Education Endowment Foundation

REVIEW OF EEF PROJECTS

August 2021

Sean Demack, Bronwen Maxwell, Mike Coldwell, Anna Stevens, Claire Wolstenholme, Sarah Reaney-Wood, Bernadette Stiell (Sheffield Institute of Education, Sheffield Hallam University)

Hugues Lortie-Forgues (University of York)

Sheffield Hallam University | Sheffield Institute of Education

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale;
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus Trust (now part of Impetus – Private Equity Foundation) and received a founding £125m grant from the Department for Education.

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

For more information about the EEF or this report please contact:

Jonathan Kay
Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

0207 802 1653

jonathan.kay@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Contents

List of tables and figures

About the reviewer

This evaluation was conducted by a team from Sheffield Institute of Education (IoE), including Sean Demack, Bronwen Maxwell, Mike Coldwell, Anna Stevens, Claire Wolstenholme, Sarah Reaney-Wood and Bernadette Stiell, together with a senior statistician from the University of York, Hugues Lortie-Forgues. The lead evaluators were Sean Demack and Bronwen Maxwell.

The Sheffield Institute of Education (IoE) at Sheffield Hallam University is a leading provider of initial and continuing teacher education; undergraduate, post-graduate and doctoral education programmes; and has a long-established track record in educational research, evaluation and knowledge exchange. Key areas of research and evaluation expertise span curriculum and pedagogy, policy and professional learning, and diversity and social justice. The IoE has extensive experience of evaluation methodologies and undertaking large-scale evaluations for a range of funders, including the Education Endowment Foundation (EEF).

Contact details:

Names of Principal Investigators (PIs): Sean Demack and Bronwen Maxwell

Address: Sheffield Institute of Education Research and Knowledge Exchange Centre

Sheffield Hallam University

Room 10101, Arundel Building

City Campus

Howard St

Sheffield

S1 1WB

Tel: 0114 225 6066

Email: SIoECDARE@shu.ac.uk

Acknowledgements

Dr Katie Shearn: Contribution to theory & evidence variable development

Lisa Clarkson, Lyn Pickerel, Rosie Smith, Noreen Drury, Charlotte Rowley, Hannah Joyce, Kellie Cook: Coding team

Executive summary

Review scope and methodology

This report presents findings from exploratory, descriptive meta-analyses of effect sizes reported by the first 82 EEF evaluations that used a randomised controlled trial (RCT) or clustered RCT impact evaluation design published up to January 2019. The review used a theoretical framework derived from literature with five overarching themes to group explanatory variables:

1 Intervention

2 Implementation & fidelity

3 Theory & evidence

4 Context

5 Evaluation design.

Meta-analyses of effect sizes reported for intention-to-treat (ITT) analyses of primary attainment outcomes, ITT analyses of secondary attainment outcomes and free school meals (FSM) subsample analyses of primary or secondary attainment outcomes are reported. Effect sizes reported for psychological outcomes were also examined but not included in the meta-analyses because of the distinct and diverse nature of these outcomes. We also present findings from trial/evaluation-level descriptive analyses of cost effectiveness of interventions and overall pupil-level attrition. An expanded summary is provided as a separate document (Demack et al., 2021).

Findings

Outcomes

- A total of 133 primary ITT effect sizes were reported by the 82 evaluations in the review and an overall meta-analysis weighted mean effect size of +0.04 standard deviations (SD) was observed.
- A total of 78 secondary ITT attainment effect sizes were reported by 35 of the 82 evaluations in the review and an overall meta-analysis weighted mean effect size of +0.01 SD was observed.
- A total of 149 FSM subsample attainment effect sizes were reported by 73 of the 82 evaluations in the review and an overall meta-analysis weighted mean effect size of +0.03 SD was observed.

Psychological outcomes were also examined but not included in the meta-analyses.[1]

- A total of 88 psychological ITT or FSM effect sizes were reported by 21 of the 82 evaluations in the review and an overall meta-analysis weighted mean effect size of +0.05 SD was observed.

Unlike effect sizes, cost effectiveness and pupil-level attrition are measured at the trial/evaluation level. A descriptive statistical approach was taken for the analyses of these trial-level variables as they were unsuited to meta-analyses.

- 40 of the 82 evaluations in the review (49%) reported evidence of a positive impact[2] and were included in the cost effectiveness outcome.

[1] The theoretical framework for the meta-analyses was developed by focusing on effect sizes reported for ITT analyses of primary outcomes across the 82 evaluations. This was adapted for the meta-analyses of effect sizes reported by secondary attainment and FSM attainment outcomes. However, because of the distinct nature and diversity of the psychological outcomes, they were unsuited to this framework. See *Presenting the outcome variables* below.
[2] Evidence of a positive impact ~ when at least half reported effect size(s) for ITT analyses of primary outcome(s) were above +0.05 SD.

- The distribution of cost effectiveness for the 40 evaluations included was highly skewed with a mean of £150 and median of £54 per pupil for an effect size of +0.10 SD.

- A pupil-level attrition rate was obtained for 79 of the 82 evaluations in the review. A mean attrition rate of 19% was observed.

Meta-analyses of reported effect sizes

- The findings from all three meta-analyses are summarised under each of the five themes. Note the number of evaluations included in the meta-analyses of primary ITT (82 evaluations), secondary ITT (35 evaluations) and FSM (79 evaluations) effect sizes. Caution is needed when interpreting the findings of the meta-analyses because differences may be an artifact of these sample differences.

Effect size and the intervention

- On average, interventions with an English curriculum focus were associated with higher primary ITT effect sizes compared with trials with a maths or cross-curriculum focus. No association was observed between FSM effect size and curriculum focus, but higher secondary ITT effect sizes were observed for interventions that had a maths focus.

- On average, primary ITT, secondary ITT and FSM effect sizes where higher for interventions in primary schools compared with secondary schools. Within primary schools, effect sizes were higher in Key Stage (KS)1 than in KS2. On average, interventions crossing the primary–secondary transition were associated with higher primary ITT and FSM (but not secondary ITT) effect size than interventions in primary or secondary schools. Caution is needed because of the limited number of transition trials and one transition trial reporting an exceptionally high effect size.

- For primary ITT and FSM outcomes, this review reflects earlier findings (Anders et al., 2017) of higher effect sizes for TA-led interventions than for other interventions, and this is likely to be associated with the mode of delivery and fidelity to the intervention. This pattern was not observed for secondary ITT effect sizes.

- On average, primary ITT, secondary ITT and FSM effect sizes were higher for interventions with a total cost of between £250k and £750k compared with interventions that were more or less expensive.

- No association was found between primary ITT, secondary ITT or FSM effect size and intervention intensity (minutes per week). Caution is needed here because of inconsistency in reporting time across the 82 evaluations.

- The perceived quality of the supporting intervention resources was not found to be associated with primary ITT effect sizes, but higher FSM and secondary ITT effect sizes were observed when the reported perceived quality of the supporting intervention resources was high.

Effect size and theory & evidence

- On average, evaluations that drew on strong empirical evidence were associated with a higher primary ITT and FSM effect size, aligned with what might be expected from evidence in the wider field of theory-based evaluation. A different pattern was seen with secondary ITT effect sizes where higher effect sizes were found when empirical evidence was limited or not present.

- No association was found between primary ITT or FSM effect size and the level of theoretical detail presented in the report. However, secondary ITT effect sizes were higher when the theory behind the intervention was highly detailed in the report. It should be noted that the level of theoretical detail is not a proxy for the strength of the underlying theory.

Effect size and context

- The meta-analyses found an association between geographical context and primary ITT effect size that showed higher effect sizes for interventions located in one or up to three geographical areas compared to interventions located in a greater number of geographical areas. A similar pattern was observed for FSM but not for secondary ITT effect. Previous reviews did not consider this effect. It is possible that observed difference relates to greater ease of consistent implementation in smaller geographical areas.

- Evaluations that mentioned the alignment of the intervention and existing practice as an enabler were – perhaps counter-intuitively – associated with lower primary ITT effect sizes, indicating that while the implementation process is likely to be easier when the new intervention is more closely aligned with existing practice, primary

outcome effect sizes are more likely to be higher when the intervention is less closely aligned to existing practice. For secondary ITT and FSM effect sizes, no association between alignment and existing practice was observed.

- On average, evaluations that mentioned staff teamwork as an enabler were associated with higher primary ITT and secondary ITT effect sizes: this may indicate that staff teamwork acts as an indicator of positive school orientation to the intervention. However, no association was observed between reported perceptions of staff teamwork and FSM effect sizes.

- On average, evaluations that mentioned specialist facilities and space or workforce capacity as a barrier were associated with higher primary ITT, secondary ITT and FSM effect sizes.

- No association was observed between primary ITT effect sizes and reported perceptions of pupil behaviour. However, evaluations that mentioned pupil behaviour as a barrier were associated with lower secondary ITT and FSM effect sizes.

- No association was observed between primary ITT and FSM effect sizes and reported perceptions of staff expectations and motivations. However, evaluations that mentioned staff expectations as an enabler were associated with higher secondary ITT effect sizes.

Effect size and implementation & fidelity

- The nine programmes that originated from schools or academy trust developers had the highest weighted mean primary ITT and FSM effect sizes, aligning with the review findings of Anders et al. (2017). However, this pattern was not observed for secondary ITT outcomes, where the eight programmes that originated from councils or local authorities had the highest mean effect size.

- No evidence was found for an association between primary ITT and FSM effect size and whether continuing professional development (CPD) was provided to support implementation of the intervention. Although some form of CPD was reported by a large majority of evaluations (77 of the 82 in the review, 94%), three interventions that did not report any CPD use were associated with higher secondary ITT effect sizes.

- Where CPD was used, the highest average primary ITT effect sizes were observed for trials that delivered CPD pre-intervention only, compared with trials where CPD also took place during an intervention. However, this pattern was not observed for secondary ITT or FSM effect sizes.

- Programmes with CPD that is subject or curriculum-specific are associated with higher average primary ITT and FSM effect sizes, resonating with research on effective professional development which indicates that subject-specific CPD is more likely than generic CPD to be effective in changing teachers' practices. However, this pattern was not observed for secondary ITT effect sizes.

- No associations were found between primary ITT, secondary ITT or FSM effect size and different modes of CPD (i.e., face-to-face training, online training, or coaching or mentoring) or whether or not a train-the-trainer model of CPD was deployed. These findings do not align with existing CPD research findings.

- Fidelity to CPD and fidelity of implementation by the direct implementer (who in most but not all trials would have taken part in the CPD) were examined separately. No evidence was found for an association between primary ITT, secondary ITT or FSM effect size and intended fidelity (for direct implementers) or actual fidelity, indicating that interventions that are intended to be adapted to context may be equally likely to lead to positive effect sizes as those designed to be faithfully adopted. High fidelity relating to CPD was observed to be associated with higher primary ITT, secondary ITT and FSM effect sizes.

- No evidence was found for an association between primary ITT or FSM effect size and whether the developer provided informal support for the intervention beyond formal CPD, or whether the senior leadership team (SLT) were supportive of the programme. Secondary ITT effect sizes were also not associated with SLT support but were associated with the provision of informal support; with the highest mean effect size found when support was provided during the intervention. Caution is needed in interpreting these findings because of the lack of consistent reporting across the 82 evaluations, resulting in missing data.

- Perhaps counter-intuitively, higher primary ITT and secondary ITT effect sizes were associated with evaluations that reported no monitoring of interventions by delivery partners than with trials where robust monitoring was reported. For FSM effect sizes, this pattern was not present. Again, caution is needed in interpretation because monitoring was not reported consistently across trials.

Effect size and evaluation design

- Whilst RCTs became less common over time, they are associated with a higher weighted mean primary ITT, secondary ITT and FSM effect size compared with clustered trials. RCTs with pupil-level randomisation had even higher weighted mean effect sizes compared with CRTs with school-level randomisation.

- No evidence of an association between type of trial (efficacy vs effectiveness) and primary ITT or FSM effect size was observed.[3] However, secondary ITT effect sizes were observed to be higher for efficacy compared with effectiveness trials.

- No association between primary ITT effect size and intervention length[4] was observed. However, both secondary ITT and FSM were associated with the length of intervention and for both the highest mean effect size was observed for interventions lasting from more than two terms up to one year, compared with longer or shorter interventions.

- In terms of sample size, primary ITT effect size was not observed to be associated with number of schools although higher effects were observed for smaller samples. However, in terms of number of pupils, smaller samples (500 or fewer) were associated with higher primary ITT effect sizes. For secondary ITT effect sizes, an association was observed with the number of schools; evaluations involving 20 or fewer having the highest observed effect size. Secondary ITT effect sizes were not associated with the number of pupils involved in the evaluation but the highest mean effect size was observed amongst evaluations involving 501 to 1000 pupils. FSM effect sizes were not found to be associated with sample size in terms of number of schools or pupils. The suggestion that smaller evaluations are associated with higher effect sizes may be a statistical echo of the higher effect sizes observed for RCT compared with clustered RCT evaluations.

- No evidence of an association between primary ITT, secondary ITT or FSM effect size and pupil-level attrition or testing burden was observed.

- On average, trials using commercial test for outcomes were associated with a higher primary ITT and FSM effect size, but lower secondary ITT effect size compared with statutory assessments. The New Group Reading Test (NGRT) developed by GL Assessment was the most common primary outcome measure and had the highest observed weighted mean effect size.

Descriptive analyses of cost effectiveness

A cost effectiveness analysis was conducted for the 40 of the 82 evaluations that reported effect sizes above +0.05 for at least half of reported ITT primary outcomes. Cost effectiveness was defined as the £ per pupil for an effect size of +0.10 standard deviations.

Descriptive analyses of pupil attrition

This variable is also measured at the trial/evaluation level and is the reported pupil level attrition relating to the primary outcome of the trial. The average pupil-level attrition reduced between 2014 and 2018. This decline in attrition rates is at least in part accounted for by a reduced use of commercial tests, along with declining attrition rates for the commercial tests that are used.

Limitations

Given the ambition and novelty of key aspects of this review, there are inevitably important limitations that need to be acknowledged. These limitations also highlight areas for further research. These include:

- *Breadth and nature of the review*: The quantity of explanatory variables selected for inclusion under the five overarching themes reflects the purposely broad nature of the review which brings methodological limitations and a need for careful interpretation of findings. We adopted a random effects model to reflect this diversity in the meta-analyses, but the bivariate nature and number of explanatory variables mean it is not appropriate to draw

[3] Note: the definition of efficacy and effectiveness trials was provided by EEF. Some inconsistency in this classification is noted in *Presenting the explanatory variables*.
[4] Please see Appendix A for how this variable was constructed. This was a combination of UCL/IoE data which was supplemented by 'intervention length' from the EEF website. As there was some variability in reporting this, it should be treated with caution.

causal conclusions from our analyses. The bivariate meta-analyses provide an initial inspection of this effect size diversity and whether/how it is associated with reported primary outcome ITT effect size(s). It is likely that associations found between explanatory variables and effect size(s) will overlap. Future reviews might want to explore this through multivariate meta-analyses.

- *Reliability and validity of the new explanatory variables*: The ambition to produce a somewhat exhaustive list of possible explanatory variables within each theme, along with review time and resource constraints and the significant variation in reporting of implementation and process evaluation (IPE) findings across the 82 trials, inevitably impacted on the reliability and validity of the variables coded for the first time in this study. In addition, a number of variables were difficult to code due to inevitable levels of subjectivity, both in the judgements of the team of coders and the judgements of the evaluators in reporting. A further limitation was that coders were obliged to make coding judgements based on what was reported, rather than on what may have actually taken place but was not adequately reported. A number of codes were set up to assess if a particular variable was perceived to be a barrier or enabler by those involved rather than simply whether or not the variable was mentioned. This was particularly difficult to assess since often there was variation in stakeholders' perceptions between schools and sometimes within schools.

- *Missing data and inconsistent evaluation reporting* was evident across all themes and this problem was largest for the theory & evidence, context and implementation variables. This led to a number of variables being dropped from the analysis that may have provided useful insights.

- *There was particular difficulty in coding against theory & evidence variables*, especially differentiating between 'strong evidence' and 'some evidence' in relation to 'prior evidence of theory'. These variables should be treated with caution and would need to be reconsidered in subsequent reviews of projects.

- *Use of statistical significance:* We accept the limitations of using statistical significance in the analyses. The effect sizes and the evaluations are not random samples. Therefore, the inferential use of statistical significance is not appropriate. We use statistical significance to help illuminate the strength of statistical association in the observed analyses which are descriptive, exploratory and mostly bivariate in nature. Interpretation of the statistical analyses drew on descriptive statistics, discussion in the research team, critical judgement and statistical significance.

- *Secondary outcomes analyses:* Secondary outcomes for the review – effect sizes for ITT analyses of trial secondary attainment outcomes, effect sizes for FSM subsample analyses of primary or secondary attainment outcomes, cost effectiveness and pupil-level attrition – were all identified during the review and following the development of the theoretical framework. This meant that explanatory variables were included in the analyses of secondary outcomes for the review in a post hoc way. Transferring the framework for the secondary attainment and FSM effect sizes was relatively straightforward. However, the pupil attrition outcome did not align well with explanatory variables that captured aspects of an intervention or how it was implemented. Overall pupil-level attrition included pupils in both the intervention and control conditions, whilst some explanatory variables related solely to the intervention condition.

- *Measuring time*: It became apparent that units of time and reporting of dates could be more clearly specified in EEF trials, and the lack of standardised metrics in reporting limits the validity of comparative analysis.

.

Introduction

Background and context

The EEF's mission is to break the link between family income and educational achievement. More specifically, the EEF aims to achieve this mission by:

- raising the attainment of 3–18 year olds, particularly those facing disadvantage;
- developing their essential life skills;
- preparing young people for the world of work and further study.

This is achieved through: summarising the best available evidence in plain language; generating new evidence of 'what works' to improve teaching and learning; and supporting teachers and school leaders to use research evidence to maximise the benefit for young people.

This study aims to support EEF in this mission by providing an analysis of 82 randomised controlled trials (RCTs) commissioned by EEF and reported by January 2019, to identify patterns related to **who** and **what** has been successful and, if possible, present evidence on **why**. The 82 trials in the review include 79 EEF RCTs published between 2014 and 2018 and three published in January 2019. The review focused only on the studies with highest evidence quality threshold, so did not include any of the following: the 20 EEF pilot studies published in this period; evaluations with a quasi-experimental (matched) design; or the Literacy Octopus trial.[5] The review therefore builds on and extends previously commissioned reviews of EEF trials (Slavin, 2016; Anders et al., 2017).

This report contributes significant new knowledge by conducting quantitative meta-analyses across a larger number of trials than previous reviews and extending the range of variables within the analyses.

A further outcome of the study is a contribution to building robust datasets that can be updated when conducting future EEF reviews. The dataset created for this project was based on the coding of the 82 reports described above under thematic areas detailed in this report. This dataset was then synthesised with extracts from:

- the EPPI database of trials being developed by Durham University;
- the database created for Anders et al.'s (2017) EEF review;
- data from the meta-analyses of Lortie-Forgues & Inglis (2019).

The outcomes of quality checks of the EPPI database undertaken by the IoE evaluation team have been fed back to the EPPI team.

Review objectives

The original objectives of the review were as follows:

[5] The Literacy Octopus trial was a large-scale multi-armed trial evaluating a range of very light touch interventions such as the use of printed and online research summaries, evidence-based practice guides, webinars, face-to-face CPD events, and access to online tool run by a range of delivery partners. No impact was observed across 12 effect sizes. In discussion with EEF, this trial was excluded from the review due to its distinct nature, complexity and size (see https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/the-literacy-octopus-communicating-and-engaging-with-research/).

1 To investigate associations between explanatory variables related to the intervention, theory & evidence underpinning the intervention, context, implementation & fidelity and evaluation design,[6] and the effect sizes for the primary outcomes of EEF efficacy and effectiveness trials, cost effectiveness and pupil-level attrition.

2 To develop a review framework, specify research questions, and identify appropriate variables and associated codes to enable the research questions to be addressed.

3 To construct an appropriate dataset and undertake meta-analyses and other quantitative analyses to address the research questions.

4 To draw out claims to support EEF grant making and scale-up, as well as more generalisable claims and reports on areas that would benefit from further exploration and analyses.

Following discussions with EEF, the review was expanded to also address the following objectives:

5 To examine effect sizes reported for secondary attainment, FSM subsample attainment and psychological outcomes.

6 To investigate associations between explanatory variables related to the intervention, theory & evidence underpinning the intervention, context, implementation & fidelity and evaluation design, and the effect sizes for the secondary attainment and FSM attainment outcomes of EEF efficacy and effectiveness trials.

Two further objectives included as part of the review are reported separately:

7 To carry out qualitative interviews with programme team members from the effectiveness trials to explore models and processes perceived to be associated with effective scale-up (Maxwell et al., 2021a).

8 To develop and pilot a measure of IPE quality (Maxwell et al., 2021b).

The review was undertaken in 2019 and 2020.

Research questions

The research questions (RQs) developed during the early stages of the study were:

RQ1 What intervention, theory & evidence, context, implementation & fidelity and evaluation design variables are associated with reported effect sizes for the intention-to-treat (ITT) analyses of primary outcomes in all EEF trials that have been reported up to January 2019?
RQ2 What relevant intervention, theory & evidence, context, implementation & fidelity and evaluation design variables are associated with the cost effectiveness of reported positive impacts across all EEF trials that have been reported up to January 2019?
RQ3 What relevant intervention, theory & evidence, context, implementation & fidelity and evaluation design variables are associated with pupil-level attrition in all EEF trials that have been reported up to January 2019?

A further two questions were included later in the study.

RQ4 What relevant intervention, theory & evidence, context, implementation & fidelity and evaluation design variables are associated with reported effect sizes for secondary ITT analyses of attainment outcomes?
RQ5 What relevant intervention, theory & evidence, context, implementation & fidelity and evaluation design variables are associated with reported effect sizes for FSM subsample analyses of primary or secondary attainment outcomes?

Ethics and data protection

The project received ethical approval from the Faculty of Social Sciences and Humanities at Sheffield Hallam University. All data used in the quantitative analyses have been coded from publicly available sources (i.e., the EEF trial reports). The only personal data held for this study was contact information for the qualitative interviews and interview data. The

[6] The grouping of factors into the five themes included in this objective and the research questions was undertaken during the initial stages of the study, but is presented here for clarity.

legal basis for processing this data is 'Public Task' as defined in Article 6(1e) of the General Data Protection Regulations (GDPR).

Table 1: Project team

Team member	Institution and title	Role/responsibilities
Sean Demack	Sheffield Hallam University, Principal Research Fellow	Co-principal investigator. Quantitative strand lead.
Professor Bronwen Maxwell	Sheffield Hallam University, Head of Commissioned Research	Co-principal investigator. Theoretical framework, operationalising variables and interpretation. Qualitative strand lead. Client liaison.
Professor Mike Coldwell	Sheffield Hallam University, Head of Centre	Co-investigator. Theoretical framework, operationalising variables and interpretation.
Anna Stevens	Sheffield Hallam University, Research Fellow	Project manager. Quantitative strand.
Claire Wolstenholme	Sheffield Hallam University, Research Fellow	Project manager. Theoretical framework, operationalising variables and coding development.
Dr Hugues Lortie-Forgues	University of York, Senior lecturer in Education	Meta-analysis lead.
Dr Sarah Reaney-Wood	Sheffield Hallam University, Research Fellow	Meta-analysis.
Bernadette Stiell	Sheffield Hallam University, Senior Research Fellow	Qualitative fieldwork.
Dr Katie Shearn	Sheffield Hallam University, Research Fellow	Theory & evidence variable development.
Lisa Clarkson, Lyn Pickerel, Rosie Smith	Associate researchers	Report coding.
Noreen Drury, Charlotte Rowley, Hannah Joyce, Kellie Cook	Student researchers	Report coding.

Review framework

This section provides an overview of the outcome measures used in the quantitative analyses and the theoretical framework developed to group the explanatory variables.

Outcome measures

Table 2 summarises the five outcomes used in the quantitative analyses of the review: primary effect size, secondary attainment effect size, FSM attainment effect size, cost effectiveness and attrition. Please see *Presenting the outcome variables* below for statistical detail on each outcome. This section also presents the overall distribution of effect sizes reported for psychological outcomes. The theoretical framework for the review was built with reference to the initial meta-analyses outcome (effect sizes for ITT analyses of primary outcomes), which was adapted for meta-analyses of secondary and FSM effect sizes for attainment outcomes and for the descriptive analyses of cost effectiveness and attrition. However, this was not feasible for effect sizes reported for psychological outcomes because of their very diverse nature. The distribution of effect sizes reported for psychological outcomes is summarised and we provide our thoughts on how these outcomes might be classified within a future review.

Table 2: Summary of outcome variables for the quantitative analyses of the review

Primary outcome	Variable	**Reported effect sizes from ITT analyses of primary outcomes for all EEF trials reported up to January 2019.**
Effect size reported for primary outcomes	Measure	Measured at the effect size level. Effect size (in units of standard deviations) and standard error of effect size (for use in the meta-analyses).
Secondary outcome 1 Effect size reported for secondary attainment outcomes	Variable	Reported effect sizes from ITT analyses of secondary outcomes measuring pupil attainment for all EEF trials reported up to January 2019.
	Measure	Measured at the effect size level. Effect size (in units of standard deviations) and standard error of effect size (for use in the meta-analyses).
Secondary outcome 2 Effect size reported for FSM attainment outcomes.	Variable	Reported effect sizes from FSM subsample analyses of primary or secondary outcomes measuring pupil attainment for all EEF trials reported up to January 2019.
	Measure	Measured at the effect size level. Effect size (in units of standard deviations) and standard error of effect size (for use in the meta-analyses).
Secondary outcome 3 Cost effectiveness	Variables	Measured at the evaluation/trial level. Probability of a trial being included in the cost effectiveness outcome (i.e., trial level probability of reporting a positive impact) in all EEF trials reported up to January 2019. Cost effectiveness of impact for trials that reported a positive impact in all EEF trials reported up to January 2019.
	Measures	Probability (0.00 to 1.00). Cost per pupil for an effect size of +0.10 SD (>0.00).
Secondary outcome 4 Attrition	Variable	Measured at the evaluation/trial level. Reported pupil-level attrition rate for ITT analyses.
	Measure	Percentage of baseline pupil sample included in ITT analyses of primary outcome.

Note: Psychological outcomes were also examined but were unsuitable for inclusion in the meta-analysis.

Effect size

The review began with a specific focus on reported effect sizes for all EEF efficacy and effectiveness trials reported to date (January 2019). This drew on data from the previous Sheffield Institute of Education (IoE) review (Anders et al., 2017) and from Lortie-Forgues & Inglis (2019) and the early stage of the construction of the EPPI database of trials. The

data were checked and completed for all 82 trials by the review team[7] and resulted in the identification of 133 effect sizes reported for the ITT analyses of primary outcome(s) across the 82 trials in the review. The results of this process informed the development of a theoretical framework and the selection of explanatory variables (see below). For this reason, the reported effect sizes for the ITT analyses of primary outcome(s) in all EEF trials published up to January 2019 constitute the primary outcome for the review. This primary outcome was measured at an effect size level and links directly to the first research question.

Follow-on meta-analyses that examined effect sizes other than primary ITT outcomes were also undertaken. In all, a total of 790 reported effect sizes were extracted; 133 related to the ITT analyses of primary outcomes reported by the 82 trials in the review; 78 related to secondary attainment outcomes reported by 35 of the 82 trials; 149 related to FSM attainment outcomes reported by 73 trials and 88 related to psychological outcomes reported by 21 trials. This accounts for 448 of the extracted effect sizes with 342 classed as 'other effect sizes'. The other effect sizes included effect sizes reported for subscales of primary or secondary outcomes; other (not FSM) subsample analyses (e.g., pupils not classed as FSM; male/female; high/medium/low attaining pupils) and measures of attendance/truancy. These are included on the data file but are outside the scope of the follow-on analyses commissioned by EEF.

Cost effectiveness

Through discussions with EEF during the first half of the review period, two additional outcomes were agreed: cost effectiveness and attrition. For an intervention to qualify for inclusion in the cost effectiveness outcome analysis, the evaluation must have reported a positive impact of the intervention; specifically, at least half of all reported effect sizes from the ITT analyses of primary outcome(s) must be above +0.05 SD. Forty of the 82 EEF evaluations in the review met the criteria and were included in the cost effectiveness outcome, which is measured at the trial level and links directly to the second research question.

The construction of the cost effectiveness outcome led to the construction of an additional outcome: the probability of being included in the outcome (i.e., the probability of reporting a positive impact). Across all 82 EEF evaluations in the review, the probability of inclusion in the cost effectiveness outcome was (40/82 =) 0.49. This additional outcome provides a second, trial-level, perspective on the reported impact of programmes evaluated by the first 82 EEF trials, which supplements the more finely grained meta-analyses of the effect size primary outcome. The additional outcome also provides contextual detail to assist interpretation of the cost effectiveness outcome. Finally, the additional outcome reflects how the cost effectiveness outcome draws on data at both effect size level and trial (cost per pupil) level.

The cost effectiveness outcome was identified following the development of the theoretical framework and the selection of explanatory variables. Explanatory variables under each of the five overarching themes were therefore selected into the analyses of cost effectiveness in a post hoc way. This was a pragmatic decision to enable the review to progress in a timely manner within the resources allocated. Future reviews may want to focus more directly on this outcome in developing their theoretical framework and selection of explanatory variables. For this reason, cost effectiveness and the probability of a trial being included in the cost effectiveness outcome are both classed as secondary outcomes for the review.

Attrition

Pupil-level attrition rates were obtained for 79 of the 82 trials in the review. This drew on data from the previous IoE review (Anders et al., 2017) and from Lortie-Forgues & Inglis (2019) and the early stage of the EPPI database of trials. The data were checked and completed for 79 trials by the review team.[8] As with the cost effectiveness outcome, pupil-level attrition was identified as an outcome following the development of the theoretical framework and the selection of explanatory variables. This also meant that explanatory variables under each of the five overarching themes were selected into the analyses of attrition in a post hoc way. This process highlighted an issue of alignment between this outcome and some explanatory variables, particularly under the intervention and implementation & fidelity themes. The attrition rate related to the sample of pupils in both the intervention and control conditions, but these explanatory variables related solely to the intervention and how it was implemented. As with cost effectiveness, future reviews may

[7] This was two researchers looking across multiple data sources and referring to evaluation reports when data were missing and/or in disagreement across the multiple data sources.
[8] This was a single senior researcher looking across the multiple data sources and referring to evaluation reports when data were missing and/or in disagreement across the multiple data sources.

want to focus more directly on pupil attrition in developing their theoretical framework and selection of explanatory variables. Further, future reviews may want to examine pupil-level attrition for the intervention sample and how this might be associated with the intervention and the way it was implemented. For these reasons, attrition is classed as a secondary outcome for the review and links directly to the third research question.

