Ugille, M., Beretvas, S.N., Ferron, J., Bunuan, R., and Van den Noortgate, W. (2017). Methods for Dealing with Multiple Outcomes in Meta-Analysis: A Comparison between Averaging Effect Sizes, Robust Variance Estimation and Multilevel Meta-Analysis'. International Journal of Social Research Methodology, 20(6), pp.559–72.

* Murphy, R., Weinhardt, F., Wyness, G. and Rolfe, H. (2017). Lesson Study: Evaluation report and executive summary. Education Endowment Foundation.

Niess, M. (2005). Oregon ESEA Title IIB MSP. Corvallis: Central Oregon Consortium.

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

* Olson, C., Matuchniak, T., Chung, H., Stumpf, R., & Farkas, G. (2017). Reducing achievement gaps in academic writing for Latinos and English learners in Grades 7–12. Journal of Educational Psychology, 109(1), pp.1-21.

Opfer, V., and Pedder, D. (2010) Benefits, status and effectiveness of Continuous Professional Development for teachers in England, The Curriculum Journal, 21(4), pp.413-431.

Pan, S. and Rickard, T. (2018). Transfer of Test-Enhanced Learning: Meta-Analytic Review and Synthesis. Psychological Bulletin. DOI 10.1037/bul0000151

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

* Parkinson, J., Salinger, T., Meakin, J., & Smith, D. (2015). Results from a three-year i3 impact evaluation of the Children's Literacy Initiative (CLI): Implementation and impact findings of an intensive professional development and coaching program. American Institutes for Research. Retrieved from www.cli.org/.

Pedder, D., Storey, A. and Opfer, V. (2008). Synthesis report: Schools and continuing professional development (CPD) in England – State of the Nation research project. Training and Development Agency for Schools

* Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth science: A comparison of three professional development programs. American Educational Research Journal, 48, 996–1025. doi:10.3102/0002831211410864

Perkins, D. and Salomon, G. (1989). Are Cognitive Skills Context-Bound? Educational Researcher, 18(1), pp.16-25.

Piper, B., Destefano, J., Kinyanjui, E., Ong'ele, S. (2018). Scaling up successfully: Lessons from Kenya's Tusome national literacy program. Journal of Educational Change (2018) 19, pp.293–321.

Powell, B., Proctor, E. and Glass, J. (2014) A Systematic Review of Strategies for Implementing Empirically Supported Mental Health Interventions. Research on Social Work Practice. 24(2): p192–212.

* Rimm-Kaufman, S. E., Baroody, A. E., Curby, T. W., Ko, M., Thomas, J. B., Merritt, E. G…. DeCoster, J. (2014). Efficacy of the Responsive Classroom Approach: Results from a 3- year, longitudinal randomized controlled trial. American Educational Research Journal, 51(3), 567-603. doi:10.3102/0002831214523821

* Rose, J. (2017). Research Learning Communities: Evaluation report and executive summary. Education Endowment Foundation.

Rosenshine, B. (2010). Principles of instruction, International Bureau of Education Educational Practices Series, 21

Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. Z. (2011). Video-based lesson analysis: Effective science PD for teacher and student learning. Journal for Research in Science Teaching, 48, 117–148. doi:10.1002/tea.20408

* Sailors, M., & Price, L. (2015). Support for the Improvement of Practices through Intensive Coaching (SIPIC): A model of coaching for improving reading instruction and reading achievement. Teaching and Teacher Education, 45, 115-127. doi: 10.1016/j.tate.2014.09.008.

* Sailors, M., & Price, L. R. (2010). Professional development that supports the teaching of cognitive reading strategy instruction. Elementary School Journal, 110, 301–322. doi:10.1086/648980

* Santagata, R., Kersting, N., Givven, K. B., & Stigler, J. W. (2011). Problem implementation as a lever for change: An experimental study of the effects of a professional development program on students' mathematics learning. Journal of Research on Educational Effectiveness, 4, 1–24. doi:10.1080/19345747.2010.498562

Saxe, G. B., Gearhart, M., & Nasir, N. I. S. (2001). Enhancing students' understanding of mathematics: A study of three contrasting approaches to professional support. Journal of Mathematics Teacher Education, 4, 55–79. doi:10.1023/A:1009935100676

Service, O., Hallsworth, M., Halpern, D., Algate, F., Gallagher, R., Nguyen, S., Ruda, S., Sanders, M. (2014). EAST: Four simple ways to apply behavioural insights. Behavioural Insights Team.