We use the terms primary outcome and secondary outcome specifically to reflect how the quantitative outcomes in the review were identified and constructed and how they influenced the selection of explanatory variables. The primary outcome (effect size) was the focus of the review from inception and was the outcome used to construct a theoretical framework and to select explanatory variables. Secondary outcomes were identified after the development of the theoretical framework and selection of explanatory variables. We do not use these terms to indicate anything other than this. All the quantitative analyses undertaken (primary and secondary outcomes) were exploratory and descriptive. The theoretical framework and resulting themes provided an analytical structure, but the selection of explanatory variables was purposefully broad to reflect the exploratory and descriptive nature of the review.

Methodology

This section outlines the methodology for the quantitative strands of the review. It sets out the theoretical framework for the review, followed by the process for identifying the explanatory variables, operationalising the variables by creating, piloting, refining and implementing a coding frame for new variables, merging these with existing datasets and conducting the quantitative analyses.

Theoretical framework

The theoretical framework for the review was developed to provide an explanatory tool to group and interpret explanatory variables likely to impact on the primary outcomes of EEF trials. The development of the theoretical framework was an iterative process conducted over the earlier stages of the project. Potential themes and subthemes were drawn out from the following sources:

- Previous EEF reviews, implementation science and wider relevant research and evaluation literature. Sources that were particularly influential in developing the theoretical framework and developing new variables and associated codes included Anders et al. (2017), Coldwell (2019), Maxwell et al. (2019), Nilsen (2015), Sharples et al. (2019), Tabak et al. (2012), Vanderkruik and McPherson (2017) and Kok et al. (2016).
- The review team's knowledge and experience in evaluating interventions, including conducting EEF trials, and their own research on evaluation design, implementation and scale-up.
- An initial review of 10 completed EEF trials to ensure coverage of factors likely to impinge on the effect size of the primary outcomes.
- Discussion with EEF colleagues, who brought knowledge and expertise from across EEF trials, previous reviews and other research supported by EEF, in particular on implementation of interventions.

The framework was further refined during the development of the explanatory variables and associated codes, which included further interrogation of the literature and piloting the coding frame (see *Presenting the explanatory variable* section).

The framework that was operationalised comprises five main themes: the intervention; theory & evidence; context; implementation & fidelity; and evaluation design. Each of these themes was divided into further subthemes (see Figure 1). Two of the themes, intervention and implementation, together comprise the programme that is the focus of each trial. The intervention theme includes characteristics of the intervention that are experienced directly by the target group; for example, this includes the subthemes of 'intensity' (such as how many lessons a pupil engages with as part of the intervention) and 'EEF intervention theme area' (such as language and literacy). The implementation theme includes characteristics of the overall programme that is subject to the trial, which are necessary to enable the direct implementers to deploy the intervention with the target group but are not experienced directly by the target group. For example, this includes the subtheme of professional development for teachers and the support of SLTs in schools where the trial is taking place. In evaluation reports and studies these two themes are sometimes conflated. We assert that it is clearer to consider intervention and implementation as two separate themes, given that the assumed causal processes and mechanisms underpinning each are distinctly different in most EEF trials.

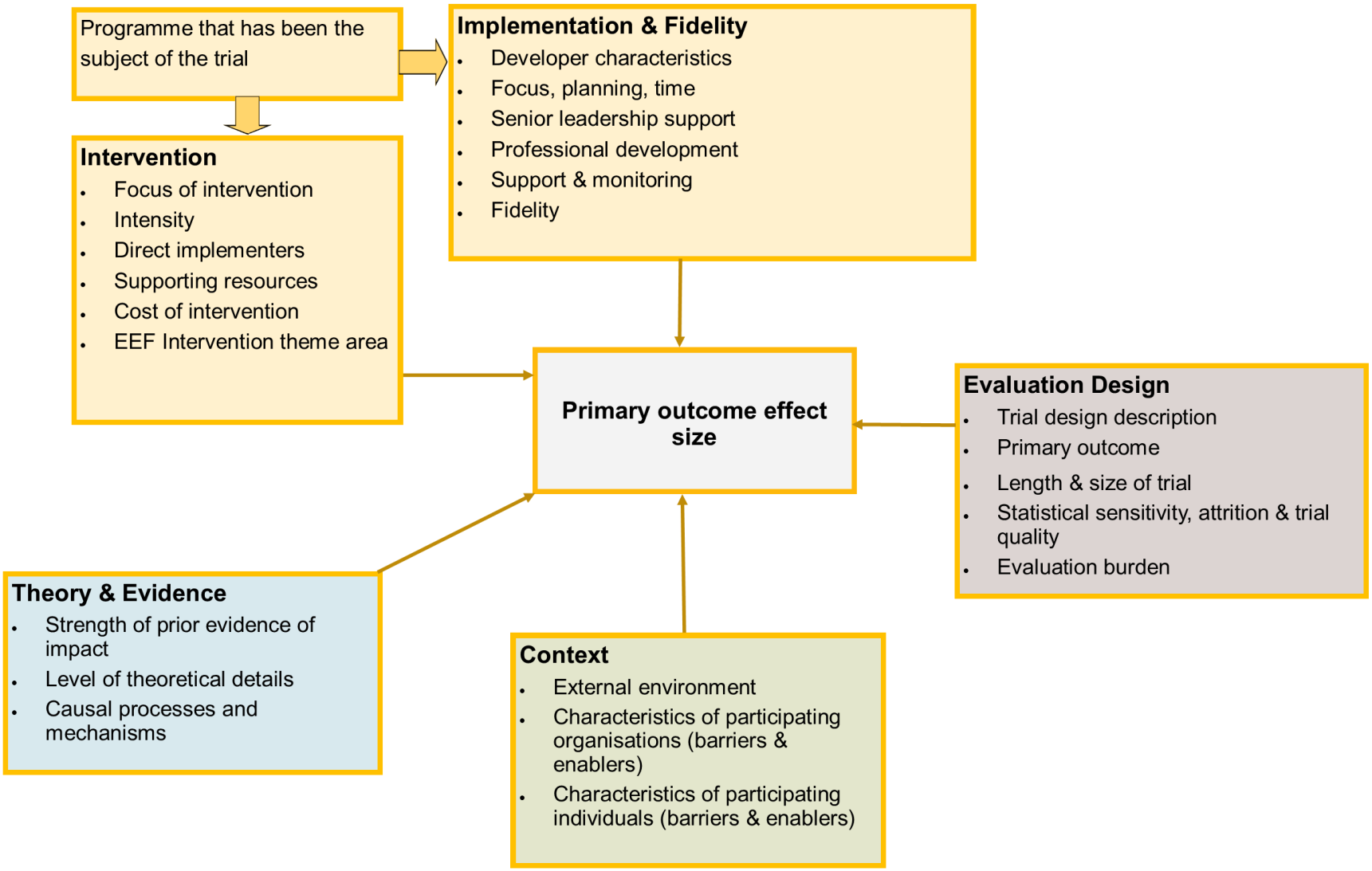
Fidelity relates to the extent to which the programme (i.e., the intervention experienced by the target group; the intervention theme), and the indirect programme activities (implementation theme) was intended to be implemented faithfully or was designed to be adapted to context and whether it was implemented as intended. For presentation purposes we have grouped implementation and fidelity into one theme: implementation & fidelity.

The theory & evidence theme relates to the extent to which the programme being trialled relates to existing theory, the strength of prior research evidence and the nature of the causal processes and mechanisms assumed to underpin the programme.

The context theme brings together contextual variables that may support or impede successful outcomes. IT is divided into three subthemes: the characteristics of the institutions (mainly schools) involved; characteristics of individuals involved; and the external environment.

The evaluation design theme groups explanatory variables associated with evaluation that may indirectly impact on the primary outcome effect size, such as evaluation burden and trial design.

Figure 1: Theoretical framework – overarching themes and subthemes



Identifying the explanatory variables

The identification of explanatory variables and construction of associated codes was undertaken as part of the iterative process of constructing the theoretical framework, the overarching themes and subthemes, and drew on the sources listed in the previous section.

Further detail is provided below in relation to identifying the explanatory variables and associated codes within each of the overarching themes in the theoretical framework.

The intervention, implementation & fidelity and context themes

A similar approach was taken to drawing up a list of explanatory variables within the intervention, implementation & fidelity and context themes. A long list of potential variables, intended to be as exhaustive as possible, was constructed for each theme. The potential variables were linked directly to prior evidence. Evidence for the intervention and implementation themes was drawn particularly from Anders et al. (2017), Slavin (2016) and Sharples et al.'s (2019) evidence-informed implementation guide to identify the variables and develop hypotheses about their likely impact on primary outcomes. Prior evidence for the contextual variables (factors associated with 'the structural – organisational, spatial and temporal – setting and the individuals involved, including their personal characteristics and inter-personal relationships' (Coldwell, 2019, p.102) drew particularly on Vanderkruik and McPherson's (2017) contextual factors framework, as well as Anders et al. (2017) and Coldwell (2019). Where our experience and discussions with EEF colleagues indicated that there may be gaps in the list of potential variables, new items were added to the long list and we undertook further investigations of the literature to support our hypotheses.

The long lists of variables were refined though grouping (into what became the subthemes within each overarching theme), revision to eliminate duplication within themes and across themes, and consideration of the feasibility of using the variables in this study. Feasibility was ascertained by first checking if the variable was already coded within one of the existing datasets that were being merged for the purpose of this study and, if not, by reviewing 10 EEF trials to make an initial assessment of the likelihood of sufficient data in the trial reports to support coding of the variable.

The refined list of variables grouped by overarching theme and subtheme remained purposively broad during the trial report coding phase (see below), to enable assessment of the feasibility of creating new variables from trial reports as well as exploring and describing patterns in intervention, implementation and contextual factors.

Lists of variables created for this review were further reduced following coding and review of univariate statistics as described in the *Operationalising the new explanatory variables* section

Theory & evidence theme

While the identification of explanatory variables for the intervention, implementation & fidelity and context themes drew strongly on prior evidence particularly from implementation science and the earlier EEF reviews, a more exploratory aspect of this review was to consider the theoretical and prior empirical evidence underpinning each of the 82 trials. There were three elements of this exploration:

1 Examining the presence of explicit theory and prior empirical evidence, and evidence for it

This element was the most straightforward to articulate, and has been included in the analyses, via the creation of two explanatory variables:

- the strength of prior evidence of impact of this or similar interventions measuring the range of evidence of impact provided in the report;
- the level of detail provided in the report about how and why the intervention will lead to the intended outcomes/impact (e.g., explicated in a visual logic model or theory of change).

2 Examining the causal process

We hypothesised that differing causal processes – combinations of underlying causal mechanisms with implementation processes – may be related to different outcomes. As indicated in 3 below, we were not able to develop an

operationalisable set of causal mechanisms, but we did develop a causal process description drawing on three variables that were included in the review:

- **Direct or training-based:** many EEF trials involve a professional development or *training* element to change practices in schools (usually of teachers or TAs) which is then expected to lead to the proximal outcomes in the immediate beneficiary (usually the pupil). Some programmes involve *direct* work from the delivery team including, for example, direct teaching provided by employees or consultants working for the delivery team, or provision of a resource such as financial reward direct to pupils.

- **The direct implementer:** usually the teacher, but sometimes a school leader, another member of school staff (typically a TA), an external coach or tutor, or a parent/carer. In some cases, a number of direct implementers are involved.

- **The main focus of change**: usually this is pupil learning, but sometimes the focus is on wider pupil outcomes, such as pupil behaviour and attitudes, or even teacher or leader change.

These three variables together yield a description of a change process: for example, the most common EEF change process combining these three variables is a **training-based, teacher-led, pupil-learning-focused intervention**. There are also many pupil-learning-focused interventions that are led by TAs and external coaches. There are some interventions that aim to change teacher practice without any specific theory of how this will improve pupil learning, and at least one programme that focused on leader change without any clear theorisation of what this would lead to in relation to teacher or pupil change.

A review of 20 EEF trials confirmed the feasibility of creating a dataset that included the two variables addressing explicit theory and prior empirical evidence and the three variables concerning causal processes set out above. However, we were not able to pursue this further within the scope of this study. To avoid confusion, we have not reported analyses for the direct/training-based variable as the variable definition is close to, but is not fully aligned with, some other variables in the intervention and implementation themes.

3 Description and meaningful allocation of causal mechanisms

This was the most exploratory element. We defined mechanisms as descriptions of causal processes that are expected to be enacted during the implementation of an intervention. In earlier work (Coldwell & Maxwell, 2018), we argued that such mechanisms can be both social (involving relations between individuals and groups) and psychological, as Lacouture et al. (2015, p. 1) suggest: ‘an element of reasoning and reactions of agents in regard to the resources available in a given context to bring about changes through the implementation of an intervention’. An initial examination of literature drawing on realist evaluation (which aims explicitly to draw out causal mechanisms; Pawson & Tilley, 1997) and implementation science, where there are a number of categorisations of implementation theories (such as Nilsen, 2015; Kok et al., 2016), did not produce a meaningful set of potential causal mechanisms. Subsequently, an exploratory assessment of 20 randomly chosen EEF trials did not yield any obvious categorisations within the time and resource constraints of this study.

As causal mechanisms could not be included in the review analyses, we examined the potential to conduct a form of realist review (Pawson et al., 2005) examining the causal mechanisms involved in projects and the contextual arrangements associated with success. The *Technical Annex* to this report lays out an initial categorisation of studies that involved external delivery. Although it was not possible to fully develop this analysis within the ambit of the current study, the categorisation does illustrate a possible approach and EEF may wish to consider commissioning a realist review to draw out the intervention and implementation characteristics that are likely to be associated with success in different groupings of EEF studies (e.g., externally delivered projects, or those delivered by TAs). This could complement the *Implementation Guidance Report* (Sharples et al., 2019) and potentially provide more precise implementation guidance for specific types of study.

Evaluation design theme

The evaluation design variables were refined and reduced following the same process as for the intervention, implementation & fidelity and context variables. First, most variables and data for this theme were drawn from the EPPI database of trials, the Anders et al. (2017) dataset or the Lortie-Forgues & Inglis (2019) dataset. Second, these variables were checked and updated using EEF evaluation reports. Third, EEF provided data that determined whether trials were

classed as efficacy or effectiveness[9] trials. Fourth, a few additional variables were identified for inclusion in this theme. Additional 'factual' variables included more finely grained measures of the trial primary outcome(s) and predicted statistical sensitivity of a trial (i.e., the minimum detectable effect size; MDES). Further variables, including how closely aligned the primary outcome was to an intervention, and the evaluation burden in terms of testing and IPE data collection activities, were drawn from judgements by reviewers of evaluation reports.

Operationalising the new explanatory variables

Where possible, explanatory variables were sourced from the EPPI database of trials, the Anders et al. (2017) dataset or the Lortie-Forgues & Inglis (2019) dataset. This section summarises the process for operationalising the explanatory variables included in the theoretical framework but were not found in the existing datasets.

The parameters of each new explanatory variable were clearly defined to facilitate consistency throughout the coding process. Decisions on each variable's definition and the identification of associated codes drew on the earlier literature reviews and review of EEF trial reports. Where possible, standardised codes were adopted in order to make the coding process manageable and to maximise reliability given the large numbers of variables coded. For example, for most contextual theme variables the coding format was whether the issue was perceived as: 'barrier', 'both barrier and enabler', 'enabler', or was 'not mentioned' or 'unclear'. More specific coding options were created where necessary. For example, the codes for 'geographical location' were: 'national', 'one geographical location', 'two or three geographical areas' and 'other'. The coding frame was carefully refined and discussed in detail in a number of internal meetings and in liaison with EEF. All new variables' definitions and codes are set out in the full details of variables (see Appendix A).

For coding purposes, the codebook was operationalised as an Excel spreadsheet containing the variables and their definitions and, where necessary, code definitions. Coders were required to select a code using a drop-down menu within the coding sheet and to record 'evidence' from the report to justify their choice.

The first stage of piloting the codebook involved members of the core evaluation team each coding two reports and comparing their responses. This led to further refinements of the codebook, including adding further detail to variable and code definitions and introducing new coding options, such as 'not mentioned' or 'unclear'. These changes were based on our comparisons of inter-coder reliability, shared understanding of the precise meanings of the variables and codes, and whether the codebook 'worked' to code the example report.

After this first refinement stage, the full coding team (eight people) attended a two-hour training session to ensure understanding of each variable's definition and related codes and the coding process. The second stage of piloting the codebook was then undertaken. The coding team were asked to code two reports (the same as completed by the core team). Their efforts were then compared across each other and to the core evaluation team's coding. The core evaluation team made an overall judgement about the 'correct' code for each variable of the report in question and the appropriate level of detail for the evidence in the coding notes to support the coding decision. This led to further changes to the coding frame and also helped to assess the coding teams' ability to interpret and understand the variables and associated codes. The finalised codebook was sent to each member of the coding team, as well as a copy of the 'correct' codebooks for the piloted reports together with individual notes identifying any miscoding and clarifying what was being asked for. Advice was also provided on the appropriate levels of evidence to be recorded. At this stage, coders were asked to code two of the reports they had been allocated and return these to the core evaluation team, who checked the coding and evidence recorded, made any corrections and provided further guidance to the coder.

Upon completion of the coding by the wider team of coders, the coding data were amalgamated, and an initial univariate analysis of variables was conducted in order to facilitate the assessment of the reliability and consistency of each variable. At this stage, a number of variables were dropped due to a large volume of the answers being recorded as 'unclear' or 'not mentioned', or due to there being little or no variation in responses. For the remaining variables the core evaluation team looked at sections of the coding across a random sample of 10 reports, and undertook further back-checking to ensure consistency across responses.

[9] There was some discrepancy in the definitions of efficacy and effectiveness trials. For example, referring to the 'Grammar for Writing' first (efficacy) trial, at the top of the EEF webpage it is stated that ‘This page covers the first (efficacy) trial of Grammar for Writing’ whilst lower down this page in the 'Evaluation Info' table the trial is also listed as an 'effectiveness' trial.

The process described above identified that certain variables needed checking across all reports, which was undertaken by the core evaluation team. This led to a further reduction in variables, for example where there were issues with a variable's definition that could not be resolved through back-checking, or where the data underpinning the coding decision was insufficient. The list of operationalised variables included in the final analysis is presented in Appendix A, including the definitions of the variables and codes in the final list. The list of variables omitted from the final analysis is presented in Appendix B.

Once the final list of variables had been ascertained, some of the categories of variables were collapsed, where there were low numbers in the categories. Similarly, certain sets of variables were amalgamated into a single variable where appropriate; for example, the numerical data recorded for the IPE burden variables were brought into a single 'IPE' burden variable.

At this stage, for analyses of secondary outcomes, the refined and checked new variables were brought together with the trial-level data file that was consolidated from data obtained from the EPPI database of trials, the previous IoE review (Anders et al., 2017), and Lortie-Forgues & Inglis (2019). This trial-level detail was then added to the consolidated effect-size-level data file for the analyses of the primary outcome (i.e., the reported effect sizes for ITT analyses of trial primary outcome(s)).

Meta-analyses of effect sizes

The primary outcome for this review was the reported effect size(s) from ITT analyses of the primary outcome(s) for all 82 evaluations in the review. Analyses were undertaken at the effect size level (i.e., the 133 primary ITT effect sizes reported by the 82 evaluations). The meta-analyses were extended to include reported effect size(s) from ITT analyses of secondary outcomes and FSM subsample analyses of primary and secondary outcome(s).

Analyses of effect sizes were conducted in two stages. First, effect sizes were examined statistically and graphically using standard descriptive statistical methods. Statistically, unweighted mean effect sizes are shown across categories of the explanatory variables under each of the five overarching thematic areas. Differences between the mean effect sizes across categories are tested using an ANOVA, and the strength of association with an explanatory variable is calculated using the eta-squared effect size measure. Given evidence of a positive skew in the distribution of effect sizes (see below), the unweighted mean effect sizes are supplemented with unweighted median effect sizes. ANOVA tests are supplemented by comparable non-parametric tests.[10]

Second, meta-analyses of effect sizes were undertaken using a random effects model. This was done to ensure that larger trials (whose effect sizes are measured with greater precision) were given more weight in the analysis. Using a random effects model is common practice in meta-analyses (see, e.g., Borenstein et al., 2010). This type of model is suitable for the present study where the effect sizes come from trials that vary substantially in terms of sample size, interventions tested, outcomes used, and participants (factors that are known to influence the magnitude of effect sizes; e.g., Cheung & Slavin, 2016). This is because a random effects model does not assume that all effect sizes are estimates of a single unique population effect (an assumption made by the other model commonly used in meta-analysis, the fixed effect model). Rather, a random effects model assumes that each trial is estimating a distinct population effect drawn from a distribution of population effects (Higgins & Green, 2011).

All the analyses involving random effects models were conducted with the R package *Metafor* (Viechtbauer, 2010). This is a free R package that is used widely in the academic literature (as of 6 November, 2019, the descriptor of the package has been cited more than 4950 times). It is also used by Steve Higgins and his team in the EPPI database of trials. *Metafor* provides various methods to quantify the amount of heterogeneity between trials – an essential step in the computation of a random effects model. In the present analyses, we used the method 'restricted maximum-likelihood estimator', which is the default method used in the package. All meta-analyses were replicated and compared for reliability purposes.

[10] The non-parametric tests used were Mann-Whitney U (when an explanatory variable had just two categories) and Kruskal-Wallis Z (when explanatory variables had three or more categories).

A post hoc power analysis for the meta-analyses was undertaken (Borenstein et al., 2009). This highlighted the value of the meta-analysis method in providing notable gains in statistical sensitivity/power compared with individual trials. Across the 133 effect sizes in the review, a statistical power of 0.80 or higher was estimated for detecting an effect size being statistically different from zero as statistically significant ($p < 0.05$, two tailed) for a weighted mean effect size of +0.022 SD or higher.[11]

In the report, we give precedence to the findings from the meta-analyses because the estimates are weighted to take account of variations in statistical uncertainty across the 133 effect sizes. We use the unweighted means and medians to supplement the meta-analyses for further descriptive detail within the *Technical Annex*.

For the additional two groupings of effect sizes (secondary and FSM attainment outcomes), the same approach was adopted as outlined above for primary ITT outcomes.

Please see Appendix D for more statistical detail on the meta-analysis approach adopted for this review.

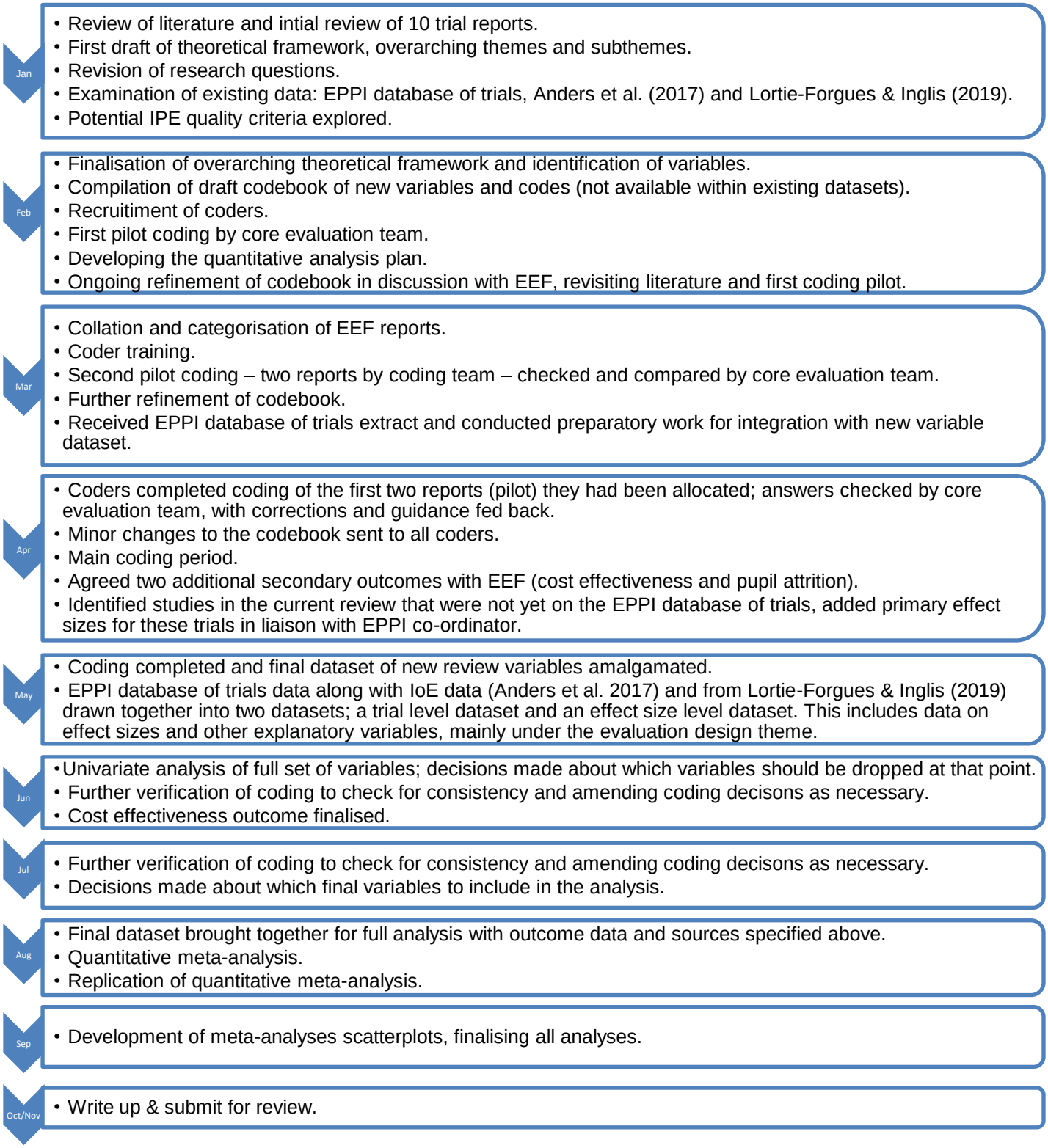
Descriptive analyses of evaluation/trial level outcomes: cost effectiveness and pupil-level % attrition

Analyses of these outcomes was undertaken using standard descriptive statistical methods. A positive skew was observed in both of the secondary outcomes (see below) and therefore parametric statistical methods (e.g., mean, ANOVA) were supplemented with non-parametric methods (median, Mann-Whitney, Kruskal-Wallis). The positive skew was most pronounced in the cost effectiveness outcome, where only a quarter of trials included in the cost effectiveness outcome have cost effectiveness at the mean (£150) or higher; three quarters of trials included had cost effectiveness below £150. This illustrates how skew can result in making the mean a misleading measure of 'average'; this is why the median is used to provide a second perspective on 'average' that is not undermined by skew. Whilst this approach brings complexity in terms of the quantity of statistics, it does avoid unhelpful complexity brought by the abstraction of transformation.

[11] Or a weighted mean effect size of −0.022 SD or lower.

Timeline

Figure 2: Timeline of study (2019)



Following this initial review, EEF commissioned further meta-analyses for the secondary ITT attainment and FSM attainment outcomes. These were undertaken in the second half of 2020 and the findings from all strands were brought together in April/May 2021.

Presenting the outcome variables

Overview

The quantitative analyses in this review focused on five outcome measures:

Primary outcome for the review:

1 Reported effect sizes for ITT analyses of trial/evaluation primary outcome(s).

Secondary outcomes for the review:

2 Reported effect sizes for ITT analyses of trial/evaluation secondary attainment outcome(s).

3 Reported effect sizes for FSM subsample analyses of trial/evaluation primary / secondary attainment outcome(s).

4 Cost effectiveness (£ per 0.10 SD effect size).

5 Pupil-level attrition (%).

The development of the cost effectiveness outcome resulted in an additional secondary outcome:

6 The probability of an EEF evaluation reporting a positive effect.[12]

This section introduces the primary and secondary outcomes and the quantitative analytic approach used for each. Each outcome is described statistically and the change over time (2014–19) is examined. When possible, the distribution for the outcome is compared with data from the previous IoE review of EEF trials.[13]

For the analyses of all outcomes, the following conventions have been applied to help identify key or interesting explanatory variables. Statistical significance is considered at three levels: $p < 0.01$ (indicated by ***); $p < 0.05$ (indicated by **) and $p < 0.10$ (indicated by *). When we judge that an observed pattern is interesting but it was not found to be statistically significant, this is indicated by #.

Reported effect sizes for ITT analyses of primary outcome(s)

There are a total of 133 effect sizes for headline ITT analyses of primary outcome(s) across the 82 trials. Most trials reported a single effect size (50 trials, 61%), but 32 trials reported more than one effect size. The use of a single (primary outcome) effect size was observed to increase over time from 29% in 2014 (eight trials) to 79% in 2018 (15 trials). All three of the trials that reported in January 2019 reported a single primary outcome effect size.

Table 3 and Figure 3 summarise statistically and graphically (respectively) the 133 effect sizes reported for primary outcomes for the main (headline) ITT analyses within the 82 trials included in the review.

The average effect size across all 133 effect sizes in the SHU review is smaller than reported by Anders et al. (2017). This is the case when analysing all 133 effect sizes (effect size level) or using a trial-level mean effect size for the 82 evaluations in the review.[14] This suggests that effect sizes have become smaller over time and this is apparent when looking at effect sizes by publication year: the mean effect size decreased from +0.08 SD in 2014 to +0.04 SD in 2018 and the three 2019 trials included had a mean of −0.02 SD.

[12] Specifically, this is the probability of a trial reporting a majority of effect sizes above a threshold of +0.05 SD (see cost effectiveness section below).
[13] Data from Anders et al. (2017) *EEF Projects Review* – but excluding evaluations with a quasi-experimental design.
[14] For trials with a single primary outcome, the trial and effect size level will be identical. Trials with multiple primary outcomes will have multiple effect sizes but a single trial-level mean. The IoE review used trial-level means but the SHU review draws on the specific effect sizes.

The longer upper tail observed in the dot plot in Figure 3 is an illustration of the positive skew in the effect size distribution. This positive skew highlights a need for caution when using parametric statistics or tests.[15] For this reason, parametric statistics (the mean) are supplemented with non-parametric statistics (the median) within the analyses. When examining the median, a pattern of declining effect sizes over time is also observed: the median effect size decreased from +0.04 SD in 2014 to +0.01 SD in 2018 and the three 2019 trials included had a median of −0.02 SD.

The meta-analyses provide a weighted mean that takes account of the statistical uncertainty for each effect size in its calculation. This is done to ensure that effect sizes from more robust trials (in terms of size and statistical sensitivity) are given more influence in the calculation of weighted mean effect sizes. From the meta-analyses, the weighted mean effect size for the 133 effect sizes in the review was +0.04 SD (95% CI: +0.03 to +0.06). The weighted mean is also seen to reduce over time from a weighted mean of +0.05 SD in 2014 to +0.02 SD in 2018 with the three 2019 trials included having a weighted mean of −0.02 SD.

The overall significance of a variable is based on an omnibus test – a test that evaluates whether the different levels of the variable differ from each other. This type of test is routinely used when comparing multiple means (e.g., ANOVA). A significant omnibus test indicates that at least two of the levels of a variable differ (Peck et al., 2015). Taking, for example, a variable with three levels (A, B and C), a significant omnibus test would indicate that at least two of the levels differs significantly; A may differ from B, B may differ from C, A may differ from C, or alternatively, each level may differ from each other. In the project, a significant test does not indicate that one or more levels of a variable significantly differ from zero.

Table 3: 133 reported effect sizes for ITT analyses of 82 EEF trials in the review: descriptive/unweighted analyses of effect sizes

	Number of trials*	Number of outcomes / effect sizes	Unweighted median (IQR)	Unweighted mean (SD)	Min	Max
UCL/IoE review	47	47	+0.06	+0.10 (0.162)	−0.14	+0.74
SHU review (trial level)	82	133	+0.03	+0.07 (0.135)	−0.13	+0.74
SHU review (ES level)	82	133	+0.03	+0.06 (0.128)	−0.14	+0.74
Year of publication (ES level)						
2014	16	28	+0.04	+0.08 (0.171)	−0.14	0.74
2015	20	32	+0.03	+0.07 (0.132)	−0.11	0.43
2016	16	33	+0.03	+0.07 (0.138)	−0.13	0.51
2017	10	18	+0.03	+0.04 (0.066)	−0.08	0.15
2018	17	19	+0.01	+0.04 (0.080)	−0.09	0.19
2019	3	3	−0.02	−0.02 (0.025)	−0.04	0.01

*Excludes quasi-experimental designs.

Figure 3: Dot plot: distribution of 133 effect sizes (effect size level)

[15] The effect size distribution significantly differs from a Gaussian/normal distribution; 1-sample Kolmogorov–Smirnov test *p*-value < 0.01.

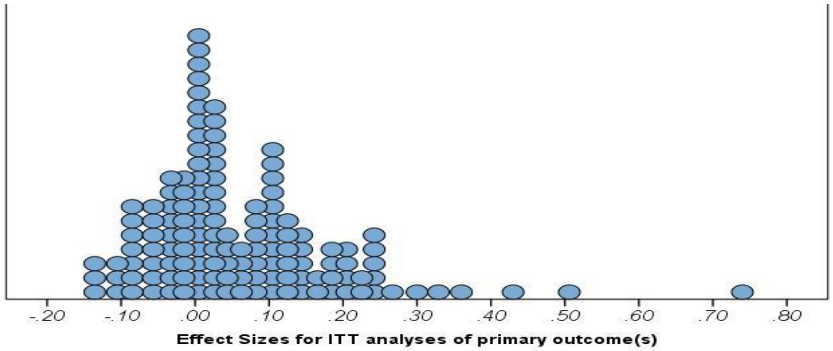


Table 4: Weighted meta-analyses of effect sizes

	Number of trials*	Number of outcomes / effect sizes	Weighted mean (SE)	Lower	Upper
SHU review (ES level)	82	133	+0.04 (0.008)	+0.03	+0.06

Secondary outcome: reported effect sizes for ITT analyses of secondary attainment outcome(s)

There was a total of 78 effect sizes for ITT analyses of secondary attainment outcomes reported by 35 of the 82 trials in the review (43%). 13 trials reported a single effect size (37% of trials reporting secondary attainment outcomes), 15 trials reported two (43%) and seven trials reported three or more secondary ITT effect sizes (20%). Note that secondary ITT effect sizes for attainment outcomes were reported by less than half of the 82 trials included in the meta-analyses of primary ITT attainment outcomes.

Table 5 and Figure 4 summarise statistically and graphically (respectively) the 78 effect sizes reported for secondary attainment outcomes. Table 6 presents the meta-analysis weighted mean and 95% CI for secondary ITT outcomes.

Table 5: Effect sizes reported for secondary ITT attainment outcomes

	Number of trials*	Number of outcomes / effect sizes	Unweighted median (IQR)	Unweighted mean (SD)	Min	Max
Secondary outcome ITT attainment effect sizes	35	78	+0.01	+0.01 (0.132)	−0.30	+0.38
Year of publication (ES level)						
2014	4	6	−0.05	−0.03 (0.098)	−0.13	+0.15
2015	5	8	−0.01	+0.07 (0.166)	−0.07	+0.38
2016	7	22	+0.01	+0.02 (0.144)	−0.28	+0.33
2017	7	18	+0.04	+0.05 (0.087)	−0.06	+0.24
2018	10	19	0.00	−0.04 (0.148)	−0.30	+0.17
2019	2	5	+0.02	+0.01 (0.041)	−0.06	+0.04

*Excludes quasi-experimental designs.

The weighted mean effect size for secondary ITT outcomes (+0.01 SD) was smaller than observed for primary ITT outcomes (+0.04 SD). Further, unlike primary ITT outcomes, the weighted mean effect size for secondary ITT outcomes was not statistically significantly greater than zero.

Figure 4: Dot plot: effect sizes reported for secondary ITT attainment outcomes

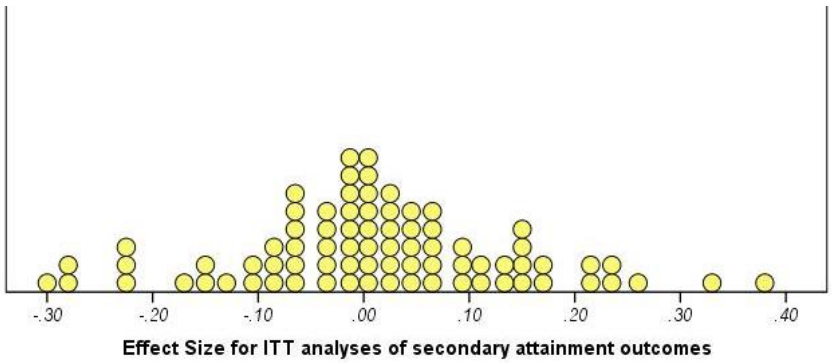


Table 6: Secondary ITT attainment outcomes; meta-analyses weighted mean and 95% CI

	Number of trials	Number of outcomes / effect sizes	Weighted mean (SE)	95% CI	
				Lower	Upper
Secondary ITT outcomes (attainment)	35	78	+0.01 (0.010)	−0.01	+0.03

Secondary outcome: reported effect sizes for FSM subsample analyses of primary / secondary attainment outcome(s)

There was a total of 149 effect sizes for FSM subsample analyses of primary or secondary attainment outcome(s) reported by 73 of the 82 trials in the review (89%).

Just under half of these 73 trials reported a single FSM effect size (34 trials, 47% of trials reporting an FSM effect size for a primary or secondary attainment outcome); 20 (27%) reported two FSM effect sizes and the remaining 19 (26%) reported three or more FSM effect sizes.

Table 7 and Figure 5 summarise statistically and graphically (respectively) the 149 FSM effect sizes.

The dot plot shows three outliers: two low outliers (–1.07 and –0.94 SD, both Nuffield Early Language Intervention (NELI) trial[16]) and one high outlier (+1.50 SD, from the first IPEEL trial).

The first IPEEL trial had a low EEF security rating (two padlocks) and (associated) very wide CI (+0.21–2.98 SD).[17] The NELI trial had four padlocks and is identified as a promising project by EEF – so there is less justification for not including the low outlier – although the evaluation report does question the accuracy of FSM effect size estimates.[18] The NELI trial reported two FSM effect sizes for two levels of the intervention (20 weeks and 30 weeks) for both primary (Language composite score) and secondary (word-level literacy score). In both cases, the reported FSM effect sizes at 20 weeks were not significantly different from zero but were significant and negative at 30 weeks.

Both of these evaluations were included in the analyses of primary ITT effect sizes. Additionally, the analyses below indicate that including these two evaluations does not result in changing the meta-analysis weighted mean effect size for FSM subsample analyses. Therefore, for consistency purposes, effect sizes from these two evaluations are included in the FSM meta-analyses.

[16] See Table 13 in the NELI evaluation report (pp. 31–32).
[17] The primary ITT effect size for the first IPEEL trial was also very high and positive (+0.74; maximum value amongst primary ITT effect sizes).
[18] From page 31 of the NELI report 'estimated treatment effects suggest both treatments are considerably less effective for pupils eligible for FSM: the treatment effect estimates for this subgroup are very large and negative, albeit insignificant. The combination of the subsample imbalances and the lack of a statistically significant effect cast further doubt on the ability of the trial to accurately reveal the effectiveness of the treatments on FSM-eligible pupils.'

Table 8 shows a statistical snapshot for distribution of the effect sizes reported from FSM subsample analyses of trial primary outcomes, that include and exclude FSM effect sizes reported by the NELI and/or the first IPEEL evaluation. This illustrates that excluding the five FSM effect sizes from the first IPEEL and NELI evaluations does not change the overall average, but does notably reduce variance around this average. The remaining 144 FSM effect sizes varied between a minimum of −0.30 and +0.48.

Table 7: Effect sizes reported for primary or secondary attainment outcomes for FSM pupil subsample

	Number of trials*	Number of outcomes / effect sizes	Unweighted median (IQR)	Unweighted mean (SD)	Min	Max
Reported FSM effect sizes	73	149	+0.02	+0.04 (0.234)	−1.07	+1.60
Reported FSM effect sizes excluding trials with high/low outliers	71	144	+0.02	+0.04 (0.152)	−0.30	+0.48
Year of publication (ES level)						
2014	13	19	+0.08	+0.16 (0.402)	−0.30	+1.60
2015	18	39	+0.07	+0.10 (0.158)	−0.21	+0.42
2016	15	34	−0.01	−0.06 (0.290)	−1.07	+0.40
2017	9	23	+0.04	+0.05 (0.080)	−0.11	+0.22
2018	15	31	0.00	−0.02 (0.123)	−0.25	+0.25
2019	3	3	+0.05	0.00 (0.127)	−0.14	+0.10

*Excludes quasi-experimental designs.

Table 8 shows meta-analysis weighted mean FSM effect sizes for the complete sample and for the sample excluding the two trials with high/low outliers. This shows that the inclusion of the trials results in no observed difference on the estimated weighted mean effect size (+0.03 SD) which was significantly greater than zero but slightly smaller than the weighted mean effect size observed for primary ITT outcomes (+0.04 SD).

Table 8: FSM attainment outcomes; meta-analyses weighted mean and 95% CI

	Number of trials	Number of outcomes / effect sizes	Weighted mean (SE)	95% CI	
				Lower	Upper
Reported FSM effect sizes	73	149	+0.03 (0.010)	+0.01	+0.05
Reported FSM effect sizes excluding trials with high/low outliers	71	144	+0.03 (0.010)	+0.01	+0.05

Figure 5: Dot plot: effect sizes reported for FSM attainment outcomes

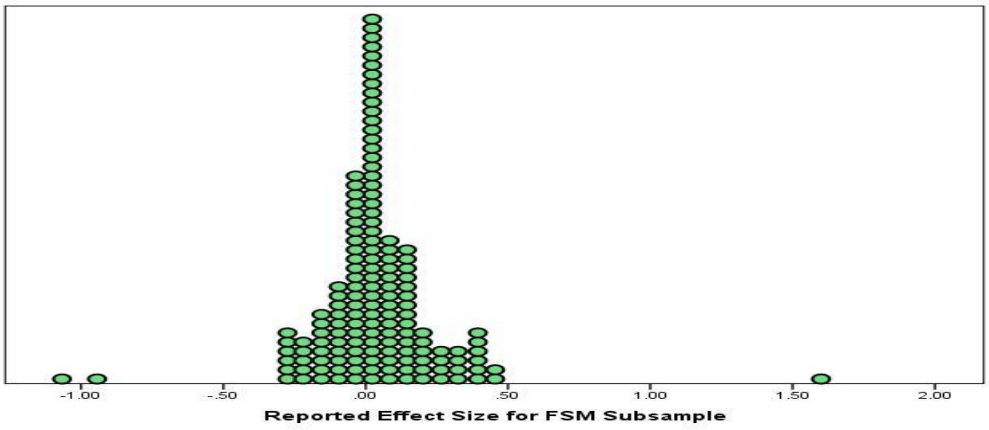
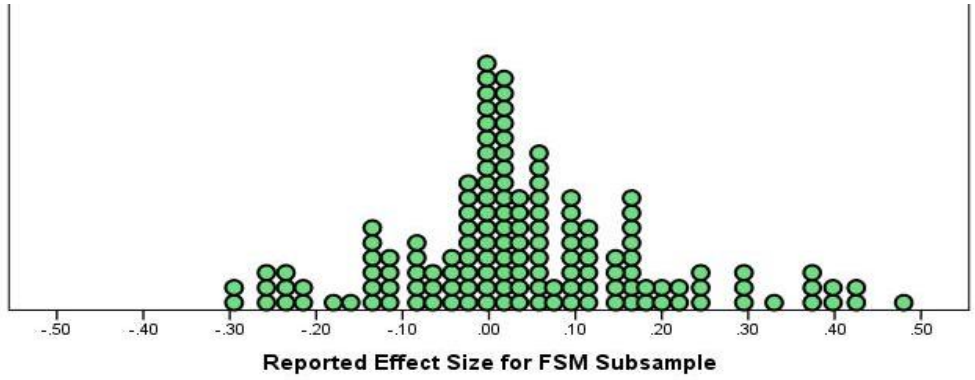


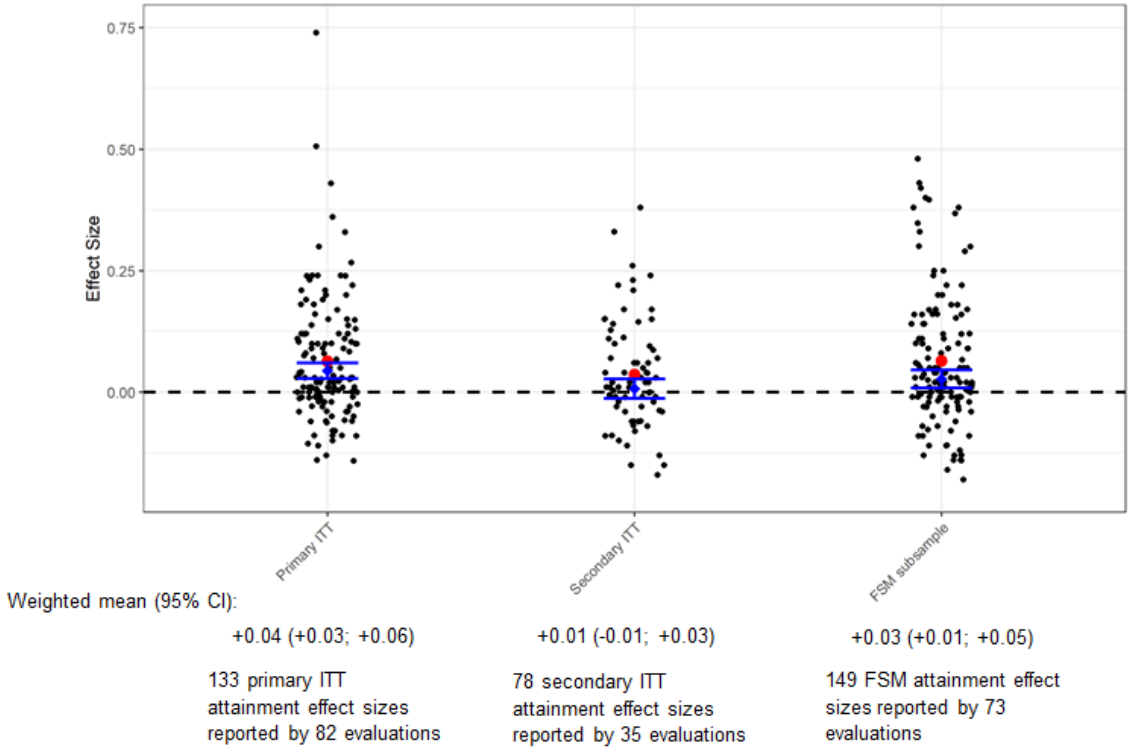
Figure 6: Dot plot: effect sizes reported for FSM attainment outcomes excluding trials with high/low outliers*



***** Excluding five FSM effect sizes reported by Nuffield Early Language & the first IPEEL evaluation (three of these are high/low outliers in Figure 6).

Finally, to summarise, the three groupings of effect sizes relating to outcomes that measured pupil attainment are illustrated in Figure 7.

Figure 7: Summary of primary ITT, secondary ITT and FSM effect sizes for attainment outcomes



Key: Individual effect sizes are shown as black dots. A single evaluation might have dot(s) in one or more of these groupings.
The red dot indicates the value for the unweighted mean.
The blue error bars represent the meta-analysis weighted mean and 95% CI for each grouping.
Note: There is some noteworthy variation in the sample size of attainment–outcome effect sizes across these three groupings, both in terms of the number of trials and reported effect sizes.

Effect sizes for ITT analyses of primary outcomes were reported by all 82 trials included in the review (133 effect sizes). Effect sizes for FSM subsample analyses of primary or secondary attainment outcomes were reported by 73 of the 82 trials included in the review (149 effect sizes). Effect sizes for secondary ITT attainment outcomes were rarer and only reported for 35 of the 82 evaluations in the review (78 effect sizes).

This brings the need for additional caution when comparing the meta-analysis findings across the three effect size groupings. This is particularly the case for secondary ITT attainment effect sizes because the analyses are restricted to less than half of the trials included in the primary ITT meta-analyses (35 trials, 43%). The FSM effect size meta-analyses

include a sizeable majority of trials included in the primary ITT meta-analyses (73 trials, 89%) but include both primary and secondary attainment outcomes.

Therefore, whilst the analyses are presented together, grouped under the five overarching themes, the descriptive statistical patterns and statistical significance of associations should be interpreted separately and any comparison between them done with caution.

Reported effect sizes for ITT analyses of psychological outcome(s)

In addition to attainment outcomes, 88 effect sizes relating to psychological outcomes were also reported by 21 of the 82 evaluations included in the review. These outcomes were even more diverse in nature than the attainment outcomes looked at above, including 'attitudes and beliefs', 'cognition and meta cognition', 'social behaviours' and 'mental health'. Given the wide variation in outcomes and mixture of teacher/pupil measurement, a meta-analysis that merely combines all psychological outcomes is of little value. Further, the psychological nature of these outcomes is distinct from the main focus of the main meta-analyses for the review (primary ITT attainment outcomes). Future research may want to focus on psychological outcomes to identify a coherent classification prior to any meta-analyses. Please see Appendix C for our initial thoughts on this.

Effect sizes for psychological outcomes for ITT (primary or secondary ITT) or FSM subsample analyses are combined for this summary.

Table 9 and Figure 8 summarise statistically and graphically the 88 effect sizes reported for psychological outcomes reported by 21 of the 82 trials in the review.

Table 9: Effect sizes reported for psychological outcomes for ITT or FSM pupil samples

	Number of trials*	Number of outcomes / effect sizes	Unweighted median (IQR)	Unweighted mean (SD)	Min	Max
Psychological outcomes	21	88	+0.04	+0.05 (0.175)	−0.78	+0.70
Year of publication (ES level)						
2014	2	6	+0.02	0.03 (0.174)	−0.20	+0.26
2015	4	19	+0.04	+0.02 (0.226)	−0.78	+0.32
2016	6	29	+0.01	+0.08 (0.213)	−0.17	+0.70
2017	1	4	–	–	–	–
2018	6	27	+0.06	+0.03 (0.095)	−0.15	+0.25
2019	2	3	–	–	–	–

*Excludes quasi-experimental designs.

Figure 8: Dot plot: effect sizes reported for psychological outcomes

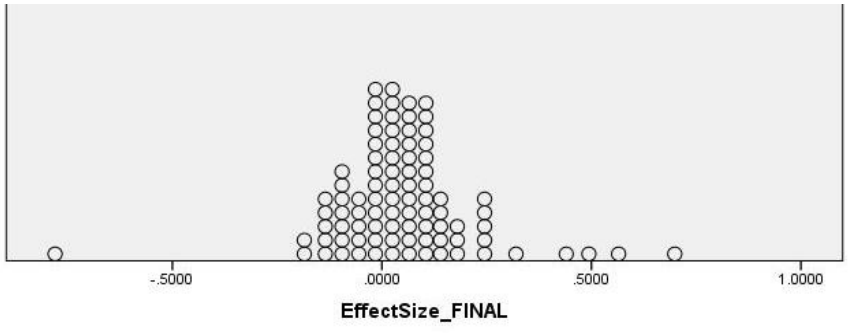


Table 10: Psychological outcomes, meta-analyses weighted mean and 95% CI

	Number of trials	Number of outcomes / effect sizes	Weighted mean (SE)	95% CI	
				Lower	Upper
Psychological outcomes	21	88	+0.05 (0.011)	+0.03	+0.07

Whilst the meta-analysis weighted mean effect size for psychological outcomes is larger than any of the weighted mean effect sizes for attainment outcomes, the disparate nature of these outcomes needs to be kept in mind. The effect sizes ranged between −0.78 SD ('mindset' outcome, FSM subsample) and +0.70 ('classroom concentration', ITT analysis) with an overall (weighted and unweighted) mean of +0.05 SD.

Secondary outcome: cost effectiveness

In consultation with EEF, a cost effectiveness measure was derived that linked the cost per pupil and reported effect size to construct a variable that measured cost effectiveness, defined as the cost per pupil for an effect size of +0.10 SD.

Cost effectiveness was derived by dividing the reported cost (per pupil averaged over three years) by the effect size to obtain the cost per effect size of 1 standard deviation. This is then divided by 10 to obtain the cost per pupil for an effect size of 0.10 standard deviations, as shown in Equation 1.

$$\text{Equation 1: Cost effectiveness} = \text{cost per pupil per 0.10 SD effect size} = \frac{\text{cost per pupil}}{10 \times \text{effect size}}$$

Although Equation 1 appears relatively simple, there is complexity in it because the numerator and denominator are measured at different levels: cost per pupil is a trial-level measure, whereas a single trial may report multiple effect sizes for ITT analyses of primary outcomes. Future reviews may allow for variation in this outcome within trials with multiple outcomes by using a multilevel analytical design[19] but this is beyond the scope of the present review. In this review, the exploratory descriptive analysis of cost effectiveness is at a trial level.

Prior to deriving the trial-level cost effectiveness measure, two criteria were agreed with EEF:

- First, at the effect size level, only effect sizes above +0.05 SD were included.[20]
- Second, at the trial level, only trials where at least half of the reported effect sizes were above +0.05 SD were included.

In applying these two criteria, only trials where at least half of the reported effect sizes for ITT analyses of primary outcomes were above +0.05 SD are included. Thus, these two criteria redefined the cost effectiveness outcome variable to be conditional on evidence of a positive effect (i.e., the cost effectiveness given that a positive impact was reported).

At the effect size level, 57 of the 133 effect sizes were above +0.05 SD. These 57 effect sizes were reported by 43 of the 82 EEF evaluations in the review. However, for three of these evaluations, less than half of reported effect sizes were above +0.05 SD and so these three trials were dropped from the cost effectiveness outcome. This resulted in identifying 40 of the 82 trials that met the criteria for inclusion. These 40 trials reported a total of 63 effect sizes, 54 of which were above +0.05 SD.

Twenty-four of the 40 trials included in the cost effectiveness outcome had a single primary outcome effect size and in these cases Equation 1 was used directly to calculate the (trial level) measure of cost effectiveness. The remaining 16

[19] With only 16 trials with multiple outcomes included in the cost effectiveness outcome, a multilevel design may struggle but could be a future possibility.
[20] There are technical and practical reasons for the need to set a lowest effect size threshold. Technically, as an effect size approaches zero, Equation 1 will approach infinity. Practically speaking, cost effectiveness makes no sense with a negative or zero effect size: spending any money to obtain no change is not cost effective! Setting a minimum effect size threshold of +0.05 SD avoids both issues.

trials had multiple effect sizes. For these trials, Equation 1 was used to calculate an effect-size-level cost effectiveness measure and, from these, a mean cost effectiveness for each of the 16 trials was calculated.

As might be expected, the 63 effect sizes reported by the 40 evaluations included in the cost effectiveness outcome have a different statistical profile to that seen overall in Figure 8 above, and this is shown in Table 11 below. The 63 effect sizes ranged between −0.11 and +0.74 SD with a mean of +0.15 and median of +0.12 SD. The 58 effect sizes reported by the 39 trials that did not report an effect size above +0.05 SD ranged between −0.14 and +0.05 SD[21] with a mean of −0.02 and median of −0.01 SD.

Table 11: Components of cost effectiveness effect size level: effect size and inclusion in the cost effectiveness outcome

		Number of trials	No. of ES	Median effect size	Mean effect size (SD)	Min	Max
All effect sizes in review		82	133	+0.03	+0.06 (0.128)	−0.14	+0.74
Inclusion in cost effectiveness:							
Not included	Zero ES > +0.05 SD	39	58	−0.01	−0.02 (0.045)	−0.14	+0.05
	Minority of ES > +0.05 SD	3	12	+0.04	+0.03 (0.048)	−0.06	+0.08
Included	50%+ of ES > +0.05	40	63	+0.12	+0.15 (0.137)	−0.11	+0.74

Table 12 shows that, on average, the cost per pupil for the 40 trials included in the cost effectiveness outcome was higher compared with the 42 trials not included. For the 40 trials included, the cost per pupil ranged between £1 and £1,750 with a mean of £246 and median of £79. For the 42 trials not included, cost per pupil ranged between £2 and £804 with a mean of £106 and median of £48.

Table 12: Trial level: cost per pupil and inclusion in the cost effectiveness outcome

	No. of trials	Median cost per pupil	Mean cost per pupil (SD)	Min	Max
All trials in review	82	£54	£174 (322.4)	£1	£1,750
Excluded from cost effective outcome	42	£48	£106 (165.8)	£2	£804
Included in cost effectiveness outcome	40	£79	£246 (420.4)	£1	£1,750

The conditionality applied to the cost effectiveness outcome provides useful information as a supplementary variable. This is the probability of a trial being included in the cost effectiveness measure (i.e., trial-level evidence of positive impact). Given that 40 of the 82 trials were included in the cost effectiveness outcome, the overall mean probability was 40/82 = 0.487. Variations around this mean probability provide a second (trial level) perspective on 'positive impact' to supplement the meta-analyses of reported effect sizes.

Figure 9 summarises the construction and distribution of the cost effectiveness outcome variable. First, a scatterplot of (primary outcome) effect size (*y*) vs (trial level) cost per pupil (*x*) is shown, which includes the reference effect size line at '+0.05 SD'. Only trials that reported an effect size above +0.05 for at least half of the primary outcomes are included in the cost effectiveness outcome. The scatterplot shows these trials highlighted in yellow. Three red effect sizes are shown above the +0.05 SD effect size reference line. These relate to three trials where one of the reported effect sizes was above +0.05 SD but the majority were not above this threshold. Below the scatterplot is a statistical summary and dot plot of the cost effectiveness outcome.

The distribution of the cost effectiveness outcome displays notable positive skew. This is illustrated visually by the long upper tail in the dot plot shown in Figure 10 and demonstrated statistically by comparing the mean (£150) with the

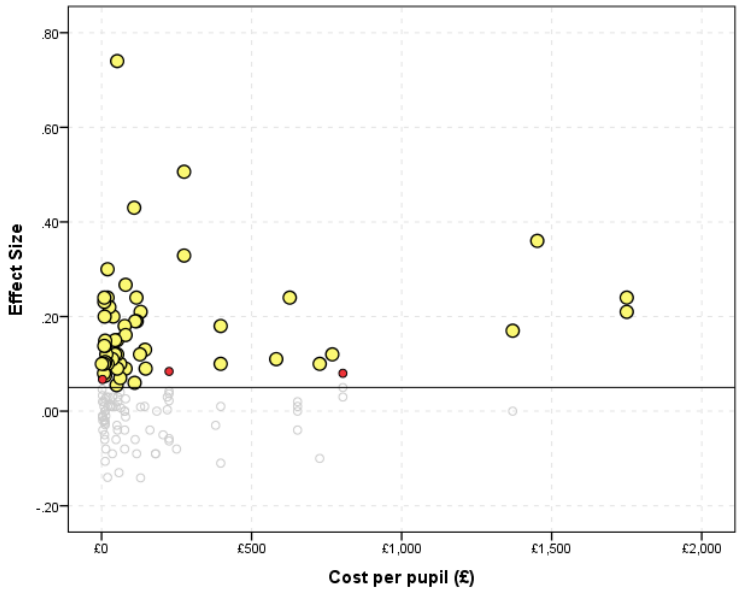
[21] This is an effect size of 0.049 rounded.

median (£54) statistics. Across the 40 trials, the distribution ranges between £1 and £806; eight trials (20% included in the outcome) had cost effectiveness above £200.

A majority of trials included in the cost effectiveness measure reported in 2014 and 2015 (23 trials; 58% of those included in the outcome) and none of the three trials reporting in early 2019 are included. This higher concentration of earlier trials is seen in Table 13 by comparing the probability of inclusion over time. The probability that a trial is included in the cost effectiveness outcome is higher in 2014 and 2015 (p = 0.63 or higher) than between 2016 and 2019 (p = 0.41 or lower).

The mean cost per pupil for an effect size of +0.05 SD is observed to fall consistently over time between 2014 and 2018. However, change over time in median cost effectiveness is less consistent, so some caution is needed here. This said, it is apparent that the 10 trials which reported a positive impact in 2014 were for less cost effective interventions (mean = £238; median = £100), compared with the 30 trials which reported a positive impact after 2014 (mean = £166 or less; median = £67 or less).

Figure 9: Deriving the cost effectiveness outcome: scatterplot of effect size vs cost per pupil

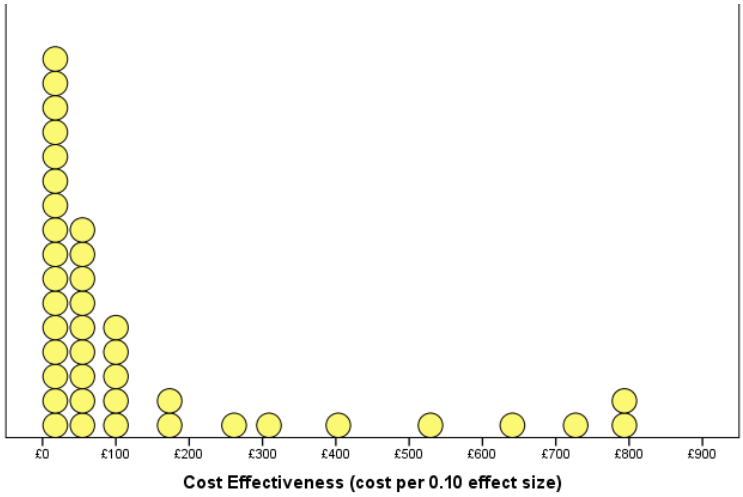


Key: Yellow – included in cost effectiveness outcome (more than 50% of ES ≥ 0.05).
Grey – ES of +0.05 or lower, not included in cost effectiveness outcome.
Red – ES > +0.05 but the majority of ES in these trials were not (so not included in cost effectiveness outcome).

Table 13: Statistical summary: cost effectiveness outcomes

	No. of trials	*p*(CE)	Median CE (£)	Mean (SD) (£)	Min	Max
Cost effectiveness						
Cost effectiveness (£ per 0.10 SD)	40	0.49	£54	£150 (229.1)	£1	£806
Cost effectiveness and publication year						
2014	10	0.63	£100	£238 (£303)	£7	£806
2015	13	0.65	£48	£166 (£224)	£4	£641
2016	6	0.38	£26	£143 (£287)	£5	£727
2017	4	0.40	£67	£83 (£59)	£34	£163
2018	7	0.41	£33	£37 (£37)	£1	£107
2019	0	0.00	–	–	–	–

Figure 10: Dot plot: cost effectiveness distribution (trial level)



Secondary outcome: % pupil-level attrition

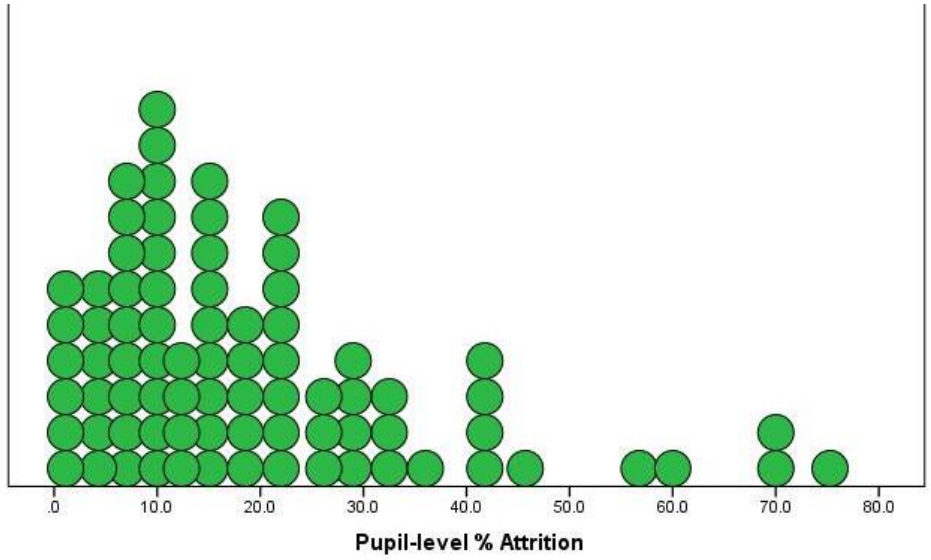
The pupil-level attrition rate was obtained for 79 of the 82 trials in the review. For the three trials where a rate was not obtained, a categorised measure was obtained from the appendix of the evaluation report (and it is this version that is used when attrition was used as a categorical explanatory variable). Table 14 and Figure 11 present the pupil attrition outcome measure and compare the distribution with the previous UCL/IoE review.

Table 14: Pupil-level attrition (%) statistical summary

	No. of trials*	Missing attrition detail	Median	Mean (SD)	Min	Max
UCL/IoE review	47	5	16.0%	19.5% (16.68)	0.0%	73.0%
SHU review (trial level)	82	3	15.2%	19.4% (16.54)	0.0%	75.3%
Year of publication (ES level)						
2014	16	0	21.3%	26.8% (22.82)	0.0	75.3
2015	20	0	16.1%	21.2% (17.37)	2.0	71.0
2016	16	0	15.4%	16.8% (11.78)	0.0	41.0
2017	10	2	9.3%	15.8% (17.40)	2.2	56.7
2018	17	1	9.5%	13.4% (10.78)	2.0	41.3
2019	3	0	26.6%	24.3% (5.29)	18.2	28.0

***Excludes quasi-experimental designs.

Figure 11: Dot plot of pupil-level attrition distribution (trial level)



The average rate of pupil attrition is slightly lower for the 79 trials with data in the SHU review compared with that found with the 42 trials in the UCL/IoE data. If the three 2019 trials are ignored, there seems to be a trend of declining rates of attrition over time between 2014 (mean = 26.8%; median = 21.3%) and 2018 (mean = 13.4%; median = 9.5%).

Presenting the explanatory variables

Overview

As set out in the review framework section, a theoretical framework that grouped explanatory variables under five overarching themes was used to structure the meta-analyses of primary outcome effect sizes:

1. The intervention
2. Evidence & theory
3. Context
4. Implementation & fidelity
5. Evaluation design.

The initial selection of explanatory variables under each theme focused initially on hypothesised association with reported effect sizes for ITT analyses of primary outcomes. The explanatory variables under each of the five themes were then assessed for inclusion or not within meta-analyses of secondary ITT and FSM effect sizes and within the descriptive analyses of cost effectiveness and attrition.

We use the term 'explanatory variable' in a purely descriptive way. We do not hypothesise that each 'explanatory variable' and effect sizes are causally linked. This is an exploratory and descriptive review and the selection of 'explanatory variables' is purposely broad. Future reviews may draw on these preliminary bivariate descriptive and exploratory meta-analyses in the formulation of more narrowly defined multivariate deductive meta-analyses. To reiterate, the analyses presented here are exploratory and descriptive and this is how they should be interpreted.

This section is in two parts. The first part introduces the explanatory variables under each theme and indicates whether they were included in the quantitative analyses of outcomes for the review. The second part describes the univariate distributions of all explanatory variables.

Nearly all of the explanatory variables were measured at an evaluation/trial level. The only exception is found under the evaluation design theme for variables that examine the type of primary outcome used. These variables link to specific detail on the type of test behind a reported effect size.

Introducing the explanatory variables

The intervention

Explanatory variables for the intervention theme were organised under seven subthemes:

1. Focus of intervention
2. Intensity (minutes per week)
3. Direct implementers
4. Perceived quality of supporting resources
5. Cost of intervention
6. EEF intervention theme areas
7. EEF rating as promising project

Table 15 lists the explanatory variables included in the intervention theme and their inclusion in the analyses of the three outcomes.

Table 15: Explanatory variables included in the intervention theme

Subtheme	Source*	Variables	Effect Size*	Cost-Effectiveness	Attrition
Focus	Data	School phase	✓	✓	✓
	Data	School Key Stage	✓	✓	✓
	Data	Curriculum focus of intervention	✓	✓	✓
Intensity	Data	Minutes per week	✓	✓	✓
Who implements with direct target?	Review	Direct implementer (Teacher / TA / external)	✓	✓	–
Perceived quality of supporting resources	Review	High / varied / low	✓	✓	–
Cost	Data	Total cost	✓	✓	–
	Data	Cost per pupil (over three years)	✓	–	–
EEF intervention themes	Data	Language and literacy	✓	✓	✓
	Data	Maths and numeracy	✓	✓	✓
	Data	Staff deployment and development	✓	✓	✓
	Data	Organising your school	✓	✓	✓
	Data	Developing effective learners	✓	✓	✓
	Data	Feedback and monitoring pupil progress	✓	✓	✓
	Data	Behaviour	✓	✓	✓
	Data	Character and essential skills	✓	✓	✓
	Data	Parental engagement	✓	✓	✓
	Data	Science	✓	✓	✓
	Data	Enrichment	✓	✓	✓
	Data	Early years	✓	✓	✓
	Data	Special educational needs	✓	✓	✓
EEF promising project	Data	Promising project	✓	✓	✓

Key: Data = amalgamated EPPI/IoE/Lortie-Forgues & Inglis with manual checks and updates to cover the 82 evaluations in the review period. **Review:** new data collected as part of review of 82 EEF evaluation reports.

The intervention theme resulted in 22 explanatory variables for analyses of primary and secondary outcomes. Of these, 20 were drawn from data amalgamated from the EPPI database of trials, the previous IoE review and Lortie-Forgues & Inglis. Most of these relate to 13 EEF intervention themes used by EEF to classify evaluations.[22] The remaining three explanatory variables were from the review of EEF evaluation reports.

Theory & evidence

Explanatory variables for the theory & evidence theme were organised under three subthemes:

1. Strength of prior evidence of impact
2. Level of theoretical detail
3. Causal processes and mechanisms.

Table 16 lists the explanatory variables included in the theory & evidence theme and their inclusion in the analyses of the three quantitative outcomes in the review.

[22] The EEF classifies evaluations into 14 EEF intervention themes but none of the 82 trials in the review were classed in the post-16/FE theme, leaving 13 EEF themes (see https://educationendowmentfoundation.org.uk/school-themes/).

Table 16: Explanatory **variables included in the theory & evidence** theme

Subtheme	Source*	Variables	ES*	CE	Att
Evidence	Review	Strength of prior evidence of impact	✓	✓	–
Theory	Review	Level of theoretical detail	✓	✓	–
Causal process	Review	Focus of change	✓	✓	✓

Key: Data = amalgamated EPPI/IoE/ Lortie-Forgues & Inglis with manual checks and updates to cover the 82 evaluations in the review period. Review**:** new data collected as part of review of 82 EEF evaluation reports.

The theory & evidence theme resulted in three explanatory variables all of which were drawn from the review of evaluation reports. All three variables were used in the meta-analyses of reported effect sizes and cost effectiveness, but the analyses of pupil attrition are limited to just focus of change.

Context

Explanatory variables for the context of the evaluations were organised under three subthemes:

1 External context

2 Characteristics of participating organisations (barriers and enablers)

3 Characteristics of participating individuals (barriers and enablers).

Table 17 lists the explanatory variables included in the context theme and their inclusion in the analyses of the three quantitative outcomes in the review.

Table 17: Explanatory variables included in the context theme

Subtheme	Source*	Variables	ES*	CE	Att
External context	Review	Geography	✓	✓	✓
	Review	Perceptions on Ofsted	✓	–	–
Characteristics of participating organisations					
	Review	Specialist facilities and space	✓	–	–
	Review	Staff time and availability	✓	–	–
	Review	Workforce capacity	✓	–	–
	Review	Alignment of intervention and current practice	✓	–	–
	Review	Staff teamwork	✓	–	–
Characteristics of participating individuals					
	Review	Pupil behaviour	✓	–	–
	Review	Staff expectations and motivations	✓	–	–

Key: Data = amalgamated EPPI/IoE/Lortie-Forgues & Inglis with manual checks and updates to cover the 82 evaluations in the review period. Review**:** new data collected as part of review of 82 EEF evaluation reports.

The context theme resulted in nine explanatory variables, all of which were drawn from the review of evaluation reports. All the data included in the context theme is perceptual (i.e., it relies on the reported perceptions of participants and other stakeholders).

Implementation & fidelity

Explanatory variables for the implementation & fidelity theme were organised under five subthemes:

1 Developer characteristics

2 Focus, planning, time and SLT support

3 Professional development

4 Support and monitoring

5 Fidelity.

Table 18 lists the explanatory variables included in the implementation *&* fidelity theme and their inclusion in the analyses of the three outcomes.

Table 18: Explanatory variables included in the implementation & fidelity theme

Subtheme	Source*	Variables	ES*	CE	Att
Developer characteristics	Data	Charity / university / private company / school or academy or MAT / Council or LA / Mixed	✓	✓	✓
Planning, time and SLT support	Review	Clarity of implementation plan	✓	✓	✓
	Review	Lead-in time for implementation	✓	–	✓
	Review	SLT support	✓	–	✓
Professional development (CPD)	Review	Whether implementation uses CPD	✓	✓	✓
	Review	Whether subject-specific or generic	✓	✓	✓
	Review	Sequencing of CPD	✓	–	✓
	Review	Who delivers CPD?	✓	–	–
	Review	Types of CPD	✓	✓	–
Implementation support and monitoring	Review	Whether developer provided support other than CPD	✓	–	✓
	Review	Monitoring of implementation	✓	–	✓
Fidelity	Review	Intended fidelity	✓	✓	✓
	Review	CPD fidelity	✓	✓	✓
	Review	Implementation fidelity	✓	✓	✓

Key: Data = amalgamated EPPI/IoE/Lortie-Forgues & Inglis with manual checks and updates to cover the 82 evaluations in the review period. Review**:** new data collected as part of review of 82 EEF evaluation reports.

The implementation & fidelity theme resulted in 15 explanatory variables for analyses of primary and secondary outcomes. Of these, 14 were drawn from the review of EEF evaluations and one was drawn from data amalgamated from the EPPI database of trials, the previous IoE review and Lortie-Forgues & Inglis.

All 15 explanatory variables were used in the meta-analyses of reported effect sizes but the analysis of cost effectiveness was limited to eight explanatory variables and the analysis of pupil attrition was limited to 12 explanatory variables.

Evaluation design

Explanatory variables for the evaluation design theme were organised under five subthemes:

1 Trial description

2 Length and size of trial

3 Statistical sensitivity, attrition and trial quality

4 Evaluation burden

5 Primary outcome.

Table 19 lists the explanatory variables included in the evaluation design theme and their inclusion in the analyses of the three outcomes.

Table 19: Explanatory variables included in the evaluation design theme

Subtheme	Source*	Variables	ES*	CE	Att
Trial description	Data	Type of trial (RCT/CRT)	✓	✓	✓
	Data	Level of randomisation	✓	✓	✓
	EEF	Efficacy/effectiveness	✓	✓	✓
	Data	Type of evaluator	✓	–	✓
Length and size of trial	Data	Intervention length (weeks)	✓	✓	✓
	Data	Number of schools	✓	✓	✓
	Data	Number of pupils	✓	✓	✓
Statistical sensitivity, attrition and trial quality	Data	Statistical sensitivity (MDES estimate)	✓	–	–
	Data	Pupil-level % attrition	✓	–	–
	Data	Trial quality (EEF padlocks)	✓	✓	✓
Evaluation burden	Review	Testing burden	✓	–	✓
	Review	IPE data collection burden	✓	–	✓
Primary outcome	Data	Type (commercial/statutory/other)	✓	✓	✓
	Data	Outcome curriculum area	✓	✓	–
	Data	Number of primary outcomes	✓	–	–
	Review	Alignment of intervention and primary outcome	✓	–	–

Key: Data = amalgamated EPPI/IoE/Lortie-Forgues & Inglis with manual checks and updates to cover the 82 evaluations in the review period. Review**:** new data collected as part of review of 82 EEF evaluation reports. EEF provided data for classifying efficacy/effectiveness trials.

The evaluation design theme resulted in 16 explanatory variables. Of these, 12 were drawn from data amalgamated from the EPPI database of trials, the previous IoE review and Lortie-Forgues & Inglis, three were from the review of EEF evaluation reports and one drew on data provided by EEF.

All 16 variables were used in the meta-analyses of reported effect sizes but the analysis of cost effectiveness was limited to nine explanatory variables and the analysis of pupil attrition was limited to 11 explanatory variables.

Describing the explanatory variables

Data tables that support the presentation and interpretation of findings below can be found in the *Technical Annex*. Variable and code descriptors are listed in Appendix A.

The intervention

Focus of intervention

The majority of interventions took place in primary schools (51 evaluations, 62%), particularly at KS2 (33 evaluations, 40%). Twenty-five interventions (30%) took place in secondary schools, most commonly at KS3 (20 evaluations, 24%).

The curriculum focus variable was obtained from the EEF webpages for the 82 evaluations in the review. The most common curriculum focus for an intervention was English or literacy (36 evaluations, 44%), followed by interventions with a cross-curriculum focus (29 evaluations, 35%) and then maths/numeracy (14 evaluations, 17%).

Intensity of intervention

The intensity of an intervention was measured using the number of minutes per week of intervention activity that was reported to take place in the classroom. Intensity data was obtained for 51 of the 82 trials in the review (62%). The overall mean intensity for these 51 interventions was just over 90 minutes per week (median of just over 70 minutes). Intensity was banded into four groups, ranging from interventions lasting 30 minutes or less per week (12 evaluations, 15% of all evaluations, 24% of evaluations with intensity detail) to those lasting over 120 minutes per week (11 evaluations, 13% of all evaluations, 22% of evaluations with intensity detail).

The previous IoE review obtained intensity detail for 30 of the 47 trials in their review (64%). These had a mean intensity of 97 minutes per week (median = 85 minutes).

There is some limited evidence of a small decrease in the intensity of interventions over the six years of the review. On average, the evaluations that reported in 2014 (median = 100 minutes per week) and 2015 (median = 85 minutes) were for more intense interventions than evaluations published more recently (median 60 minutes per week or less).

The relationship between the intensity of an intervention (minutes per week) and length of intervention (in weeks) was examined. This highlighted that the 31 evaluations with no intensity detail were longer on average (mean length = 55 weeks; median = 45 weeks) compared with the 51 trials that did have detail of the intensity of the intervention (mean length = 28 weeks; median = 23 weeks). Among the 51 trials with intensity detail, differences in mean length of intervention were less striking, ranging from a mean length of 23 weeks for interventions with an intensity of 61–120 minutes per week up to a mean length of 34 weeks for interventions with an intensity of 30 minutes or less per week.

Direct implementers

Teacher-led interventions were the most common (37 evaluations, 45%) followed by externally-led interventions (i.e., led by the delivery partner or consultants they employed) (18 evaluations, 22%) and TA-led interventions (12 evaluations, 15%). Trials led by parents, resources or other school staff are few (<3 trials each). The distribution of EEF trials direct by implementer has not previously been reported.

Perceived quality of intervention support resources

The review found that a sizeable minority of evaluations provided no detail on how supporting resources were perceived by implementers (30 evaluations, 37%) but of the 52 that did, 20 evaluations (24% of all evaluations, 38% of evaluations that reported perceptions) found that perceived quality was high; five evaluations (6% overall, 9% of evaluations that reported perceptions) found the quality to be low; and 27 evaluations (33% overall, 52% of evaluations that reported perceptions) reported variations in perceptions on support resources. Although this distribution has not previously been reported for EEF trials, the qualitative review by Anders et al. (2017) conveyed a general dissatisfaction with programme resources. The findings in this review are slightly more positive in that stakeholders in some trials (38% of evaluations that reported on this variable) perceived resources to be high quality. The EEF may wish to consider whether it should play a more active role in setting standards for the quality of resources used in EEF trials and whether it should instruct

evaluators to systematically report data on the perceived quality of resources to enable a more accurate understanding of the extent and nature of the issue.

Cost

The mean total cost of delivery of interventions across the 82 trials in the review was just under £500k (median ~£470k). Total cost was categorised into six bands, ranging from interventions under £100k (4 evaluations, 5%) to those above £1 million (6 evaluations, 7%).

The total cost of delivery is seen to fluctuate over time from a mean and median under £370k in 2014 and 2015 increasing to a peak of over £700k (median just over £600k) in 2017, before falling to just under £500k (median = £450k) in 2018. The three evaluations that reported in January 2019 were for interventions with a mean total cost of just over £400k (median just over £500k).

The cost per pupil scale variable was used in conjunction with effect size to derive the cost effectiveness outcome variable for this review (see *Outcome variables* above). The mean cost per pupil was £174 (median = £54). Here, as an explanatory variable under the intervention theme, it has been categorised into seven bands, ranging from interventions costing less than £10 per pupil (12 trials, 15%) to interventions costing over £1,000 per pupil (3 trials, 4%).

Data from the previous IoE review obtained cost per pupil detail for 46 of the 47 trials in their review. These had a mean cost per pupil of £260 (median = £111) for evaluations published up to 2016. This suggests that cost per pupil has decreased over time. In terms of publication year, a pattern of falling cost per pupil is observed. The highest average cost was found for evaluations published in 2014 (mean = £312 per pupil; median = £115) but this is seen to fall consistently between 2015 (mean = £225; median = £85) and 2018 (mean = £46; median = £36). The three evaluations that reported in 2019 had a mean cost per pupil of £34 (median = £39).

EEF intervention themes

The 82 evaluations in the review were linked to one or more of 13 EEF intervention themes.[23] The 'Language and literacy' theme was the most common (38 evaluations, 46%) followed by 'Staff deployment and development' (36 evaluations, 44%), 'Organising your school' (18 evaluations, 22%), 'Developing effective learners' (17 evaluations, 21%) and 'Mathematics' (16 evaluations, 20%). Evaluations included in the remaining eight themes were rarer (10 evaluations or fewer).

Each of 19 evaluations (23%) were placed in just one of the 13 intervention themes, but the remaining 63 evaluations were placed in two themes (41 evaluations, 50%) or three or more themes (22 evaluations, 27%).

EEF promising interventions

The review includes 17 interventions classed as promising[24] by EEF (21% of all trials in the review) and which reported a total of 30 ITT primary outcome effect sizes.

Theory & evidence

Empirical evidence and theoretical detail

Evaluations with strong empirical evidence were equally as common as trials with highly detailed theoretical discussion (17 trials, 21%). However, only five evaluations (6%) had both strong evidence and highly detailed theory whilst 29 evaluations (35%) had either strong evidence or highly detailed theory, but not both.

[23] There are 14 EEF intervention themes but the review did not include any interventions in the relatively new post-16/FE theme (see https://educationendowmentfoundation.org.uk/school-themes/).
[24] Classed as promising by EEF during the review period.

Evaluations with minimal or no theoretical detail were more common (37 evaluations, 45%) than evaluations with minimal or no empirical evidence (9 evaluations, 11%). Only seven evaluations (9%) had both minimal or no theoretical detail and minimal or no empirical evidence.

All of the evaluations with minimal or no empirical evidence were published before 2018, whilst the proportion of evaluations with strong empirical evidence is seen to increase over time from only one of the 16 evaluations published in 2014 (6%) to five of the 17 evaluations published in 2018 (29%).

Theoretical detail is also seen to increase over time. Between 2014 and 2017, over half of the evaluations had minimal or no theoretical detail (34 evaluations, 55% published between 2014 and 2017) but this was rarer in more recent publications (3 evaluations, 15% published between 2018 and 2019). The number of evaluations with highly detailed discussion of theory was observed to increase in 2018 (6 evaluations, 35% published in 2018). All three of the evaluations published in January 2019 discuss underlying theory in detail.

It is important to note that this variable does not provide an accurate measure of empirical evidence and the strength of theory supporting the intervention, because it is dependent on the evaluator providing a good quality account of the underpinning evidence and theory. The increase in empirical and theoretical evidence found may reflect that the evaluators have been striving to provide more detail over time, rather than there being stronger empirical and theoretical foundations for the trialled interventions. To undertake a more valid assessment of the impact of prior evidence and theory on outcomes would require clearer guidelines to evaluators to enable consistent reporting and capture of prior evidence.

Causal processes and mechanisms

A large majority of interventions in the review had a main focus on learner change (69 evaluations, 84%) with nine evaluations (11%) focusing mainly on change in wider pupil outcomes, and four evaluations focusing mainly on teacher change (3%), rather than a direct focus on learners.

Context

External context

The geographical contexts of the trials were commonly national (25 trials, 30%) or within two or three geographical areas (22 trials, 27%).

Preparing for and being subject to Ofsted inspection visits, or staff perceptions that Ofsted required particular ways of planning, delivering or assessing learning that conflicted with the intervention methods, was mentioned as a barrier to implementation in 16 evaluation reports (20%).

Organisational characteristics

The most commonly perceived organisational barrier was staff time and availability (66% of evaluations), followed by specialist facilities and space (43%) and workforce capacity (38%). A similar proportion of evaluations identified staff teamwork (27% of evaluations) and the alignment of an intervention and existing practice (23%) as enablers to implementation.

Individual characteristics

Pupil behaviour was mentioned as a barrier to implementation in 32% of evaluations. Perceptions on senior leadership team (SLT) buy-in and staff expectations/motivations were more mixed. In 29 evaluations, the interviewees or survey respondents perceived that SLT support for or championing of the intervention was an enabler in their school. Conversely, in 19 evaluations, a lack of SLT support was perceived to be a barrier. There is an intersection of 11 trials where SLT buy-in was mentioned as both an enabler and a barrier. Positive staff expectations and/or motivations were perceived to be enabling factors in 33 evaluations, while less positive expectations and/or weaker motivation were perceived as barriers in 27 evaluations (there is an intersection of 15 trials where staff expectations and motivations were mentioned as both an enabler and barrier).

Implementation & fidelity

Developers

Charities were the most common developers of interventions across the 82 evaluations in the review (32 evaluations, 39%), followed by universities (19 evaluations, 23%), private companies and individual schools / academy trusts (both with 9 evaluations, 11%) and local authorities (8 evaluations, 10%).

Implementation planning and time

The review identified 33 evaluations (40%) where the implementation plan was clearly understood (i.e., the data indicated that all or nearly all stakeholders understood how the intervention and (when appropriate) any associated training was to be implemented). However, 23 evaluations (28%) were identified that reported varied perceptions on the clarity of the plan, and in 26 evaluations (32%) it was either not clear whether the plan was understood or there was no mention of understanding of the plan. A robust analysis of the effect of the extent to which the implementation plan was understood would require more systematic reporting in IPEs.

The review found that the 'lead-in time' for preparing for implementation was mentioned in a small majority of evaluation reports (43 evaluations, 52%). Aligning with previous implementation literature, when mentioned, having insufficient time was the more commonly reported issue (24 evaluations, 29% of all evaluations, 56% of evaluations where 'lead-in time' was mentioned). It was rare for evaluations to report that there was sufficient time (5 evaluations, 6% of all evaluations,12% of evaluations where 'lead-in time' was mentioned).

Use of continuing professional development (CPD)

The vast majority of programmes in the review provided one or more forms of CPD to support implementation of the intervention (77 trials, 94%).[25]

The review found that it was most common for the CPD to take place both before **and** during the delivery of the intervention to the direct targets (47 evaluations, 57%). Eighteen evaluations (22%) reported that CPD took place only before the delivery of the intervention and 10 evaluations (12%) reported CPD taking place only during the delivery of the intervention.[26]

A majority of interventions were predominantly subject-specific or curriculum-specific (49 evaluations, 60%) but a notable proportion were more generic (22 evaluations, 27%). Seven interventions (9%) had a mixed subject and generic focus.

The CPD was commonly delivered directly by the developers (53 evaluations, 65%).

Details on four types of CPD were obtained: the most common form was face-to-face CPD (74 evaluations, 90%), followed by train-the-trainer CPD (16 evaluations, 20%), coaching or mentoring (13 evaluations, 16%) and online CPD (11 evaluations, 13%). Please note that these types of CPD are not mutually exclusive and the implementation of one intervention can be classed across multiple categories. For example, 10 of the 11 evaluations that reported online CPD also reported face-to-face CPD.

Implementation support and monitoring

The review found that a majority of developers provided support other than group training sessions before and/or during the intervention (60 evaluations, 73%). This support took various forms, spanning more intensive support, such as classroom visits and feeding back to direct implementers, through to lighter touch approaches, such as email support. In some instances, support was provided to the direct implementer; in others to the school lead.

Just over half of the evaluations reported some form of monitoring of implementation (42 evaluations, 51%); this varied in nature and intensity across the trials. Most commonly this monitoring was done by delivery partners and there was

[25] Note that this includes interventions that involved direct delivery by external organisations; in these cases, the CPD was for staff in these external delivery organisations
[26] See the *Limitations* section for issues relating to what constitutes 'before' and 'during' intervention, and the *Recommendations* section for thoughts on improving the clarity of reporting key time points in an evaluation.

some limited evidence of monitoring by school staff. The lack of evidence on in-school monitoring is likely to be a consequence of this variable being outside the scope of most evaluations.

The review found that SLT support was not mentioned in the majority of reports (44 evaluations, 54%). Strong support was reported for the implementation of 11 interventions (13% of all evaluations, 29% of evaluations that mentioned SLT support). Having some support was more commonly reported (22 evaluations, 27% of all trials, 58% of evaluations that mentioned SLT support). Evaluations that reported limited or minimal SLT support were rare (5 evaluations, 6% of all trials, 13% of evaluations that mentioned SLT support). This is a surprising distribution of responses because across the evaluations of educational programmes more generally the lack of SLT support is frequently mentioned as a barrier to effective implementation, as was the case in the qualitative review by Anders et al. (2017). The limited reporting of a lack of SLT support impeding successful implementation in EEF trials may have occurred simply because this was not a focus for investigation within the IPE. A further possible explanation is that in EEF trials SLT receive full details of the requirements of both the intervention and the trial, and are formally asked to commit to these obligations (usually through a memorandum of understanding) before they are classified as participants. In these circumstances it is likely that SLT who are less committed to implementing the intervention would choose not to participate. In other types of evaluation, where there may be less clarity about what will be required from the school at an early stage, or an individual teacher rather than the SLT agrees to try out an intervention in their classroom, SLT commitment may be weaker.

Fidelity

Fidelity of implementation was examined in terms of fidelity to CPD, intended fidelity (by the direct implementer), and the actual fidelity of implementation (by the direct implementer).

The review found that 68 evaluations (83%) mentioned intended fidelity (by the direct implementer). Whilst the majority did indicate that the intervention was intended to be adopted faithfully (37 evaluations, 45% of all evaluations, 54% that mentioned intended fidelity), a notable proportion reported that the intention was adaptation to context (31 evaluations, 38% of all evaluations, 46% that mentioned intended fidelity).

The review found that sufficient data and/or interpretation were available in 44 evaluations (54%) for the review team to be confident in making an assessment of fidelity relating to CPD. Whilst 12 evaluations reported high CPD fidelity (15% of all evaluations, 27% that mentioned CPD fidelity), the majority reported moderate/varied CPD fidelity (26 evaluations, 32% of all evaluations, 59% that mentioned CPD fidelity). Six evaluations did report limited CPD fidelity (7% of all evaluations, 14% that mentioned CPD fidelity). The new IPE guidance (2019) should contribute to ensuring that a more complete dataset is available in the future.

The review found that 73 evaluations (89%) provided sufficient data and/or interpretation for the review team to be confident in making an assessment of intervention implementation fidelity. Whilst 13 evaluations reported high implementation fidelity (16% of all evaluations, 18% that mentioned implementation fidelity), the majority reported moderate/varied fidelity of implementation (46 evaluations, 56% of all evaluations, 63% that mentioned implementation fidelity). Fourteen evaluations did report limited implementation fidelity (17% of all evaluations, 19% that mentioned implementation fidelity). It is important to note here that the assessment is based on the comparison of intended and actual fidelity, so where the intention was faithful adoption, it would only be categorised as high implementation fidelity if there was no or limited variation to the intervention set out by the developers. Where the intention was to adapt the intervention, high implementation fidelity would include appropriate adaptation to context.

Evaluation design

Trial design description

The majority of evaluations in the review had a clustered RCT trial design with school-level randomisation (48 trials, 59%). Using the classification provided by EEF,[27] half of the evaluations were classed as efficacy trials and half were

[27] There is some variation on defining efficacy and effectiveness trials (e.g., at the top of the EEF webpage it is stated that ‘This page covers the first (efficacy) trial of Grammar for Writing’ whilst lower down this page in the 'Evaluation Info' Table the trial is listed as an 'effectiveness' trial. We therefore have used the classification provided by EEF. It should be noted that the EEF

effectiveness trials. A majority of evaluations were undertaken by a university (52 trials, 63%).

Clustered RCT (CRT) designs are observed to become more common over time. For evaluations published between 2014 and 2016, just under half of evaluations used a classic RCT design (24 evaluations, 46% that published between 2014 and 2016). Evaluations published after 2016 were much more likely to use a CRT design (27 evaluations, 90% that published between 2017 and 2019).

Intervention length and size

The length of intervention was obtained by bringing together differing units of time (weeks, months, terms, years) into a standardised 'weeks' scale, as described in Appendix A. Across the 82 trials, the length of intervention ranged between four and 97 weeks, with a mean of 38 weeks (SD = 30.0). These were categorised into four bands ranging from trials lasting 15 weeks or less (23 trials, 28%) to those lasting over 45 weeks (17 trials, 21%).

Comparing the IoE review, on average the first 47 EEF trials in that review were shorter (mean and median = 24 weeks) than the 82 trials included in the present review. This suggests that, on average, trials may have increased in length over time. Looking at this closely, the length of intervention was observed to fluctuate over time. The mean length was 16 weeks (median = 13.5 weeks) for the first 16 trials published in 2014 up to a peak of 77 weeks (median = 97 weeks) in 2017 before falling to 48.5 weeks (median = 45 weeks) in 2018. The three trials published in January 2019 had a mean and median length of 45 weeks.

Across the 82 evaluations, the number of participating schools ranged between three and 205 with a mean of 64 schools. These were categorised into six bands ranging from trials with 20 schools or fewer (15 trials, 18%) to trials involving over 100 schools (17 trials, 21%).

The number of individual participants (usually pupils) ranged between 36 and 25,000, with a mean of about 3,700 individuals. This distribution was categorised into five bands ranging from trials with 500 individuals or fewer (19 trials, 23%) to trials involving over 5,000 individuals (20 trials, 24%).

The length and size of trial is clearly associated with the type of trial design. On average, the 27 evaluations that used an RCT design were shorter (mean = 18 weeks) and involved fewer schools (mean = 25) and individuals (mean = 618). The 55 evaluations that used a clustered RCT (CRT) design were longer (mean = 49 weeks) and involved a greater number of schools (mean = 83) and individuals (mean = 5,200).

The overlap between trial size and trial design may account for the observed increasing size of trials over the six years of the review. The mean numbers of participating schools increased from 38 in 2014 up to 102 in 2018 (153 for the three 2019 trials). The mean number of participating individuals increased from just under 1,500 in 2014 to nearly 6,000 in 2018 (just under 3,500 for the three 2019 trials).

Statistical sensitivity, attrition and trial quality

The statistical sensitivity of an RCT design is estimated using the reported MDES, assuming a statistical significance of 0.05 and statistical power of 0.80. MDES estimates were obtained for 78 of the 82 evaluations in the review. MDES estimates ranged from 0.07–0.45 SD with a mean of 0.22 SD (SD = 0.081). The MDES estimates were categorised into four bands ranging from 11 trials (13%) with an MDES below 0.15 SD, to six trials (7%) with an MDES of 0.35 SD or higher. Data from the IoE review show a slightly higher mean MDES estimate (0.24 SD, 44 evaluations) for trials published between 2014 and 2016.

MDES estimates were relatively stable over time for the six years of the review. The highest mean MDES was observed in 2014 (0.25 SD) but between 2015 and 2019, the estimates ranged between 0.20 and 0.23.

Attrition rates were obtained for 79 of the 82 trials in the review. Pupil-level attrition rates ranged between 0% and 75% with a mean of 19% (median = 16%). The scale version of pupil-level attrition was used as a secondary outcome in the review (see above). When used as an explanatory variable, it was categorised into five bands, ranging from three trials

classification contrasts with the one used in the previous IoE review. Restricting the sample to RCT/CRT designs, the IoE data show 26 efficacy and 21 effectiveness trials. The EEF classification lists four of the IoE efficacy trials as effectiveness trials and three of the IoE effectiveness trials as efficacy trials.

(4%) reporting zero attrition to 17 trials (21%) reporting an attrition rate of 35% or higher. The categorised version includes all 82 trials in the review, because categorical data were available for the three trials where a specific percentage rate was not available. Data from the IoE review show a similar average pupil-level attrition rate for trials published between 2014 and 2016 (mean = 20%; median = 16%, 42 evaluations).

A fall in attrition rates was observed between 2014 (mean = 27%; median = 21%) and 2018 (mean = 13%; median = 10%) with higher attrition reported for the three trials that published in 2019 (mean = 24%; median = 27%).

The EEF padlock rating was used to measure the security of trial findings. Padlock ratings ranged between 0 and 5 with a mean of 3.1 padlocks (SD = 1.25). Nine of the 82 trials in the review (11%) that reported 11 of the 133 primary outcome effect sizes (8%) were awarded EEFs highest 5 padlocks. Data from the IoE review show a slightly lower average padlock rating for trials published between 2014 and 2016 (mean = 2.8, 47 trials).

The mean padlock rating was observed to increase over time between 2014 (mean = 2.3) and 2018 (mean = 3.7) but was lower for the three trials published in 2019 (mean = 3.0).

Evaluation burden

Evaluations were found to commonly result in quite a high burden on schools in terms of testing and IPE data collection.

Only nine evaluations (11%) undertook no testing at baseline, interim and/or outcome stages and drew exclusively on NPD data. A majority of 49 evaluations (60%) collected two or more external tests from schools and the remaining 24 evaluations (29%) used a single test.

Twelve evaluations (15%) did not collect any survey or interview data from schools for the IPE. A small majority of 43 evaluations (52%) collected both survey and interview data from schools and the remaining 27 evaluations (33%) collected survey or interview data but not both.

Drawing on both the testing and the IPE data collection burden, 22 evaluations (27%) are identified as having the highest level of burden in both aspects; 12 evaluations (15%) have the lowest burden in both.

Trial ITT analysis of primary outcome(s)

This subtheme was the only one in the review to include variables measured at the primary outcome / effect size level. These effect-size-level variables are used to measure specific detail about the type of primary outcome used across the 133 effect sizes in the review. These effect-size-level variables were included alongside three trial-level variables.

At a trial/evaluation level, the majority of evaluations had a single primary outcome (50 evaluations, 61%). However, the remaining 29 evaluations (39%) with two or more primary outcomes did provide the majority of effect sizes (83 effect sizes, 62%) for the review. Single primary outcomes became more common over time: 27 of the 52 evaluations published between 2014 and 2016 (52%), compared with 23 of the 30 evaluations published between 2017 and 2019 (77%) used a single primary outcome.

Also at the trial level, in most evaluations, the review found a direct match between the intervention focus and primary outcome(s) (47 evaluations, 57%), but in 10 evaluations (12%) this alignment was limited.

At an effect size level, the majority of primary outcomes were attainment measures relating to English or literacy (77 effect sizes, 58%) with maths being the second most common (38 effect sizes, 29%). Looking at the measures more closely, a majority of primary outcomes were commercial tests (79 effect sizes across 51 evaluations) with tests from GL Assessment being the most common (46 effect sizes across 33 evaluations). Official data on Key Stage attainment (or absence) were also common (45 effect sizes across 22 evaluations) with KS2 attainment being the most common (30 effect sizes across 15 evaluations).

More specific detail was sought on the type of primary outcome, but here the data become sparse. Thirteen specific outcomes[28] accounted for 91 of the 133 primary outcome effect sizes in the review (68%), with the GL NGRT being the most common primary outcome (23 effect sizes, 17%) followed by GL's Progress in English/Progress Test in English (13 effect sizes, 10%) and KS2 maths results (9 effect sizes, 7%).

A final trial-level variable drew on data from both the evaluation/trial level and the effect size level. At the trial level, 24 evaluations (29%) had a cross-curriculum focus. Effect-size-level detail was used to distinguish cross-curriculum trials that used composite/cross-curriculum outcome variables (e.g., total GCSE/KS2 score) from cross-curriculum trials that used 2+ primary outcomes in separate subject areas (e.g., attainment in GCSE maths and GCSE English separately). Of the 24 cross-curriculum trials, eight used composite cross-curriculum outcomes (10% of all trials, 33% of cross-curriculum trials) and 16 used 2+ outcomes in separate subjects (20% of all trials, 67% of cross-curriculum trials). Forty English trials (49%) were found to use just English attainment primary outcome(s); 15 maths trials (18%) were found to use just maths attainment primary outcome(s); and three science trials (4%) were found to use just science attainment primary outcome(s).

There was a close alignment between the curriculum focus variable in the intervention theme and the primary outcome subtheme of evaluation design. These measures were constructed differently. Under the intervention theme, the curriculum focus was ascertained through the review of evaluation reports. Under the evaluation design theme, the primary outcome was constructed from specific detail about the primary outcome (e.g., test) that drew on the EPPI database of trials, data from Lortie-Forgues & Inglis (2019) and additional data and quality checks across the 82 trials in the review. The close alignment between the two measures can be seen as a reliability cross-check for the review.

The 29 evaluations of interventions with a cross-curriculum focus included all 24 evaluations that were identified as cross-curriculum from their primary outcome measure(s). A further four interventions are identified as having a cross-curriculum focus here but only used English/literacy primary outcome(s). One intervention is identified as having a cross-curriculum focus here but only used maths/numeracy primary outcome(s). Aside from these five differences, the two variables are aligned for 77 of the 82 trials in the review (94%). This close agreement between the two data sources serves to boost confidence in the reliability of these measures.

[28] Whilst there is communality in the grouping of these 13 specific outcomes, they should not be considered 'the same'. For example, the GL PiM or PTM category will include both earlier PiM and current PTM tests across different age groups (PTM13, PiM11 etc). The 13 groupings are: seven commercial: three GL Assessments (NGRT; PiE or PTE; PiM or PTM); three CEM (InCAS maths; InCAS reading; CEM InCAS reading & maths); and one Hodder (GRT); six NPD/official (KS2 maths; KS2 reading; KS2 writing; GCSE maths; GCSE English; GCSE overall).

Findings 1: Meta-analyses of reported effect sizes for primary, secondary and FSM subsample outcomes

Introduction

The review focused on meta-analyses of the primary outcome: reported effect size(s) for ITT analyses of primary outcomes for the 82 EEF evaluations reported up to January 2019. Please see the *Technical Annex* for more comprehensive detail on this outcome and the meta-analysis approach. To summarise here, meta-analyses of effect sizes were undertaken at the effect size level (N_{ES} = 133), and used standard errors to account for variations in statistical uncertainty across these 133 effect sizes. This process results in effect sizes from RCT designs with relatively high statistical precision being accorded higher weight than effect sizes from RCT designs with relatively low statistical precision in the meta-analyses. Specifically, the meta-analyses constructed random effects models (Borenstein et al., 2009) to acknowledge that the 133 effect sizes in the review stem from a wide breadth of interventions and outcome measures. Fixed effects models assume the existence of a single 'real' effect size and are suited to meta-analyses of trials that all have a similar focus and outcome variable.

Whilst the theoretical framework and five overarching thematic areas provided a structure, the selection of explanatory variables under each theme was purposely broad to reflect the exploratory and descriptive nature of the review. This resulted in a sizeable quantity of explanatory variables (and hence analyses). This section succinctly summarises the key findings of these analyses; the complete analyses can be found in the *Technical Annex*.

The analyses of primary and secondary outcomes are summarised in a table for each of the five overarching thematic areas, presented at the start of each subsection. These tables are supplemented by scatter plots and interpretation. The values in these five tables are annotated as follows:

Table 20: Guide to analysis tables

When a statistically significant association was observed…	
… $p \leq 0.01$	✓***
… $p \leq 0.05$	✓**
… $p \leq 0.10$	✓*
Interesting pattern observed but $p > 0.10$	✓#
Explanatory variable included in the analyses but no obvious association observed	✓

Effect sizes and the intervention

Summary

Table 21: Summary of meta-analyses of ITT effect sizes and intervention

Subtheme	Explanatory variable	Primary ITT	Secondary ITT	FSM
Focus of the intervention	School phase	✓#	✓#	✓**
	School Key Stage	✓#	✓***	✓**
	Curriculum focus of intervention	✓*	✓***	✓#
Intensity	Minutes per week	✓#	✓#	✓#
Who implements with direct target?	Direct implementer (teacher / TA / external)	✓***	✓***	✓#
Perceived quality of supporting resources	High / varied / low	✓	✓***	✓***

Subtheme	Explanatory variable	Primary ITT	Secondary ITT	FSM
Cost	Total cost	✓***	✓***	✓***
	Cost per pupil (over three years)	✓**	✓***	✓#
EEF intervention themes	Language and literacy	✓	✓	✓
	Maths and numeracy	✓	✓	✓
	Staff deployment and development	✓	✓	✓
	Organising your school	✓	✓	✓
	Developing effective learners	✓	✓	✓
	Feedback and monitoring pupil progress	✓	✓	✓
	Behaviour	✓	✓	✓
	Character and essential skills	✓	✓	✓
	Parental engagement	✓	✓	✓
	Science	✓	✓	✓
	Enrichment	✓	✓	✓
	Early years	✓	✓	✓
	Special educational needs	✓	✓	✓
EEF promising project	Whether identified as promising on EEF website	✓***	✓***	✓***

Focus of the intervention

School phase and key stage

Primary ITT

On average, for interventions in primary schools, effect sizes were higher in KS1 (weighted mean = +0.08 SD; 95% CI: +0.04 to +0.11) compared with KS2 (weighted mean = +0.03 SD; 95% CI: +0.01 to +0.05). On average, for interventions in secondary schools, effect sizes were higher in KS3 (weighted mean = +0.06 SD; 95% CI: −0.01 to +0.12) compared with KS4 (weighted mean = +0.04 SD; 95% CI: +0.01 to +0.07).

On average, interventions with a focus on Y6–Y7 primary–secondary transition were associated with a higher effect size (weighted mean = +0.12 SD; 95% CI: −0.01 to +0.25) compared with interventions in primary or in secondary schools (weighted mean = +0.04 SD or lower), although the distinction was not statistically significant. This contrasts with the qualitative findings of Anders et al. (2017) that trials in the transition phase were less likely to be successful. However, caution is needed in interpreting the findings of this review due to the limited number of transition trials (n_t = 6) and primary ITT effect sizes that these trials reported (n_{es} = 7). Additionally, one transition trial reported an exceptionally high effect size.[29]

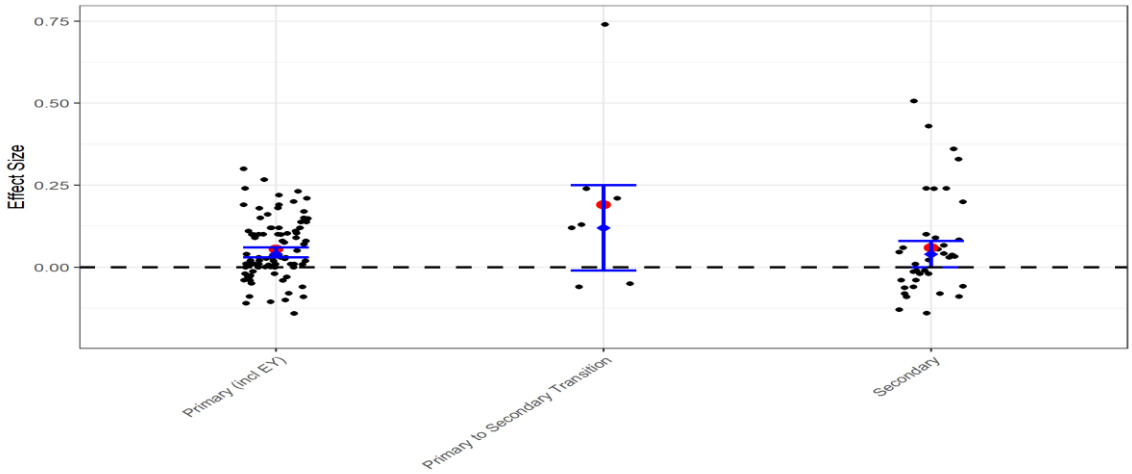
Table 22: Effect size by school phase primary ITT attainment outcomes

Factor	*n*	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Primary (including Early Years)	88	+0.05	0.093	+0.03	+0.06
Primary to secondary transition	7	+0.12	0.067	−0.01	+0.25
Secondary	38	+0.04	0.020	+0.00	+0.08

Meta *p*-value < 0.10*. Overall weighted mean = +0.04 SD.

[29] An effect size of 0.74 SD reported for a GL PiE (writing subscale) outcome in the first IPEEL efficacy trial (Torgerson, 2014; SHU ID 96). This trial also reported an effect size of +1.60 SD for FSM pupils. However, this trial was awarded a relatively low 2 EEF padlocks for trial quality and so further caution is warranted.

Figure 12: Effect size by school phase primary ITT attainment outcomes



Secondary ITT

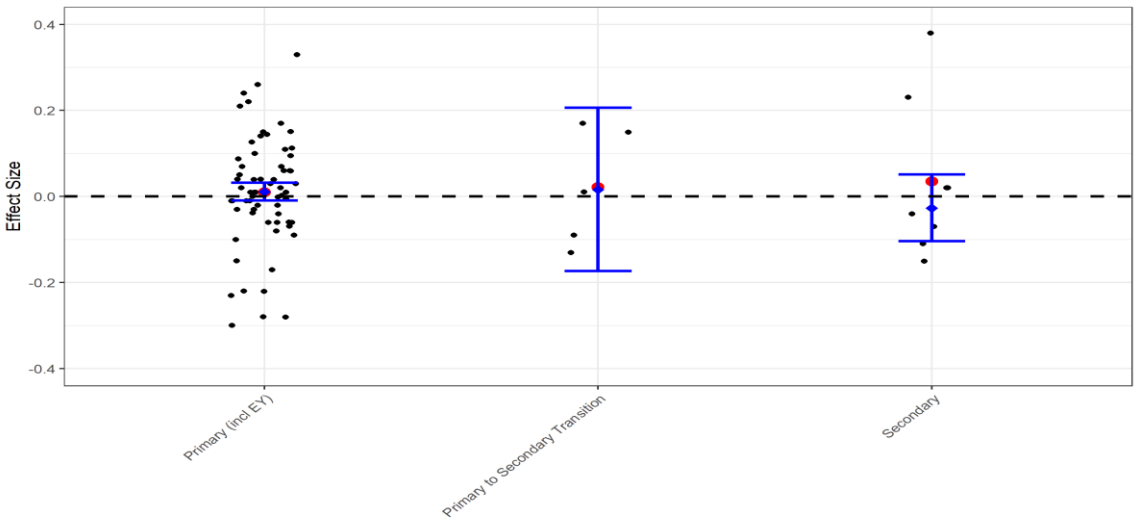
On average, for interventions in primary schools, effect sizes were higher in KS1 (weighted mean = +0.09 SD; 95% CI: +0.05 to +0.12) compared with KS2 (weighted mean = −0.01 SD; 95% CI: −0.02 to +0.01). On average, for interventions in secondary schools (weighted mean = −0.03 SD; 95% CI: −0.10 to +0.05), effect sizes were lower than interventions in primary schools (weighted mean = +0.01 SD; 95% CI: −0.01 to +0.03) but the difference between these is not statistically significant.

Table 23: Effect size by school phase secondary ITT attainment outcomes

Factor	*n*	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Primary (including Early Years)	65	+0.01	0.010	−0.01	+0.03
Primary to secondary transition	5	+0.02	0.097	−0.17	+0.21
Secondary	8	−0.03	0.040	−0.10	+0.05

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.01 SD.

Figure 13: Effect size by school phase secondary ITT attainment outcomes



FSM

On average, for interventions in secondary schools (weighted mean = 0.00 SD; 95% CI: −0.04 to +0.03), effect sizes were significantly lower than interventions in primary schools (weighted mean = +0.04 SD; 95% CI: +0.02 to +0.06). Within primary schools, on average, FSM effect sizes were higher in KS1 (weighted mean = +0.07 SD; 95% CI: +0.02 to +0.12) compared with KS2 (weighted mean = +0.04 SD; 95% CI: +0.01 to +0.06). Within secondary schools, on

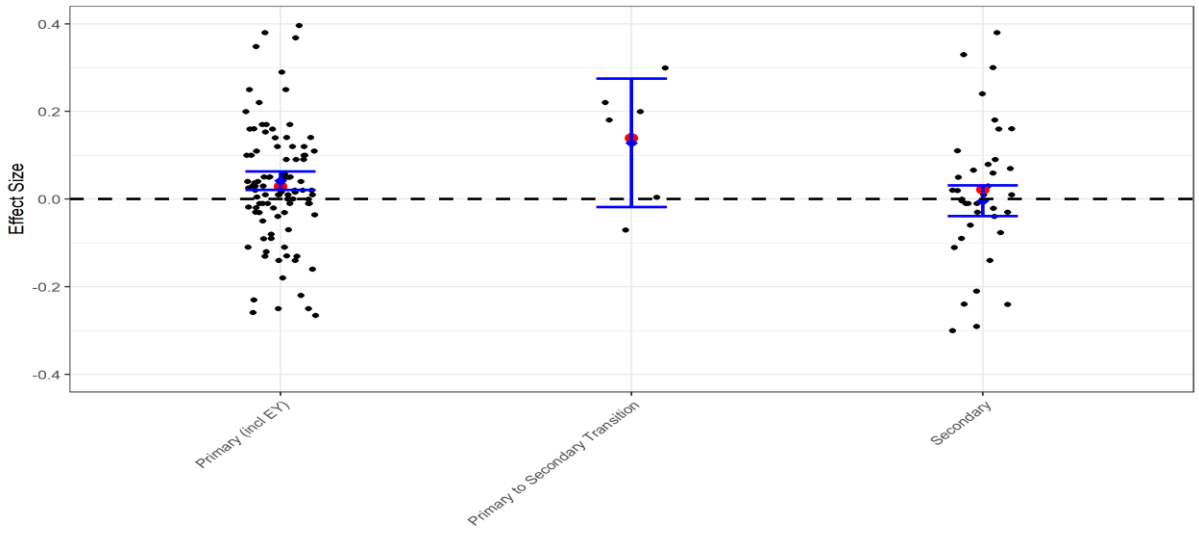
average, FSM effect sizes were higher in KS4 (weighted mean = +0.04 SD; 95% CI: 0.00 to +0.08) compared with KS3 (weighted mean = −0.01 SD; 95% CI: −0.08 to +0.06). Similar to what is seen with primary ITT effect sizes, the mean FSM effect sizes reported for interventions that focused on Y6–Y7 transition was notably high (+0.13 SD) but the wide CI (95% CI: −0.02; +0.28) highlight the small number of effect sizes in this group (*n* = 7). Further, this grouping also includes an exceptionally high FSM effect size reported by the same trial noted above for primary ITT effect sizes.

Table 24: Effect size by school phase FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Primary (including Early Years)	102	+0.04	+0.011	+0.02	+0.06
Primary to secondary transition	7	+0.13	+0.075	−0.02	+0.28
Secondary	40	0.00	+0.018	−0.04	+0.03

Meta *p*-value < 0.05**. Overall weighted mean = +0.03 SD.

Figure 14: Effect size by school phase FSM attainment outcomes



Curriculum focus

Primary ITT

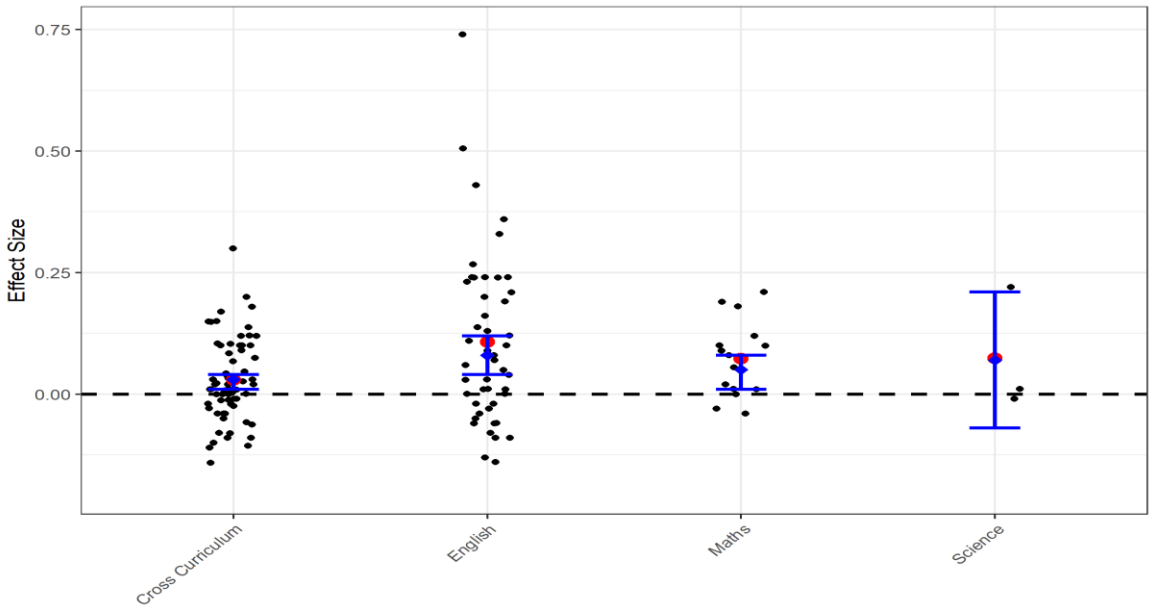
On average, interventions with an English curriculum focus were associated with higher effect sizes (weighted mean = +0.08 SD; 95% CI: +0.04 to +0.12) compared with trials with a maths curriculum focus (weighted mean = +0.05 SD; 95% CI: +0.01 to +0.08) or cross-curriculum focus (weighted mean = +0.03 SD; 95% CI: +0.01 to +0.04). The association between the curriculum focus of an intervention and effect size was statistically significant at the 10% level ($p < 0.10$) (Figure 15, Table 25). The weaker effect size for cross-curriculum interventions is potentially an area for future investigation. A number of factors may explain the finding, including implementing a cross-curriculum intervention requires a higher level of co-ordination than a single subject intervention; maintaining fidelity across a school may be more challenging, particularly in large secondary schools; and cross-curriculum change may take longer to become embedded. A further possible explanation for the finding may be weaker alignment between the cross-curriculum interventions and the primary outcome test than is possible for single-subject interventions.

Table 25: Effect size by curriculum area primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Cross-curriculum	67	+0.03	0.009	+0.01	+0.04
English	48	+0.08	0.021	+0.04	+0.12
Maths	15	+0.05	0.017	+0.01	+0.08
Science	3	–	–	–	–

Meta *p*-value <0.10*. Overall weighted mean = +0.04 SD.

Figure 15: Effect size by curriculum area primary ITT attainment outcomes



Secondary ITT

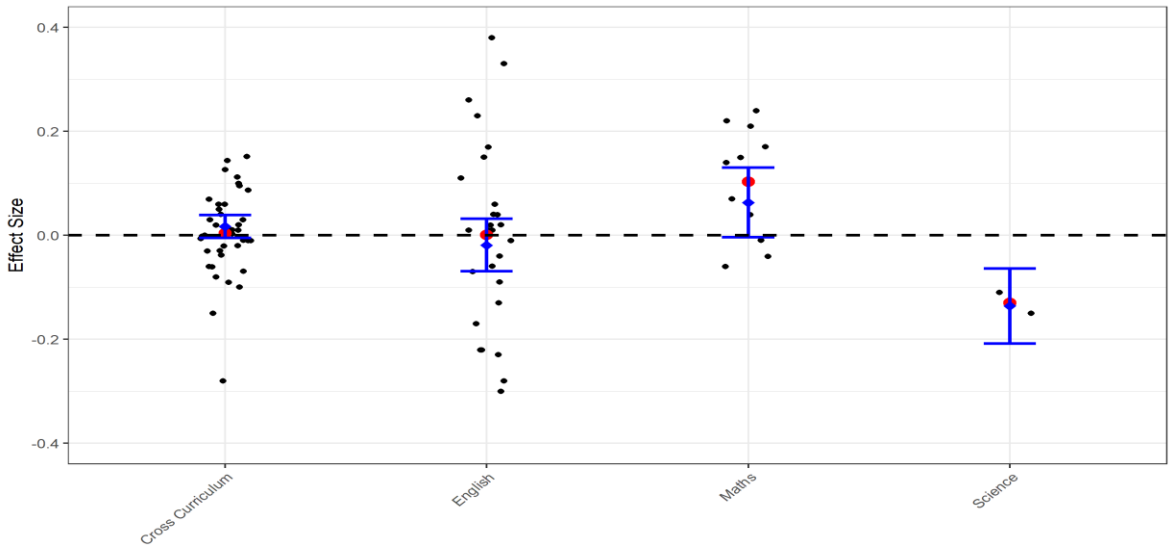
On average, interventions with a maths curriculum focus were associated with higher secondary ITT effect sizes (weighted mean = +0.06 SD; 95% CI: 0.00 to +0.13) compared with trials with a cross-curriculum focus (weighted mean = +0.02 SD; 95% CI: −0.01 to +0.04) or English curriculum focus (weighted mean = −0.02 SD; 95% CI: −0.07 to +0.03).

Table 26: Effect size by curriculum area secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Cross-curriculum	38	+0.02	0.011	−0.01	+0.04
English	27	−0.02	0.026	−0.07	+0.03
Maths	11	+0.06	0.034	0.00	+0.13
Science	2	–	–	–	–

Meta *p*-value <0.01***. Overall weighted mean = +0.01 SD.

Figure 16: Effect size by curriculum area secondary ITT attainment outcomes



FSM

No evidence of statistically significantly different FSM effect sizes was observed for interventions with differing curriculum focus (effect sizes ranged between +0.03 and +0.04 SD).

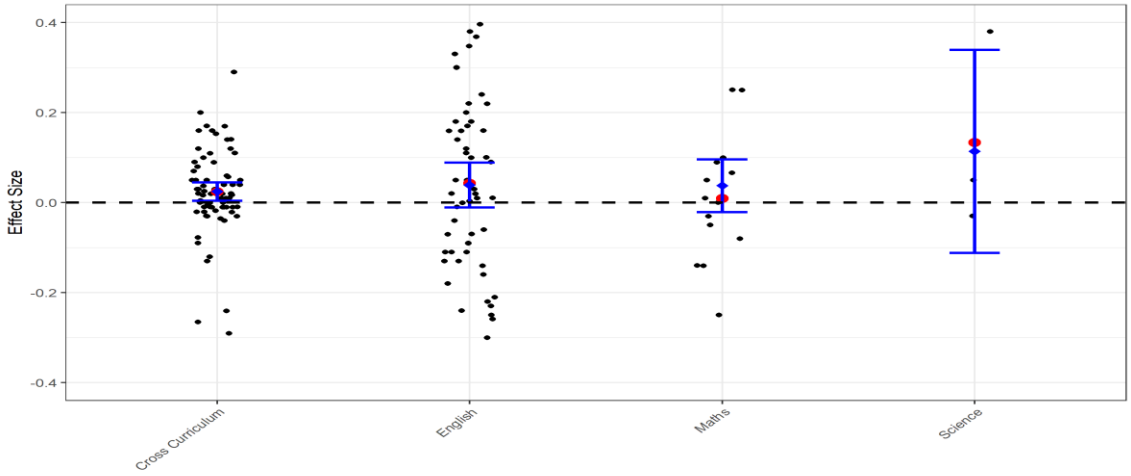
Table 27: Effect size by curriculum area FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Cross-curriculum	70	+0.03	0.010	0.00	+0.05
English	61	+0.04	0.025	−0.01	+0.09
Maths	15	+0.04	0.030	−0.02	+0.10
Science	3	–	–	–	–

Meta *p*-value >0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 17: Effect size by curriculum area FSM attainment outcomes

Intensity of intervention (minutes per week)



Primary ITT

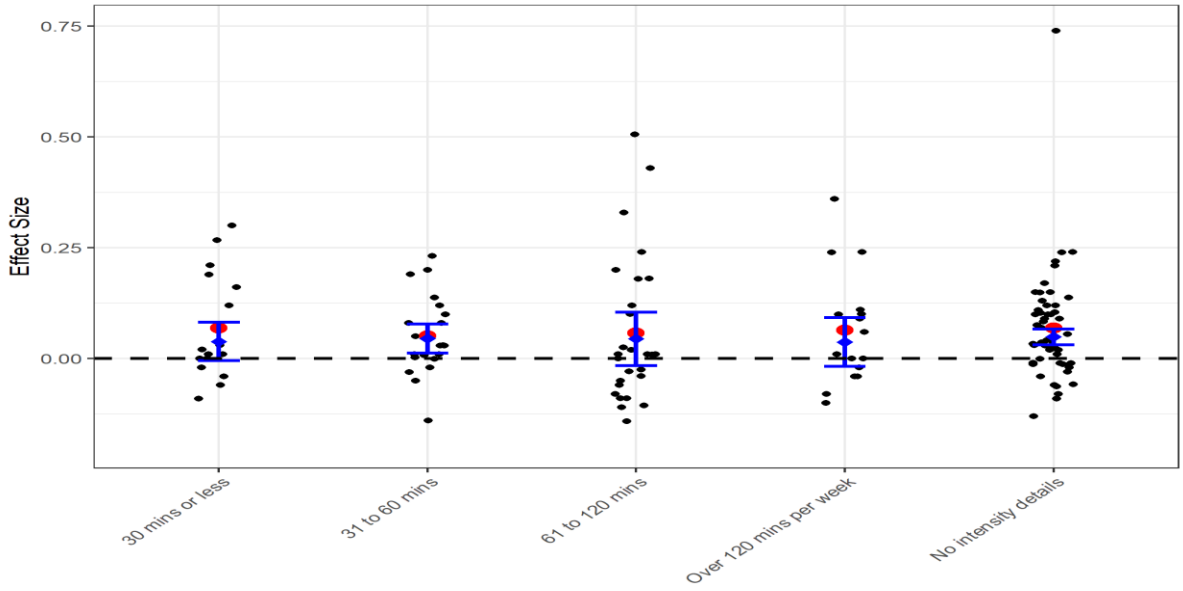
No evidence of an association between the intensity of an intervention and effect size was observed. Across the categories of intensity, the weighted mean effect size ranged between +0.04 and +0.05 SD.

Table 28: Effect size by intensity of intervention primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Up to 30 min/week	16	+0.04	0.022	+0.00	+0.08
31–60 min/week	21	+0.05	0.017	+0.01	+0.08
61–120 min/week	27	+0.04	0.031	−0.02	+0.10
Over 121 min/week	16	+0.04	0.029	−0.02	+0.09
No intensity detail	53	+0.05	0.009	+0.03	+0.07

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Figure 18: Effect size by intensity of intervention primary ITT attainment outcomes



Secondary ITT

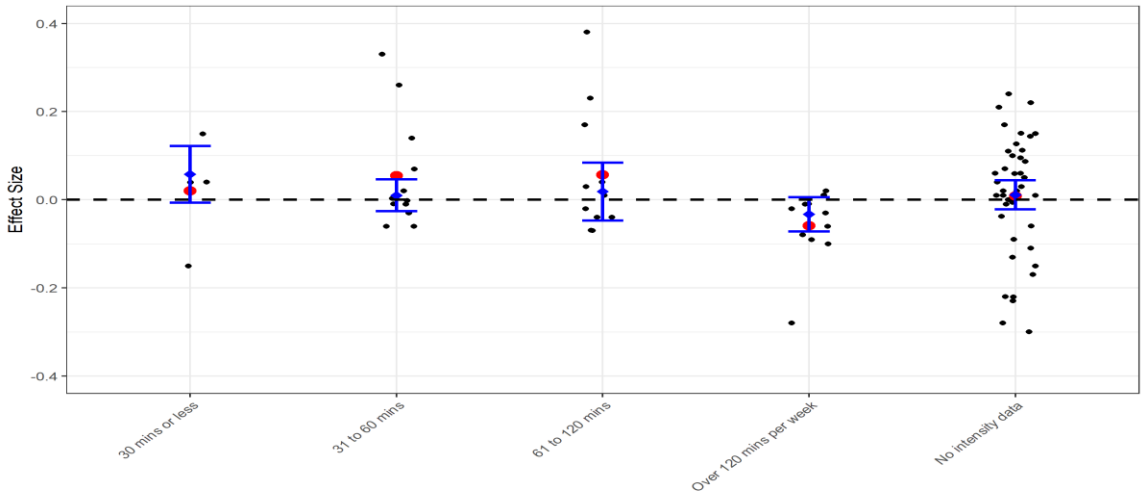
No evidence of an association between the intensity of an intervention and secondary ITT effect size was observed. Across the categories of intensity, the weighted mean effect size ranged between +0.01 and +0.06 SD.

Table 29: Effect size by intensity of intervention secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Up to 30 min/week	4	+0.06	0.033	−0.01	+0.12
31–60 min/week	12	+0.01	0.019	−0.03	+0.05
61–120 min/week	11	+0.02	0.033	−0.05	+0.08
Over 121 min/week	11	−0.03	0.020	−0.07	+0.01
No intensity detail	40	+0.01	0.017	−0.02	+0.04

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.01 SD.

Figure 19: Effect size by intensity of intervention secondary ITT attainment outcomes



FSM

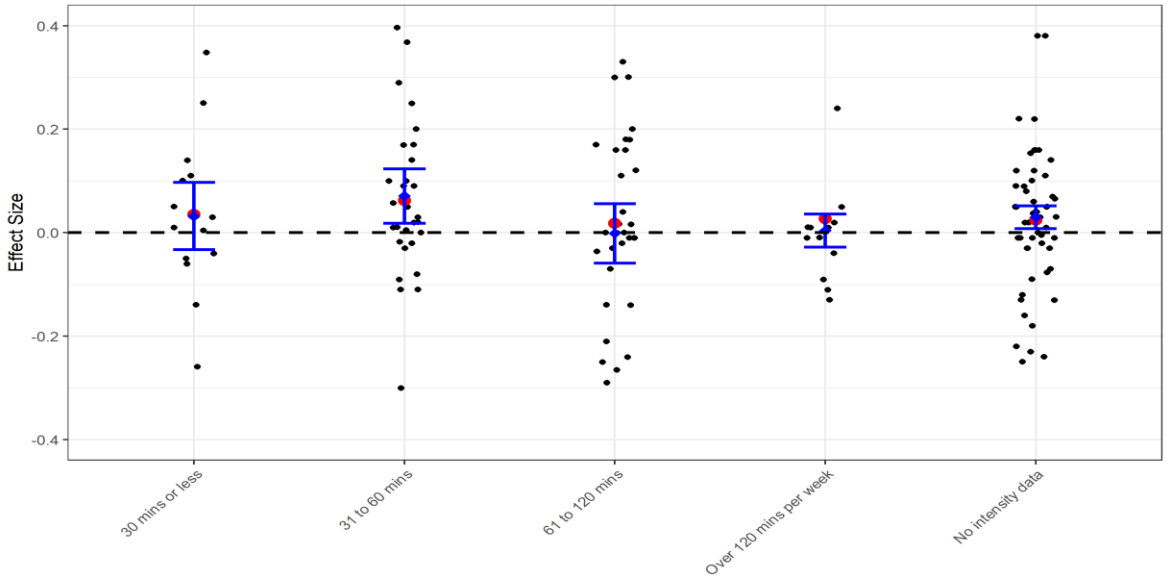
No evidence of an association between the intensity of an intervention and FSM effect size was observed. Across the categories of intensity, the weighted mean effect size ranged between 0.00 and +0.07 SD.

Table 30: Effect size by intensity of intervention FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Up to 30 min/week	18	+0.03	0.033	−0.03	+0.10
31–60 min/week	29	+0.07	0.027	+0.02	+0.12
61–120 min/week	32	0.00	0.029	−0.06	+0.06
Over 121 min/week	13	0.00	0.016	−0.03	+0.04
No intensity detail	57	+0.03	0.011	+0.01	+0.05

Meta *p*-value >0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 20: Effect size by intensity of intervention FSM attainment outcomes



Who implements with direct target?

Primary ITT

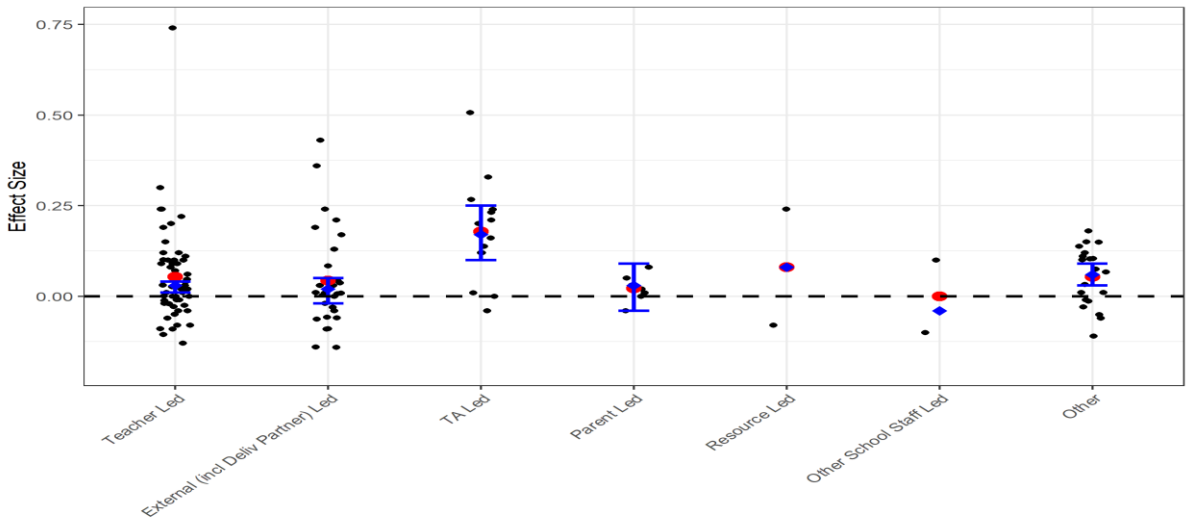
On average, the 12 TA-led interventions (reporting a total of 15 effect sizes) are associated with statistically significantly ($p < 0.01$) higher effect sizes (weighted mean = +0.17 SD; 95% CI: +0.10 to +0.25) compared with interventions led by others (weighted mean = +0.03 SD or lower) (Figure 21, Table 31). This aligns with earlier findings and is likely to be associated with the mode of delivery and fidelity to the intervention. TA-led interventions tend to be delivered on a one-to-one basis, which have also been shown to be associated with high effect sizes. TA interventions were usually fairly tightly codified and all TAs used the same resources and so these programmes were more likely to be implemented with fidelity. Tight codification may be particularly important for TA-led interventions, where research has shown that the effectiveness of TA support is increased when clearly defined structured interventions that are aligned to pupil's needs are used (Sharples et al., 2015).

Table 31: Effect size by direct implementer primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Teacher-led	57	+0.03	0.009	+0.01	+0.04
Externally-led (incl. delivery partner)	30	+0.02	0.018	−0.02	+0.05
TA-led	15	+0.17	0.037	+0.10	+0.25
Parent-led	7	+0.03	0.032	−0.04	+0.09
Resource-led	2	–	–	–	–
Other school staff-led	2	–	–	–	–
Other	20	+0.06	0.017	+0.03	+0.09

Meta *p*-value < 0.01***. Overall weighted mean = +0.04 SD.

Figure 21: Effect size by direct implementer primary ITT attainment outcomes



Secondary ITT

TA-led interventions were associated with higher secondary attainment effect sizes (weighted mean = +0.03 SD; 95% CI: −0.04 to +0.09; 5 effect sizes reported by 4 evaluations) compared with interventions led by teachers, delivery partners and other school staff (weighted mean = +0.01 SD or lower). However, the weighted mean effect size for TA-led interventions and the difference between this and the weighted mean effect size of interventions led by others is less than observed with the primary outcomes, interventions with direct implementers other than TAs, teachers, delivery

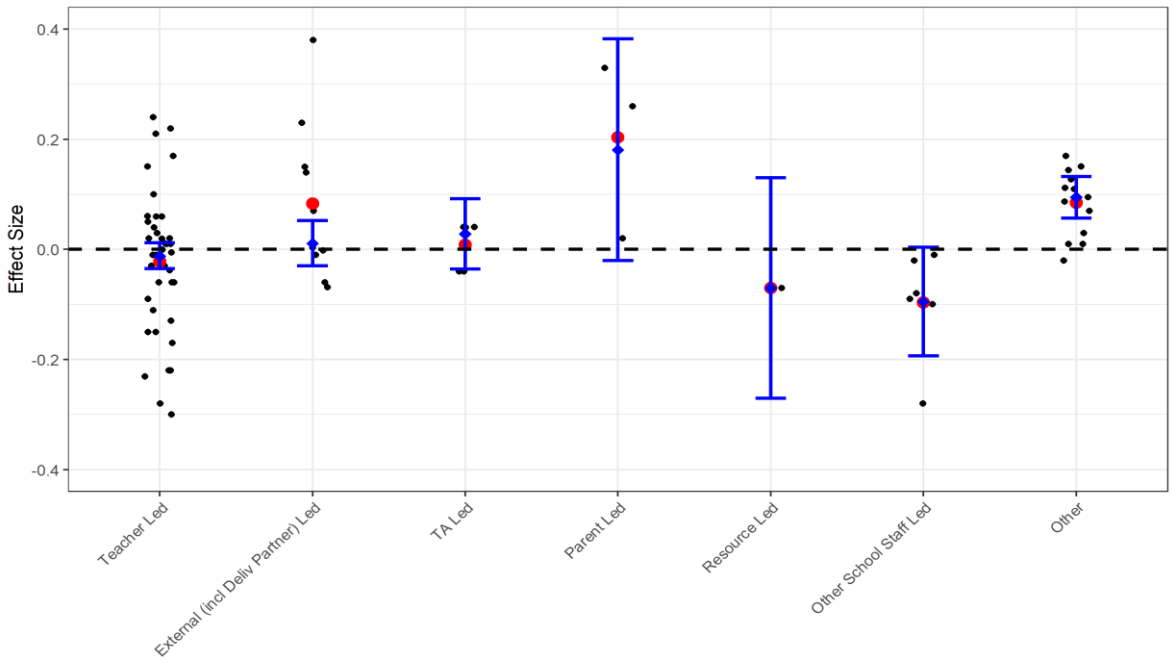
partners, other school staff, parents or resources were observed to be the largest (weighted mean = +0.10 SD; 95% CI: +0.06 to +0.13). This may arise where specialist organisations are employed to deliver an intervention that is very closely aligned to the secondary outcome/s.

Table 32: Effect size by direct implementer secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Teacher-led	40	−0.01	0.012	−0.04	+0.01
Externally-led (incl. delivery partner)	10	+0.01	0.021	−0.03	+0.05
TA-led	5	+0.03	0.033	−0.04	+0.09
Parent-led	3	–	–	–	–
Resource-led	1	–	–	–	–
Other school staff-led	6	−0.10	0.050	−0.19	0.00
Other	13	+0.10	0.019	+0.06	+0.13

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Figure 22: Effect size by direct implementer secondary ITT attainment outcomes



FSM

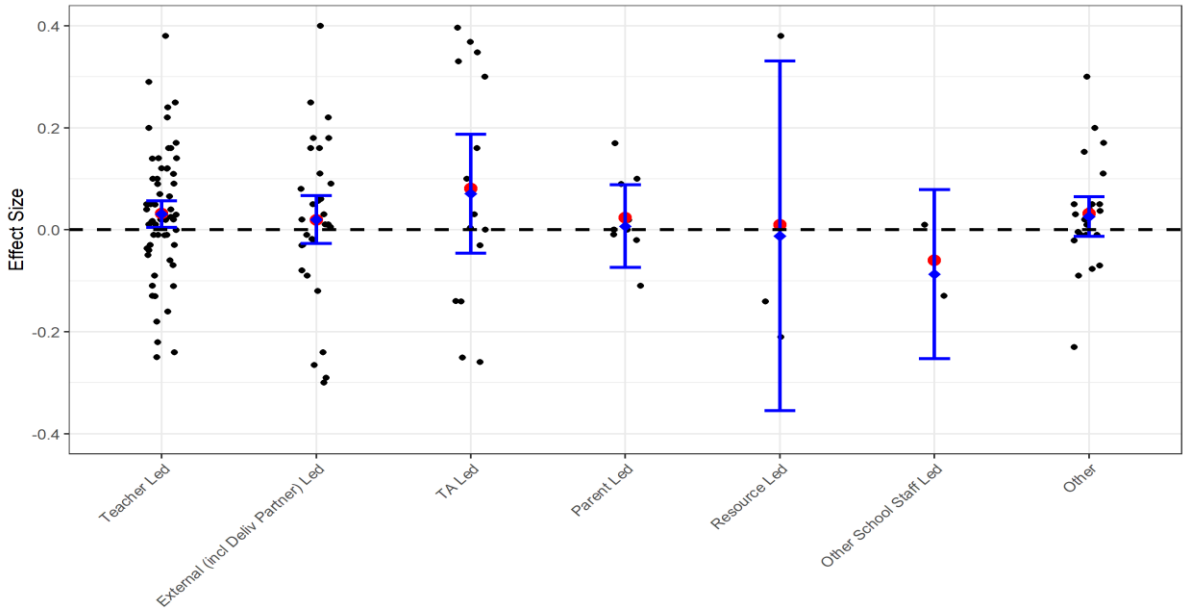
No evidence of an association between direct implementers and FSM effect size was observed. Across the categories, the weighted mean effect size ranged between +0.01 and +0.07 SD, with the highest weighted mean being attributed to the 10 TA-led interventions (reporting 17 effect sizes). However, the lack of statistical significance of association indicates that where the focus of the intervention is to improve attainment of pupils in receipt of FSM rather than all pupils, TA-led interventions may not be any more effective that interventions led by others.

Table 33: Effect size by direct implementer FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Teacher-led	63	+0.03	0.013	+0.01	+0.06
Externally-led (incl. delivery partner)	32	+0.02	0.024	−0.03	+0.07
TA-led	17	+0.07	0.059	−0.05	+0.19
Parent-led	10	+0.01	0.041	−0.07	+0.09
Resource-led	3	–	–	–	–
Other school staff-led	2	–	–	–	–
Other	22	+0.03	0.020	−0.01	+0.07

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 23: Effect size by direct implementer FSM attainment outcomes



Perceived quality of supporting resources

Primary ITT

No evidence of an association between the perceived quality of supporting resources and effect size was observed.

Table 34: Effect size by perceived quality of supporting resources primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
High	27	+0.06	0.016	+0.02	+0.09
Variation	40	+0.05	0.020	+0.01	+0.08
Low	6	+0.03	0.028	−0.03	+0.09
Not mentioned	60	+0.04	0.011	+0.02	+0.06

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Secondary ITT

The association between secondary ITT effect size and perceptions on the quality of supporting resources was statistically significant at the 1% level ($p < 0.01$). When reported perceptions on the quality of supporting resources were considered to be low or perceptions varied across participants/schools, the mean secondary ITT effect size was lower (−0.04 SD or lower) than when reported perceptions were high (+0.01 SD) or when perceptions on quality of resources were not mentioned (+0.03 SD).

Table 35: Effect size by perceived quality of supporting resources secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
High	25	+0.01	0.010	−0.01	+0.03
Variation	19	−0.04	0.027	−0.10	+0.01
Low	4	−0.12	0.032	−0.18	−0.05
Not mentioned	30	+0.03	0.020	−0.01	+0.07

Meta *p*-value > 0.01***. Overall weighted mean = +0.01 SD.

FSM

The association between FSM effect size and perceptions on the quality of supporting resources was statistically significant at the 1% level ($p < 0.01$). A complex pattern is observed with the highest mean effect size was when perceptions on quality of supporting resources were high (+0.10 SD; 95% CI: +0.05 to +0.15) and the second highest was when perceptions on quality of supporting resources were low (+0.05 SD; 95% CI: −0.02 to +0.12). Lower mean effect sizes were observed when perceptions on quality of supporting resources were varied or not mentioned (+0.02 or lower).

Table 36: Effect size by perceived quality of supporting resources FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
High	30	+0.10	0.025	+0.05	+0.15
Variation	59	0.00	0.011	−0.02	+0.02
Low	5	+0.05	0.035	−0.02	+0.12
Not mentioned	55	+0.02	0.017	−0.01	+0.05

Meta *p*-value < 0.01***. Overall weighted mean = +0.03 SD.

Cost of the intervention

Total cost

Primary ITT

The association between effect size and the total cost of an intervention is complex and statistically significant ($p < 0.01$). On average, higher effect sizes are observed for interventions that cost between £250k and less than £500k (weighted mean effect size = +0.09 SD; 95% CI: +0.06 to +0.11), compared with cheaper interventions (weighted mean = +0.02 SD or lower) or more expensive interventions (weighted mean = +0.05 SD or lower).

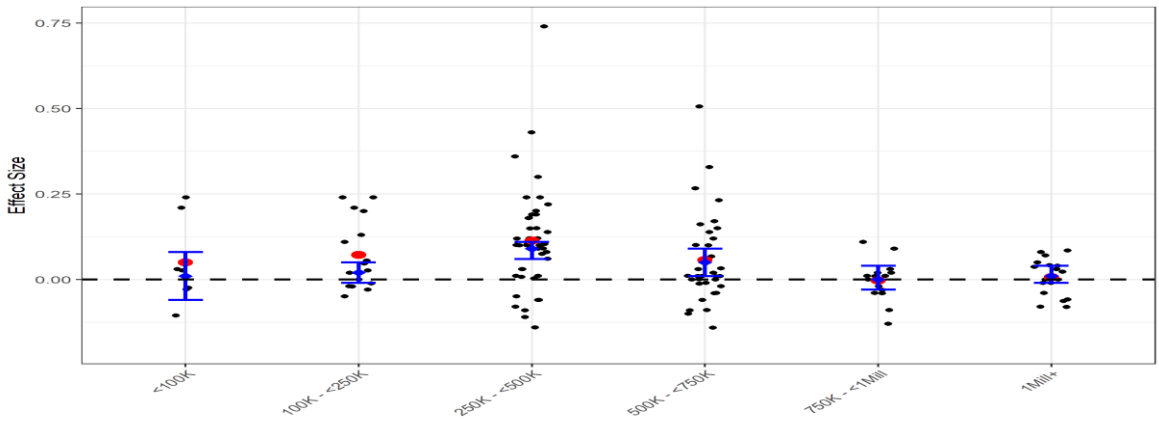
Table 37: Effect size by total cost primary ITT attainment outcomes

Factor	*n*	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high

<100k	7	+0.01	0.035	−0.06	+0.08
100k–<250k	16	+0.02	0.015	−0.01	+0.05
250k–<500k	44	+0.09	0.014	+0.06	+0.11
500k–<750k	33	+0.05	0.021	+0.01	+0.09
750k–<1 million	15	0.00	+0.018	−0.03	+0.04
1 million+	18	+0.01	+0.014	−0.01	+0.04

Meta *p*-value < 0.01***. Overall weighted mean = +0.04 SD.

Figure 24: Effect size by total cost primary ITT attainment outcomes



Secondary ITT

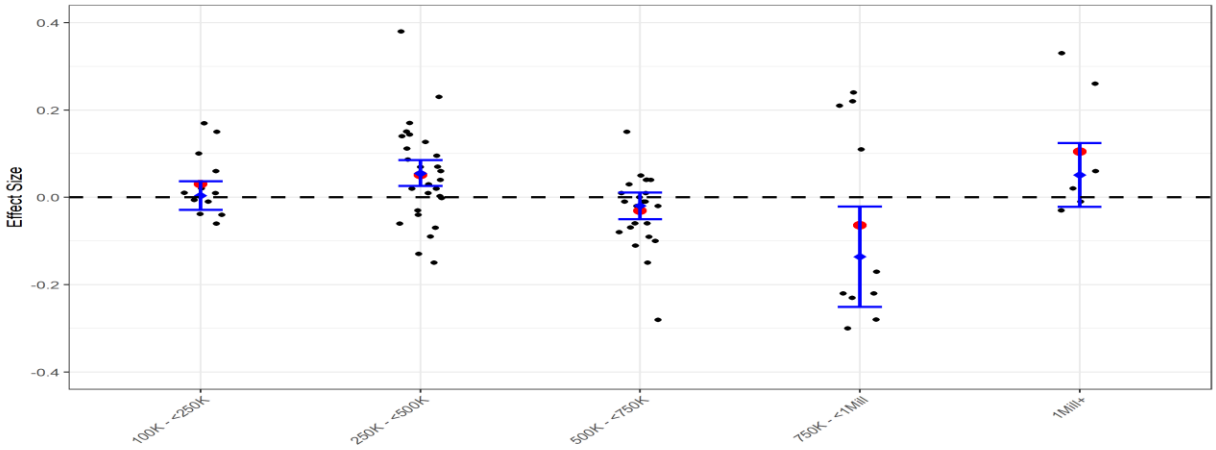
The association between secondary ITT effect size and the total cost of an intervention is also complex and statistically significant ($p < 0.01$). On average, higher effect sizes are observed for interventions that cost between £250k and less than £500k (weighted mean effect size = +0.06 SD; 95% CI: +0.03 to +0.08) compared with cheaper interventions (weighted mean = 0.00 SD; 95% CI: −0.03 to +0.04) or more expensive interventions (weighted mean = −0.02 SD or lower).

Table 38: Effect size by total cost secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
<100k	0	–	–	–	–
100k–<250k	12	0.00	0.017	−0.03	+0.04
250k–<500k	27	+0.06	0.015	+0.03	+0.08
500k–<750k	23	−0.02	0.015	−0.05	+0.01
750k–<1 million	10	−0.14	0.059	−0.25	−0.02
1 million +	6	+0.05	0.037	−0.02	+0.12

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Figure 25: Effect size by total cost secondary ITT attainment outcomes



* **Note**: there were no secondary ITT attainment outcomes for interventions with a total cost less than £100k.

FSM

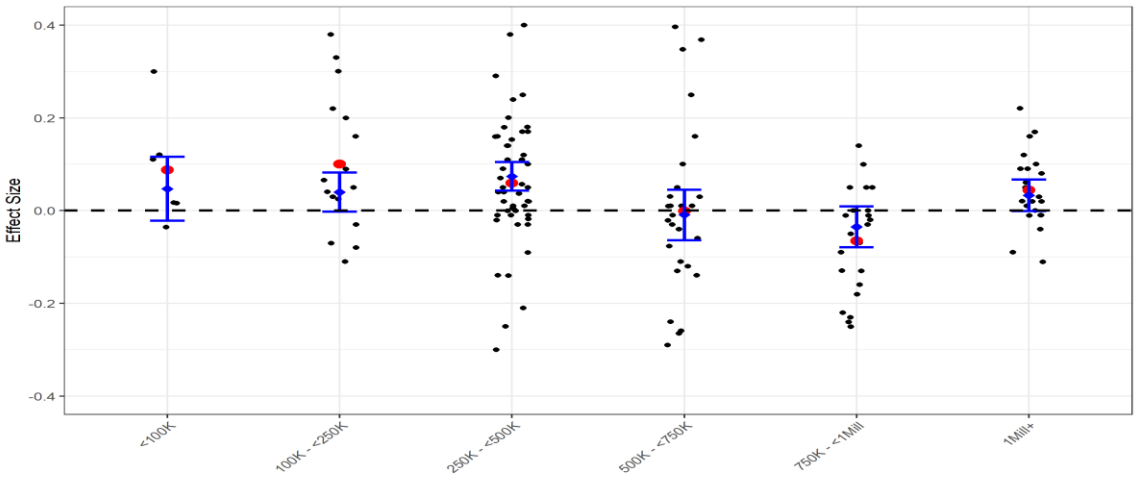
The association between FSM effect size and the total cost of an intervention is also complex and statistically significant ($p < 0.01$). On average, higher effect sizes are observed for interventions that cost between £250k and less than £500k (weighted mean effect size = +0.07 SD; 95% CI: +0.04 to +0.11) compared with cheaper interventions (weighted mean = +0.05 or lower) or more expensive interventions (weighted mean = +0.03 SD or lower).

Table 39: Effect size by total cost FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
<100k	6	+0.05	0.035	−0.02	+0.12
100k–<250k	16	+0.04	0.022	0.00	+0.08
250k–<500k	53	+0.07	0.016	+0.04	+0.11
500k–<750k	30	−0.01	0.028	−0.06	+0.05
750k–<1 million	22	−0.04	0.022	−0.08	+0.01
1 million +	22	+0.03	0.017	0.00	+0.07

Meta *p*-value < 0.01***. Overall weighted mean = +0.03 SD.

Figure 26: Effect size by total cost FSM attainment outcomes



Cost per pupil (over three years)

Primary ITT

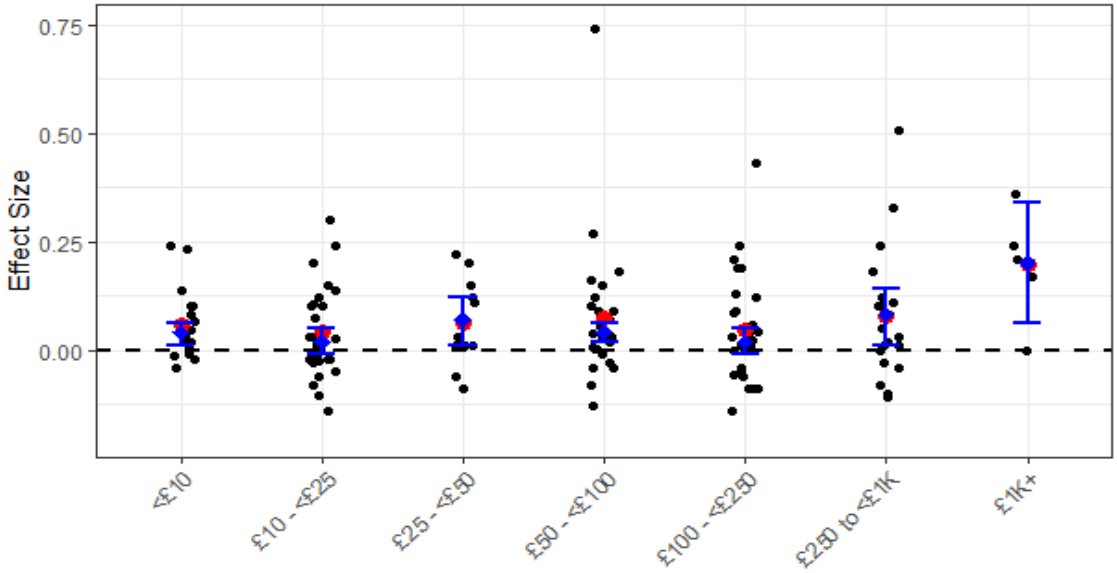
The association between effect size and the cost per pupil is also complex and statistically significant ($p < 0.05$). On average, higher effect sizes are observed for the three interventions that cost £1,000 or more per pupil (weighted mean effect size = +0.20 SD; 95% CI: +0.06 to +0.34) compared with cheaper interventions (weighted mean = +0.08 SD or lower).

Table 40: Effect size by cost per pupil primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
<£10	17	+0.04	0.013	+0.01	+0.06
£10–<£25	28	+0.02	0.015	−0.01	+0.05
£25–<£50	12	+0.07	0.028	+0.01	+0.12
£50–<£100	24	+0.04	0.012	+0.02	+0.06
£100–<£250	27	+0.02	0.018	−0.01	+0.05
£250–<£1,000	20	+0.08	0.035	+0.01	+0.14
£1,000+	5	+0.20	0.073	+0.06	+0.34

Meta *p*-value < 0.05**. Overall weighted mean = +0.04 SD.

Figure 27: Effect size by cost per pupil primary ITT attainment outcomes



Secondary ITT

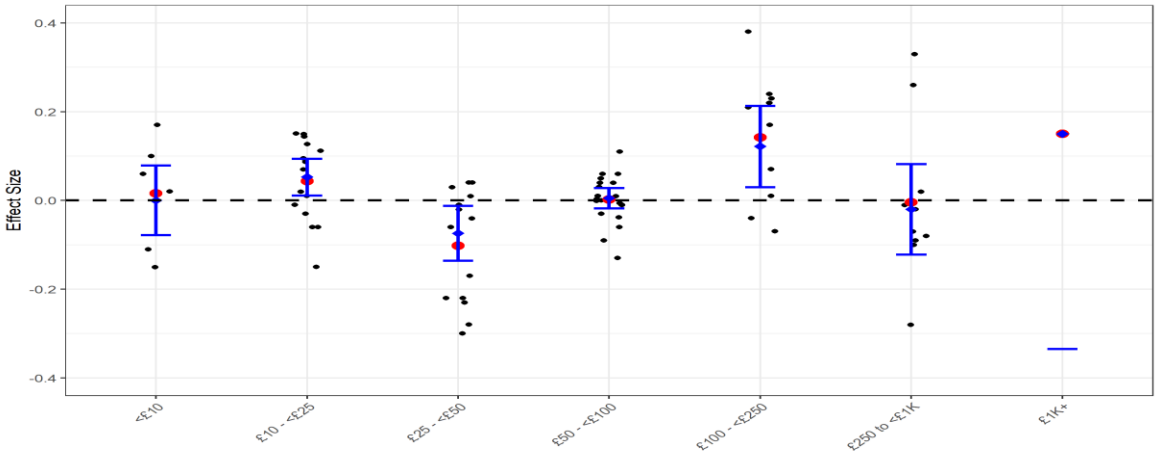
The association between secondary ITT effect size and the cost per pupil is also complex and statistically significant ($p < 0.05$). On average, higher effect sizes are observed for the interventions that cost £100–£250 per pupil (weighted mean effect size = +0.12 SD; 95% CI: +0.03 to +0.21) and interventions that cost £10–£25 per pupil (weighted mean effect size = +0.05 SD; 95% CI: +0.01 to +0.09), with other cost levels having lower mean effect sizes (between −0.07 and 0.00 SD).

Table 41: Effect size by cost per pupil secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
<£10	7	0.00	0.040	−0.08	+0.08
£10–<£25	15	+0.05	0.021	+0.01	+0.09
£25–<£50	14	−0.07	0.032	−0.14	−0.01
£50–<£100	20	+0.01	0.012	−0.02	+0.03
£100–<£250	11	+0.12	0.047	+0.03	+0.21
£250–<£1,000	10	−0.02	0.052	−0.12	+0.08
£1,000+	1	–	–	–	–

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Figure 28: Effect size by cost per pupil secondary ITT attainment outcomes



FSM

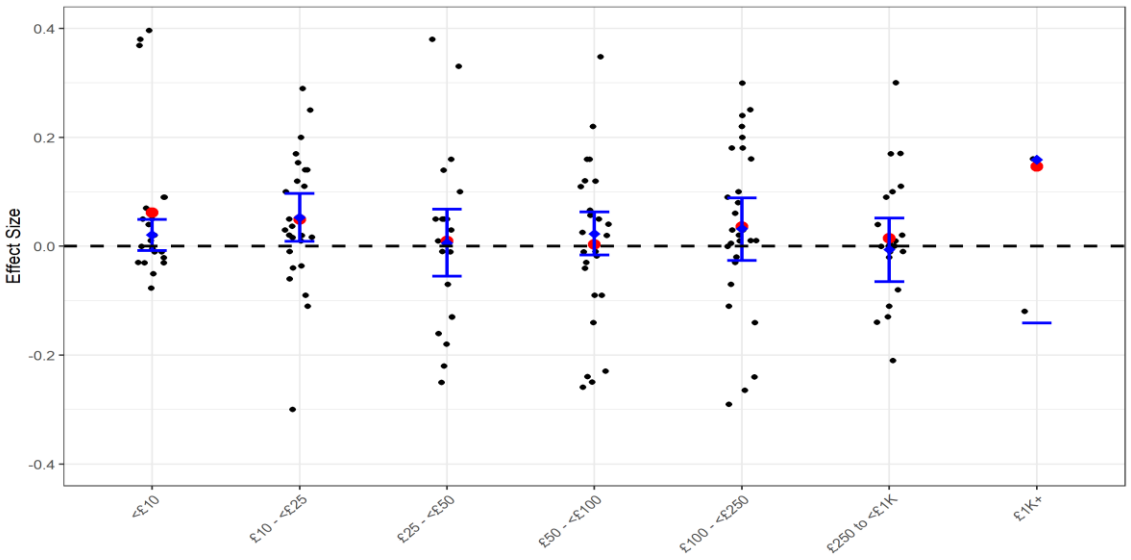
For FSM effect sizes, there is no evidence of a statistically significant association with cost per pupil, with effect sizes ranging between +0.01 and +0.05 across categories.

Table 42: Effect size by cost per pupil FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
<£10	21	+0.02	0.015	−0.01	+0.05
£10–<£25	25	+0.05	0.022	+0.01	+0.10
£25–<£50	20	+0.01	0.031	−0.06	+0.07
£50–<£100	28	+0.02	0.020	−0.02	+0.06
£100–<£250	31	+0.03	0.030	−0.03	+0.09
£250–<£1,000	21	−0.01	0.030	−0.07	+0.05
£1000+	3	–	–	–	–

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 29: Effect size by cost per pupil FSM attainment outcomes



Primary ITT

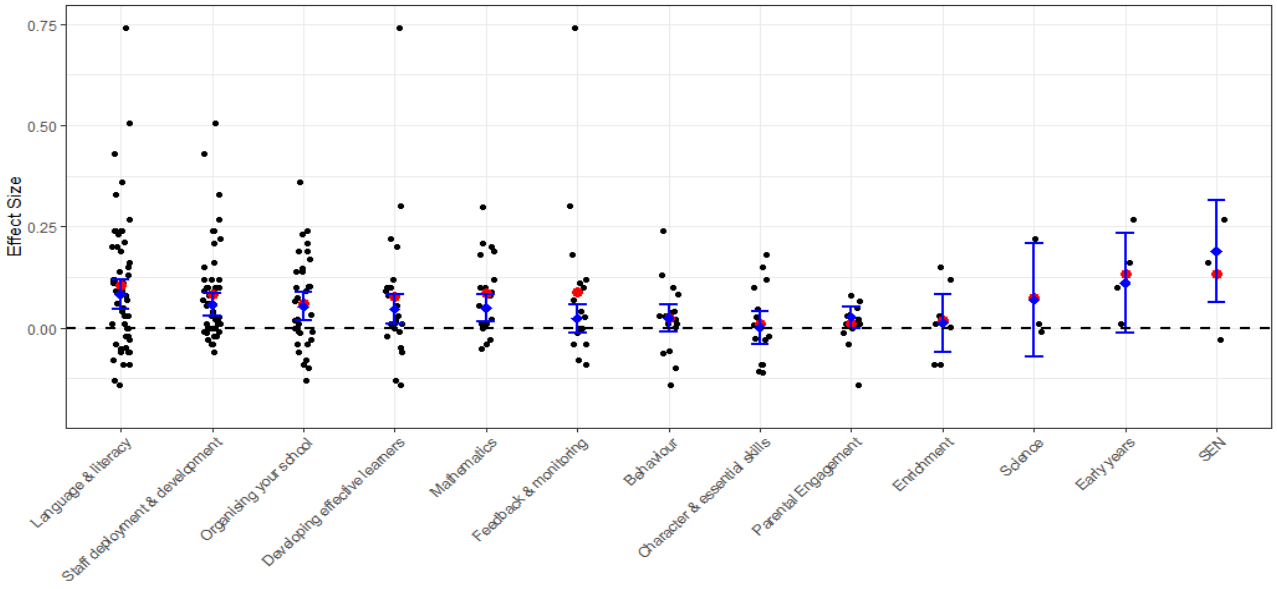
Across the 11 EEF intervention themes with sufficient data, the weighted mean effect size ranged between 0.00 (character and essential skills) and +0.11 (early years).

Table 43: Effect size by EEF intervention themes classification primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Language and literacy	53	+0.09	0.019	+0.05	+0.12
Staff deployment and development	46	+0.06	0.014	+0.03	+0.09
Organising your school	33	+0.05	0.018	+0.02	+0.09
Developing effective learners	23	+0.05	0.018	+0.01	+0.08
Mathematics	18	+0.05	0.016	+0.02	+0.06
Feedback and monitoring pupil progress	16	+0.02	0.017	−0.01	+0.06
Behaviour	16	+0.02	0.017	−0.01	+0.06
Character and essential skills	15	0.00	0.021	−0.04	+0.04
Parental engagement	14	+0.03	0.014	0.00	+0.05
Enrichment	7	+0.01	0.037	−0.06	+0.08
Science	3	–	–	–	–
Early years	4	+0.11	0.063	−0.01	+0.23
Special educational needs and disabilities	3	–	–	–	–

Overall weighted mean = +0.04 SD.

Figure 30: Primary ITT effect size by EEF intervention theme



Note: The EEF intervention themes are not mutually exclusive and so a single effect size can appear in multiple intervention themes.

Secondary ITT

Across the eight EEF intervention themes with sufficient data, the weighted mean secondary ITT effect size ranged between −0.07 (feedback and monitoring, 18 secondary attainment effect sizes; behaviour, 8 effect sizes) and +0.07 (parental engagement, 6 effect sizes).

Table 44: Effect size by EEF intervention themes classification secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Language and literacy	31	−0.01	0.019	−0.05	+0.03
Staff deployment and development	32	+0.01	0.011	−0.01	+0.03
Organising your school	21	+0.06	0.026	0.00	+0.11
Developing effective learners	12	−0.02	0.046	−0.11	+0.07
Mathematics	14	+0.04	0.046	−0.02	+0.10
Feedback and monitoring pupil progress	18	−0.07	0.029	−0.02	−0.01
Behaviour	8	−0.07	0.036	−0.14	0.00
Character and essential skills	2	–	–	–	–
Parental engagement	6	+0.07	0.052	−0.03	+0.18
Enrichment	2	–	–	–	–
Science	2	–	–	–	–
Early years	1	–	–	–	–
Special educational needs and disabilities	1	–	–	–	–

Overall weighted mean = +0.01 SD.

FSM

Across the 11 EEF intervention themes with sufficient data, the weighted mean FSM effect size ranged between −0.06 (enrichment, 10 FSM effect sizes) and +0.07 (developing effective learners, 21 effect sizes).

Table 45: Effect size by EEF intervention themes classification FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Language and literacy	68	+0.05	0.022	0.00	+0.09
Staff deployment and development	52	+0.04	0.012	+0.01	+0.06
Organising your school	32	+0.04	0.023	−0.01	+0.09
Developing effective learners	21	+0.07	0.032	+0.01	+0.14
Mathematics	18	+0.05	0.025	0.00	+0.10
Feedback and monitoring pupil progress	26	−0.02	0.020	−0.06	+0.02
Behaviour	17	+0.04	0.023	−0.01	+0.08
Character and essential skills	16	0.00	0.031	−0.07	+0.06
Parental engagement	16	−0.01	0.015	−0.04	+0.02
Enrichment	10	−0.06	0.047	−0.15	+0.03
Science	3	–	–	–	–
Early years	5	+0.01	0.023	−0.04	+0.05
Special educational needs and disabilities	6	−0.15	0.203	−0.55	+0.25

Overall weighted mean = +0.03 SD.

EEF promising interventions

Primary ITT

Interventions classed as promising by EEF had a statistically significantly ($p < 0.01$) larger effect size (weighted mean = +0.12 SD; 95% CI: +0.09 to +0.15) compared with interventions not classed as promising (weighted mean = +0.01 SD; 95% CI: 0.00 to +0.03). Whilst perhaps unsurprising, these findings serve to quantify the distinction between interventions that are or are not classed as promising by EEF.

Table 46: Effect size by EEF promising intervention classification primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Promising	30	+0.12	0.016	+0.09	+0.15
Other	103	+0.01	0.007	0.00	+0.03

Meta *p*-value < 0.01***. Overall weighted mean = +0.04 SD.

Secondary ITT

Interventions classed as promising by EEF had a statistically significantly ($p < 0.01$) larger secondary ITT effect size (weighted mean = +0.08 SD; 95% CI: +0.05 to +0.11) compared with other interventions not classed as promising (weighted mean = −0.01 SD; 95% CI: −0.03 to +0.01).

Table 47: Effect size by EEF promising intervention classification secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Promising	16	+0.08	0.016	+0.05	+0.11
Other	62	−0.01	0.010	−0.03	+0.01

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

FSM

Interventions classed as promising by EEF had a statistically significantly ($p < 0.01$) larger FSM effect size (weighted mean = +0.11 SD; 95% CI: +0.06 to +0.15) compared with other interventions not classed as promising (weighted mean = +0.01 SD; 95% CI: −0.01 to +0.03).

Table 48: Effect size by EEF promising intervention FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Promising	35	+0.11	0.025	+0.06	+0.15
Other	114	+0.01	0.009	−0.01	+0.03

Meta *p*-value < 0.01***. Overall weighted mean = +0.03 SD.

Effect sizes and theory & evidence

Summary

Table 49: Summary of meta-analyses of ITT effect sizes and theory & evidence

Subtheme	Explanatory variable	Primary ITT	Secondary ITT	FSM
Empirical evidence	Strength of prior evidence of impact	✓*	✓***	✓***
Theory	Level of theoretical detail	✓#	✓***	✓#
Causal process	Focus of change (learning, teacher or wider outcomes)	✓***	n/a	✓***

Most of the explanatory variables in the theory & evidence theme were included in the meta-analyses of secondary ITT and FSM attainment outcomes. There was one exception for secondary ITT effect size, where there were insufficient cases ($n < 4$) to allow the analyses examining association with causal processes (focus of change). Essentially the learning focus was so common, this variable does not discriminate well – this was also the case for primary and secondary ITT effect size but not to the extent of precluding analyses (in both, 80% of effect sizes related to interventions with a learning focus).

Empirical evidence

Primary ITT

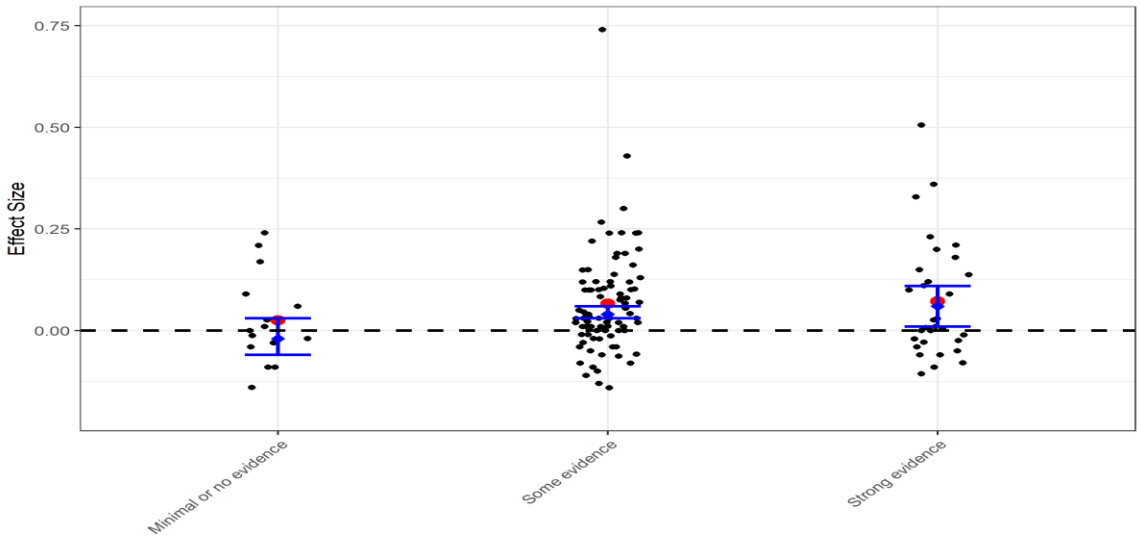
On average, evaluations that drew on strong empirical evidence were associated with a higher effect size (weighted mean = +0.06 SD; 95% CI: +0.01 to +0.11) compared with evaluations with some empirical evidence (weighted mean = +0.04 SD; 95% CI: +0.03 to +0.06) or evaluations with minimal/no empirical evidence (weighted mean = −0.02 SD; 95% CI: −0.06 to +0.03). The association between strength of empirical evidence and effect size was statistically significant at the 10% level ($p < 0.10$). This finding is aligned with evidence in the wider 'theory-based evaluation' field (for example, Weiss, 1995).

Table 50: Effect size by strength of prior evidence of impact primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Minimal or no evidence	15	−0.02	0.022	−0.06	+0.03
Some evidence	87	+0.05	0.008	+0.03	+0.06
Strong evidence	31	+0.06	0.024	+0.01	+0.11

Meta *p*-value <0.10*. Overall weighted mean = +0.04 SD.

Figure 31: Effect size by strength of prior evidence of impact (primary ITT attainment outcomes)



Secondary ITT

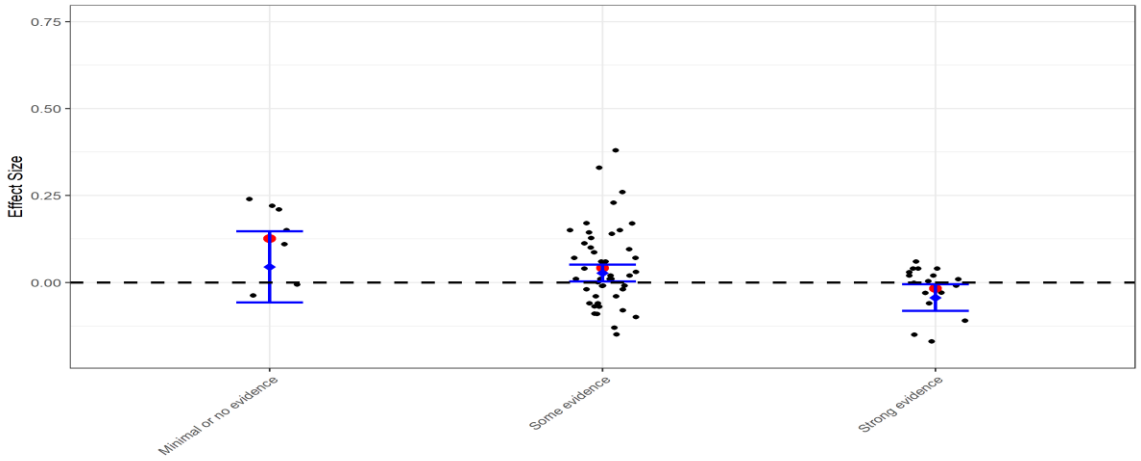
On average, evaluations that drew on strong empirical evidence were associated with a lower secondary ITT effect size (weighted mean = −0.04 SD; 95% CI: −0.08 to −0.01) compared with evaluations with some empirical evidence (weighted mean = +0.03 SD; 95% CI: 0.00 to +0.05) or evaluations with minimal/no empirical evidence (weighted mean = +0.04 SD; 95% CI: −0.06 to +0.15). The association between strength of empirical evidence and effect size was statistically significant at the 5% level ($p < 0.05$). It is unclear why this should be the case but, as noted above, secondary ITT attainment effect sizes were reported by less than half (35) of the 82 trials included in the review. Therefore, greater caution is needed when drawing conclusions from statistically significant findings and making comparisons with other effect size groupings.

Table 51: Effect size by strength of prior evidence of impact secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Minimal or no evidence	7	+0.04	0.052	−0.06	+0.15
Some evidence	49	+0.03	0.012	0.00	+0.05
Strong evidence	22	−0.04	0.020	−0.08	−0.01

Meta *p*-value <0.01***. Overall weighted mean = +0.01 SD.

Figure 32: Effect size by strength of prior evidence of impact secondary ITT attainment outcomes



FSM

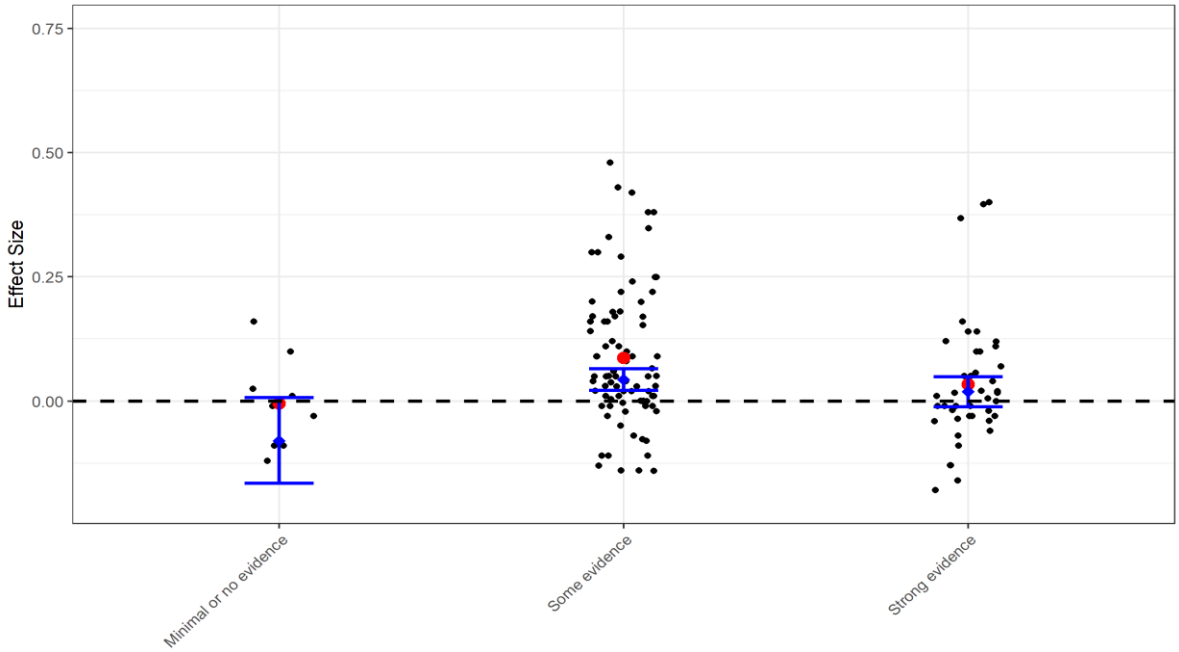
On average, evaluations that drew on some empirical evidence were associated with a higher FSM effect size (weighted mean = +0.04 SD; 95% CI: +0.02 to +0.07) compared to evaluations with strong empirical evidence (weighted mean = +0.02 SD; 95% CI: −0.01 to +0.05) or evaluations with minimal/no empirical evidence (weighted mean = −0.08 SD; 95% CI: −0.17 to +0.01). The association between strength of empirical evidence and effect size was statistically significant at the 1% level ($p < 0.01$). The lower effect size for evaluations with minimal/no empirical evidence compared with others aligns with Weiss et al. (1995) and other related literature as indicated above.

Table 52: Effect size by strength of prior evidence of impact FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Minimal or no evidence	13	−0.08	0.044	−0.17	+0.01
Some evidence	91	+0.04	0.011	+0.02	+0.07
Strong evidence	45	+0.02	0.015	−0.01	+0.05

Meta *p*-value <0.01***. Overall weighted mean = +0.03 SD

Figure 33: Effect size by strength of prior evidence of impact FSM attainment outcomes



Theoretical detail

Primary ITT

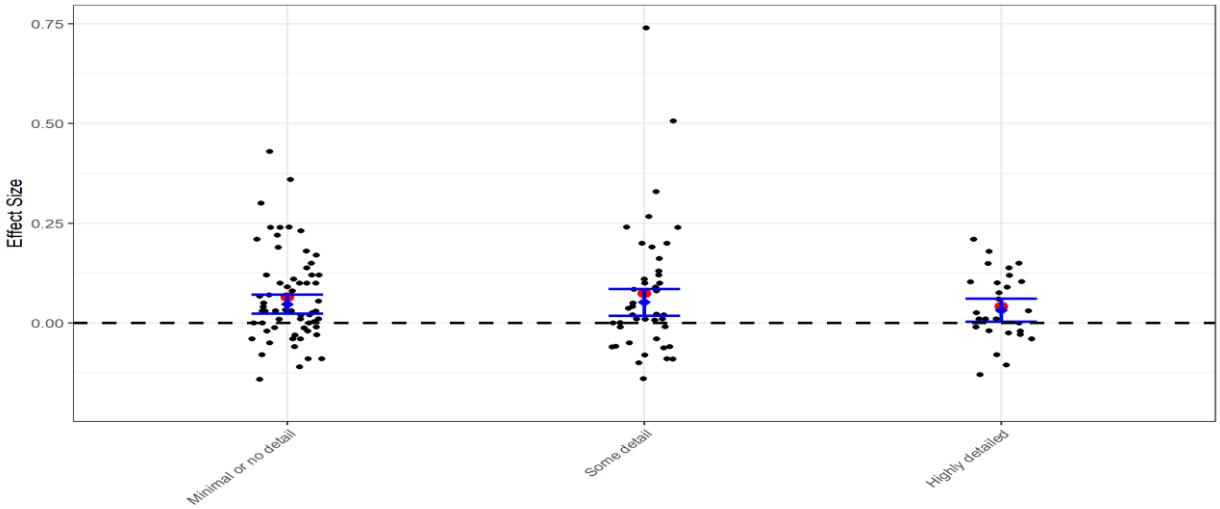
No evidence of an association between effect size and theoretical detail was observed. Across the categories of the theory variable, the weighted mean effect size ranged between +0.03 and +0.05 SD. It is important to note that the level of theoretical detail presented in the trial reports is dependent on evaluators' engagement with the underpinning theory and so should not be taken as a proxy for the strength of the theory. Therefore, the analyses for all outcomes should be treated with caution.

Table 53: Effect size by theoretical detail primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Minimal or no detail	62	+0.05	0.012	+0.02	+0.07
Some detail	44	+0.05	0.017	+0.02	+0.08
Highly detailed	27	+0.03	0.015	0.00	+0.06

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Figure 34: Effect size by theoretical detail primary ITT attainment outcomes



Secondary ITT

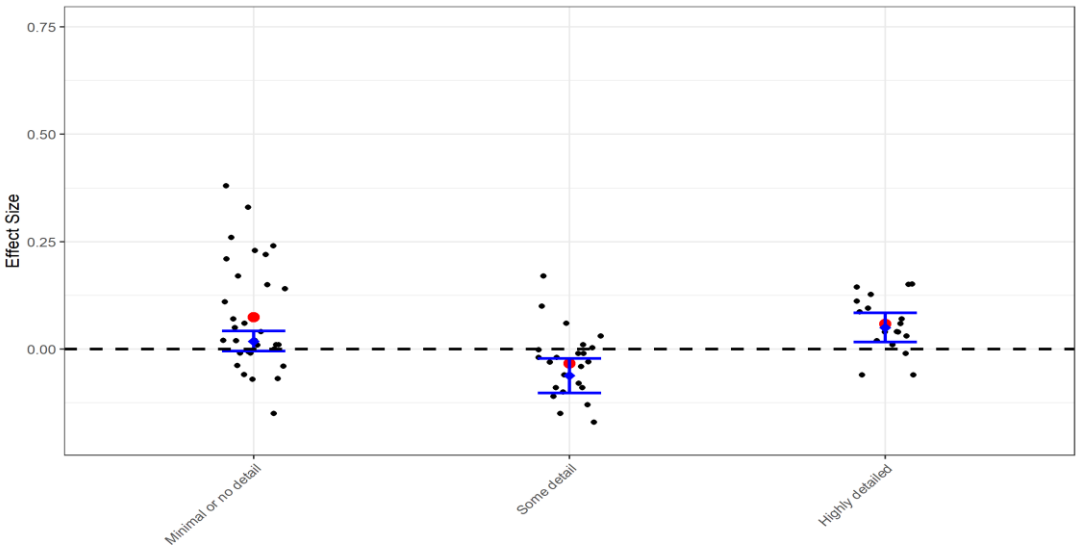
On average, evaluations where the theory of change was highly detailed were associated with a higher secondary ITT effect size (weighted mean = +0.05 SD; 95% CI: +0.02 to +0.08) compared with evaluations with some theoretical detail (weighted mean = −0.06 SD; 95% CI: −0.10 to −0.02) or evaluations with minimal/no theoretical detail (weighted mean = +0.02 SD; 95% CI: −0.01; +0.04). The association between theoretical detail and secondary ITT effect size was statistically significant at the 5% level ($p < 0.05$).

Table 54: Effect size by theoretical detail secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Minimal or no detail	31	+0.02	0.012	−0.01	+0.04
Some detail	29	−0.06	0.020	−0.10	−0.02
Highly detailed	18	+0.05	0.017	+0.02	+0.08

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Figure 35: Effect size by theoretical detail secondary ITT attainment outcomes



FSM

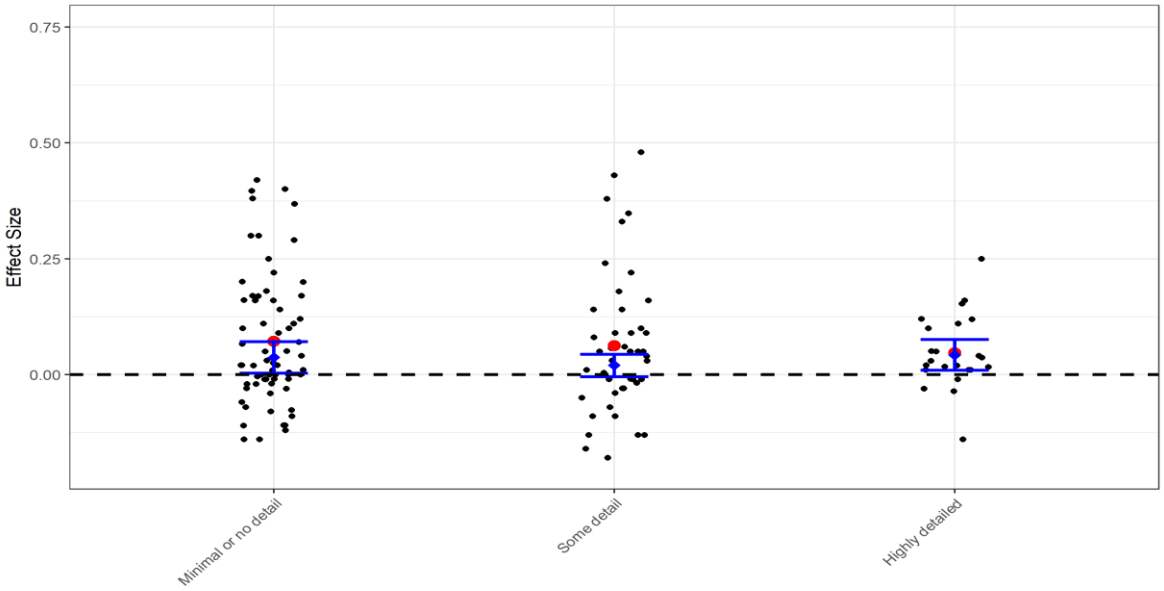
No evidence of an association between FSM effect size and theoretical detail was observed. Across the categories of the theory variable, the weighted mean effect size ranged between +0.02 and +0.04 SD.

Table 55: Effect size by theoretical detail FSM attainment outcomes

Factor	*n*	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Minimal or no detail	72	+0.04	0.018	0.00	+0.07
Some detail	51	+0.02	0.012	−0.01	+0.04
Highly detailed	26	+0.04	0.017	0.00	+0.07

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 36: Effect size by theoretical detail FSM attainment outcomes


0.75
0.50
0.25
0.00
Effect Size
Minimal or no detail
Some detail
Highly detailed

Causal processes

Primary ITT

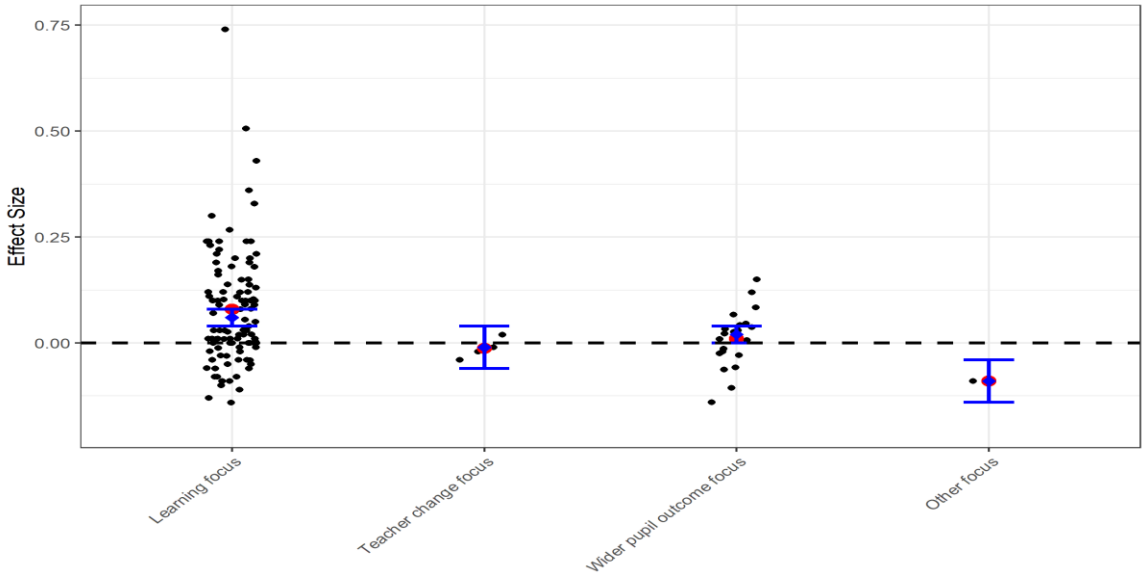
The vast majority of effect sizes in the review were from evaluations of interventions with a learning focus (106 effect sizes across 69 evaluations) and, on average, these are associated with a statistically significantly ($p < 0.01$) higher effect size (weighted mean = +0.06 SD; 95% CI: +0.04 to +0.08) compared with the nine evaluations of interventions (reporting 21 effect sizes) with a focus on wider pupil outcomes (weighted mean = +0.02 SD; 95% CI: 0.00 to +0.04). A closer relationship between higher effect sizes and programmes with a learning focus as opposed to other areas of change was also found by Slavin (2016).

Table 56: Effect size by main focus of change primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Learning focus	106	+0.06	0.012	+0.04	+0.08
Teacher change focus	4	−0.01	0.025	−0.06	+0.04
Wider pupil outcome focus	21	+0.02	0.010	−0.01	+0.04
Other focus	2	–	–	–	–

Meta *p*-value < 0.01***. Overall weighted mean = +0.04 SD.
Note: Analyses of secondary ITT effect size by main focus of change was not possible because only a single category (learning focus) had sufficient cases (of four or more).

Figure 37: Effect size by main focus of change primary ITT attainment outcomes



Note: Analyses of secondary ITT effect size by main focus of change was not possible because only a single category (learning focus) had sufficient cases (of four or more).

Secondary ITT

The number of secondary ITT effect sizes for interventions with a teacher change or wider pupil outcome focus was too small to enable analyses (both, $n = 2$). The vast majority of secondary ITT effect sizes are for interventions with a learning focus (74 effect sizes; weighted mean = +0.01; 95% CI: −0.02; +0.03).

FSM

The vast majority of FSM effect sizes in the review were from evaluations of interventions with a learning focus (119 effect sizes) and, on average, these are associated with a statistically significantly ($p < 0.01$) higher effect size (weighted

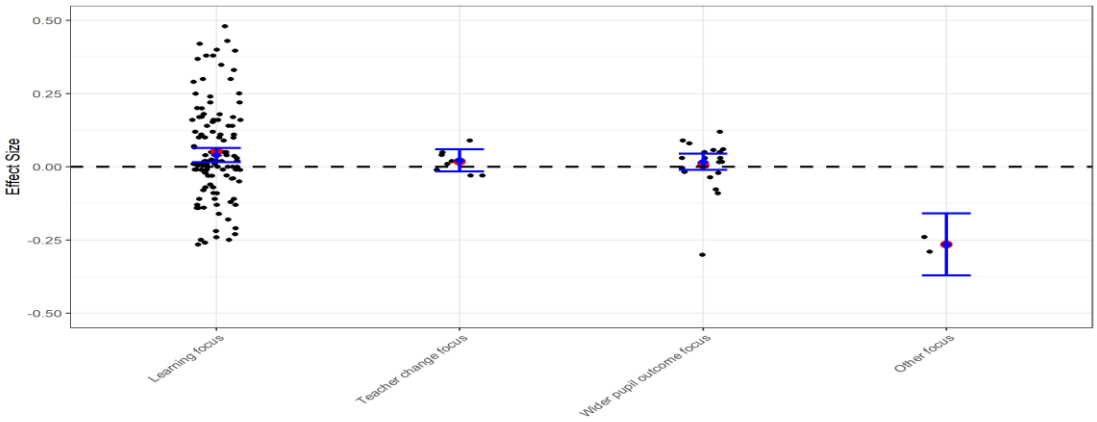
mean = +0.04 SD; 95% CI: +0.02 to +0.06) compared with interventions with a focus on teacher change or wider pupil outcomes (both weighted mean = +0.02 SD). Again, this aligns with Slavin's (2016) earlier work.

Table 57: Effect size by main focus of change FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Learning focus	119	+0.04	0.017	+0.01	+0.08
Teacher change focus	8	+0.02	0.019	−0.02	+0.06
Wider pupil outcome focus	20	+0.02	0.014	−0.01	+0.05
Other focus	2	–	–	–	–

Meta *p*-value < 0.01***. Overall weighted mean = +0.03 SD.

Figure 38: Effect size by main focus of change FSM attainment outcomes



Effect sizes and context

Summary

Table 58: Summary of meta-analyses of ITT effect size and context

Subtheme	Explanatory variable	Primary ITT	Secondary ITT	FSM
External environment	Geography	✓*	✓	✓#
	Perceptions on Ofsted	✓#	✓	✓
Characteristics of participating organisations	Specialist facilities and space	✓***	✓***	✓**
	Staff time and availability	✓	✓	✓
	Workforce capacity)	✓	✓***	✓
	Alignment of intervention and current practice	✓*	✓	✓
	Staff teamwork	✓*	✓***	✓#
Characteristics of participating individuals	Pupil behaviour	✓	✓***	✓*
	SLT buy in	✓	✓	✓
	Staff expectations and motivations	✓	✓***	✓#

All of the explanatory variables included in the context theme were included in the meta-analyses of secondary ITT and FSM attainment outcomes.

Geography

Primary ITT

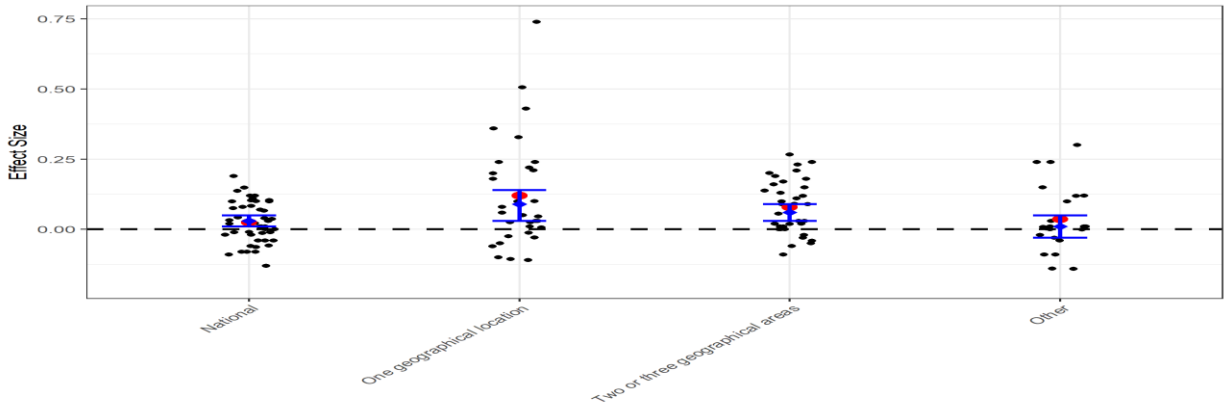
On average, trials located in one geographical area are associated with higher effect sizes (weighted mean = +0.09 SD; 95% CI: +0.03 to +0.14) compared with trials that cover a wider geographical area (weighted mean = +0.06 SD or lower). The meta-analyses found the association between geographical context and effect size to be statistically significant at the 10% level ($p < 0.10$). Previous reviews did not consider this relationship. It is possible that the difference relates to greater ease of consistent implementation in smaller geographical areas (aligned somewhat with the finding of Anders et al. (2017) that implementation was easier in school groups than standalone schools).

Table 59: Effect size by geography primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
National	45	+0.03	0.009	+0.01	+0.05
One geographical location	31	+0.09	0.028	+0.03	+0.14
Two or three geographical areas	35	+0.06	0.015	+0.04	+0.09
Other	22	+0.02	0.021	−0.02	+0.06

Meta *p*-value <0.10*. Overall weighted mean = +0.04 SD.

Figure 39: Effect size by geography primary ITT attainment outcomes



Secondary ITT

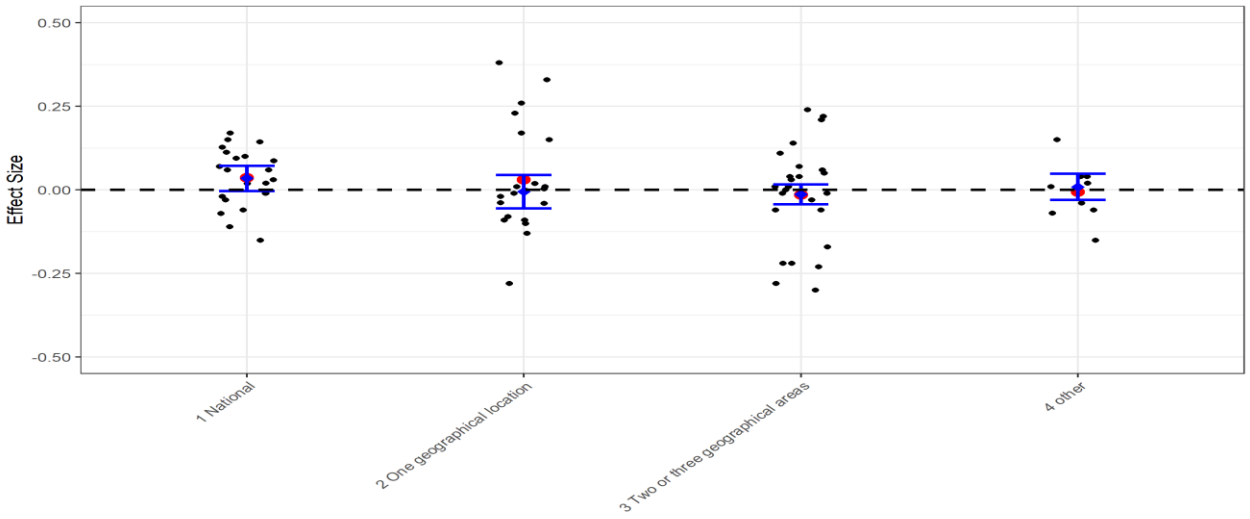
No evidence of an association between geographical context and secondary ITT effect size was observed. Across categories, mean effect sizes ranged between −0.01 and +0.04 SD.

Table 60: Effect size by geography secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
National	22	+0.04	0.019	0.00	+0.07
One geographical location	22	−0.01	0.025	−0.06	+0.05
Two or three geographical areas	25	−0.01	0.015	−0.04	+0.02
Other	9	+0.01	0.020	−0.03	+0.05

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.01 SD.

Figure 40: Effect size by geography secondary ITT attainment outcomes



FSM

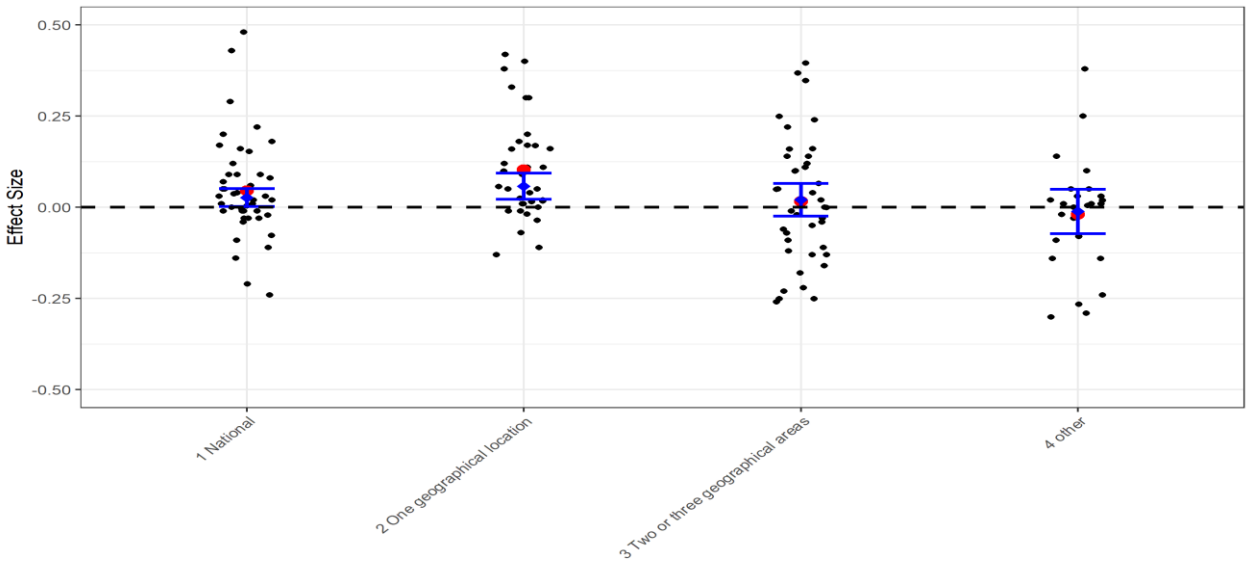
No evidence of an association between geographical context and FSM effect size was observed. Across categories, mean effect sizes ranged between −0.01 and +0.06 SD with the highest weighted mean effect size observed for trials located in one geographical area.

Table 61: Effect size by geography FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
National	46	+0.03	0.012	0.00	+0.05
One geographical location	36	+0.06	0.018	+0.02	+0.09
Two or three geographical areas	42	+0.02	0.023	−0.02	+0.07
Other	25	−0.01	0.031	−0.07	+0.05

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 41: Effect size by geography FSM attainment outcomes



Perceptions of Ofsted

Primary ITT

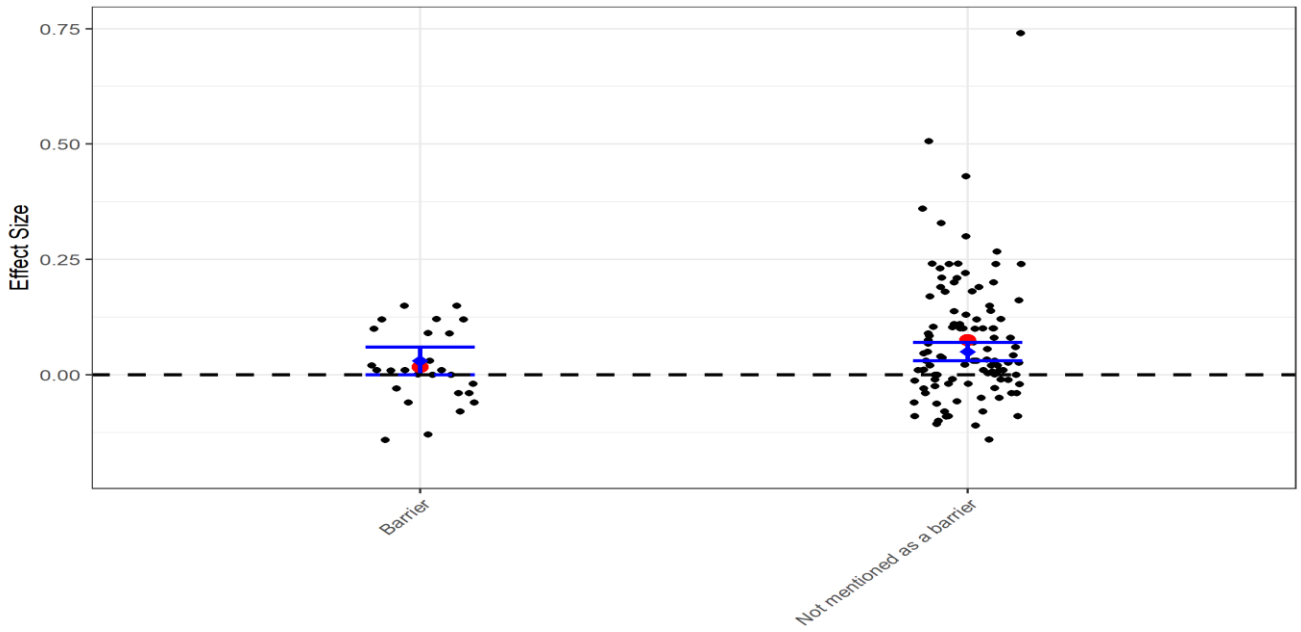
Turning to perceptions of Ofsted as an aspect of the external environment, on average, evaluations reporting that Ofsted was perceived as a barrier for the intervention were associated with slightly lower effect sizes (weighted mean = +0.03 SD) compared with evaluations that did not mention Ofsted as a barrier (weighted mean = +0.05 SD), although this difference was not statistically significant.

Table 62: Effect size by perceptions on Ofsted primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	26	+0.03	0.016	0.00	+0.06
Not mentioned as a barrier	107	+0.05	0.009	+0.03	+0.07

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Figure 42: Effect size by perceptions on Ofsted primary ITT attainment outcomes



Secondary ITT

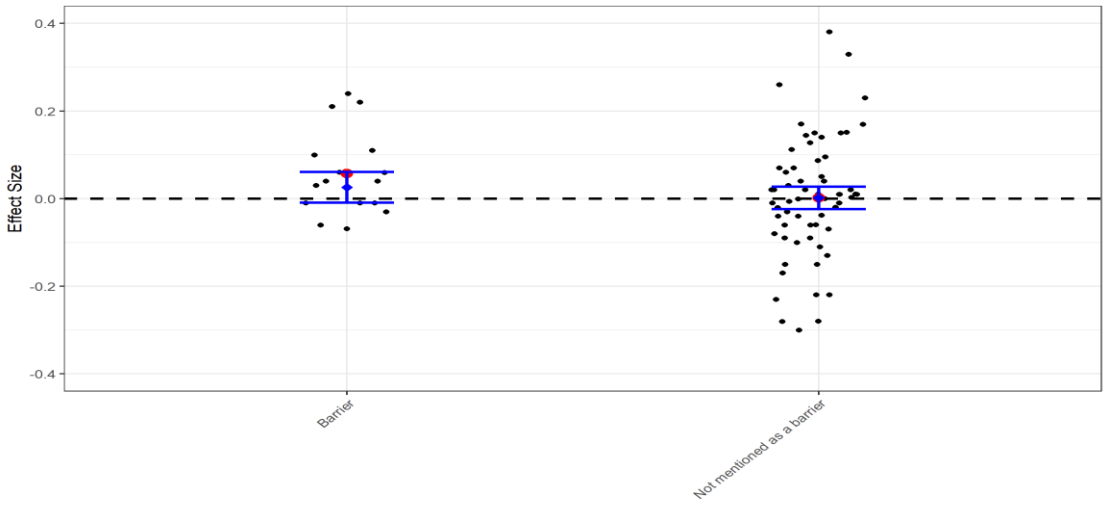
No evidence of an association between perceptions of Ofsted and secondary ITT effect size was observed.

Table 63: Effect size by perceptions on Ofsted secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	16	+0.03	0.018	−0.01	+0.06
Not mentioned as a barrier	62	0.00	0.013	−0.02	+0.03

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.01 SD.

Figure 43: Effect size by perceptions on Ofsted secondary ITT attainment outcomes



FSM

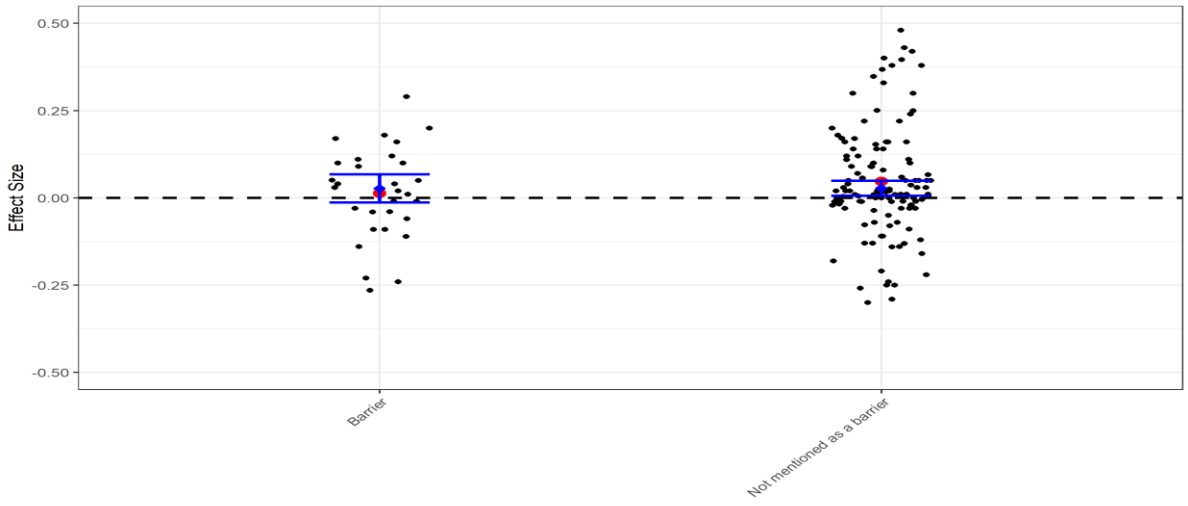
No evidence of an association between perceptions of Ofsted and FSM effect size was observed.

Table 64: Effect size by perceptions on Ofsted FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	30	+0.03	0.020	−0.01	+0.07
Not mentioned as a barrier	119	+0.03	0.011	+0.01	+0.05

Meta *p*-value >0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 44: Effect size by perceptions on Ofsted FSM attainment outcomes



Characteristics of participating organisations

Specialist facilities and space

Primary ITT

The meta-analyses found the association between specialist facilities and space and effect size to be statistically significant at the 5% level ($p < 0.05$). Perhaps contradictory to what would be expected, reports that did **not** mention

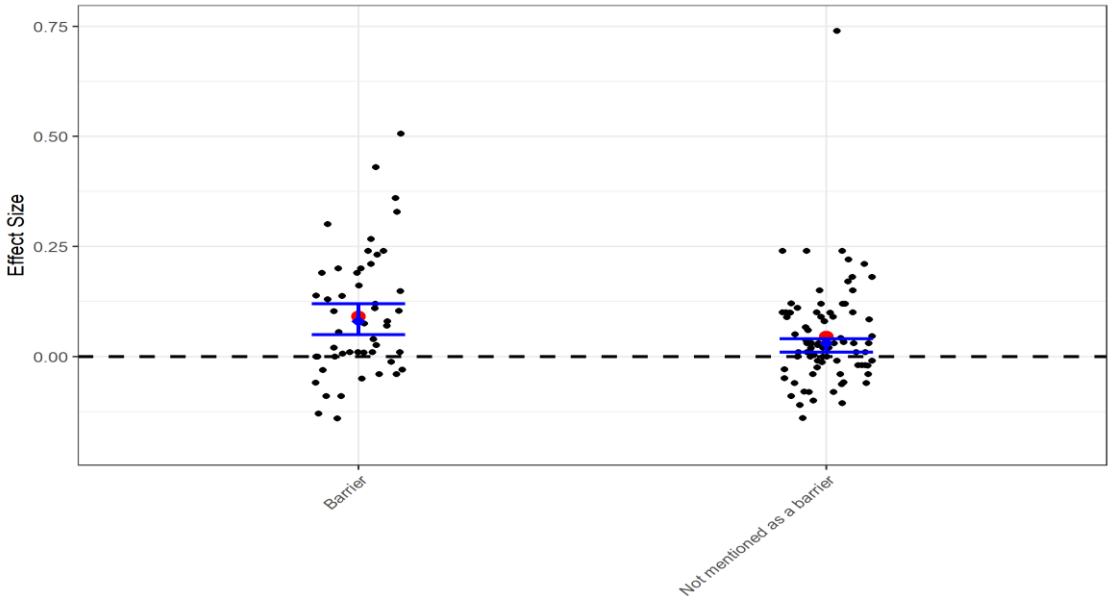
specialist facilities and space to be a barrier had a lower mean effect size (0.03) compared with reports that did mention this as a barrier (0.08). This finding should be treated with caution due to issues with coding this variable (please see Appendix A).

Table 65: Effect size by specialist facilities and space primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	51	+0.08	0.018	+0.05	+0.12
Not mentioned as a barrier	82	+0.03	0.008	+0.01	+0.04

Meta *p*-value < 0.05**. Overall weighted mean = +0.04 SD.

Figure 45: Effect size by specialist facilities and space primary ITT attainment outcomes



Secondary ITT