* Siegle, D. and McCoach, B. (2007). Increasing Student Mathematics Self-Efficacy Through Teacher Training. Journal of Advanced Academics, 18(2), pp.278–312.

Simpson, A. (2017). The misdirection of public policy: comparing and combining standardised effect sizes. Journal of Education Policy, 32(4), pp.450-466.

Sims, S., and Fletcher-Wood, H. (under review). Characteristics of effective teacher professional development: what we know, what we don't, how we can find out. Available from https://improvingteaching.co.uk/characteristics-cpd/.

Slavin, R. (2019). The Fabulous 20%: Programs Proven Effective in Rigorous Research. Robert Slavin's blog, 18th April, https://robertslavinsblog.wordpress.com/2019/04/18/the-fabulous-20-programs-proven-effective-in-rigorous-research/ [accessed, 22nd May, 2019]

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

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

Staiger, D.O. and Rockoff, J.E., (2010). Searching for effective teachers with imperfect information. Journal of Economic perspectives, 24(3), pp.97-118.

Staufenberg, J. (2019). Investigation: The highs (and occasional lows) of academy CEO pay. *Schools Week*, 1st March.

* Stokes, L., Hudson-Sharp, N., Dorsett, R., Rolfe, H., Anders, J., George, A., Buzzeo, J., and Munro-Lott, N (2018). Mathematical Reasoning: Evaluation report and executive summary. Education Endowment Foundation

Supovitz, J. (2013) The Linking Study: An Experiment to Strengthen Teachers' Engagement With Data on Teaching and Learning. CPRE Working Papers.

Teacher Tapp (2018a). What teachers Tapped this Week #63 – 10th December 2018. Teacher Tapp [blog], https://teachertapp.co.uk/2018/12/what-teachers-tapped-this-week-63-10th-december-2018/

Teacher Tapp (2018b). What teachers Tapped this Week #56 – 22nd October 2018. Teacher Tapp [blog], https://teachertapp.co.uk/what-teachers-tapped-this-week-56-22nd-october-2018/

Teacher Tapp (2019). How clustered is our sample? Teacher Tapp [blog] https://teachertapp.co.uk/how-clustered-is-our-sample/

* Thurston, A. (2016). Talk of the Town: Evaluation report and executive summary. Education Endowment Fund

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

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

Van den Noortgate, W., López-López, J.A., Marín-Martínez, F., and Sánchez-Meca, J. (2013). Three-Level Meta-Analysis of Dependent Effect Sizes. Behavior Research Methods 45(2), pp.576–94.

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

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.

Vescio, V., Ross, D., Adams, A. (2007) A review of research on the impact of professional learning communities on teaching practice and student learning. Teaching and Teacher Education 24, pp.80–91

Viechtbauer, W., (2010). Conducting Meta-Analyses in R with the Metafor Package. Journal of Statistical Software, 36(1), pp.1–48.

* Vignoles, A., Jerrim, J. and Cowan, R. (2015). Mathematics Mastery: Primary Evaluation Report. Education Endowment Foundation

* Villar, A. and Strong, M. (2007). Is Mentoring Worth the Money? A Benefit-Cost Analysis and Five-year Rate of Return of a Comprehensive Mentoring Program for Beginning Teachers. ERS Spectrum. 25(3), pp. 1-17

Watlington, E., Shockley, R., Guglielmino, P., Felsher, R. (2010). The High Cost of Leaving: An Analysis of the Cost of Teacher Turnover. Journal of Education Finance, 36(1), pp. 22-37.

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

West, M., Ainscow, M., Wigelsworth, M., Troncoso, P. (2017). Challenge the Gap: Evaluation report and executive summary. Educational Endowment Fund.

* Wiggins, M. Jerrim, J., Tripney, J., Khatwa, M. and Gough, D. (2019). The RISE Project: Evidence-informed school improvement: Evaluation Report. Education Endowment Fund.

* Wiggins, M., Parrao, C., Austerberry, H. and Ingold, A. (2017). Foreign Language Learning in Primary School: Evaluation report and executive summary. Education Endowment Foundation.

Wiliam, D. (2016) Leadership for Teacher Learning: Creating a Culture Where All Teachers Improve So That All Students Succeed. Learning Sciences International.

* Worth, J., Sizmur, J., Walker, M., Bradshaw, S., Styles, B. (2017). Teacher Observation: Evaluation report and executive summary. Education Endowment Foundation.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). Reviewing the evidence on how teacher professional development affects student achievement (Issues & Answers Report, REL 2007–No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs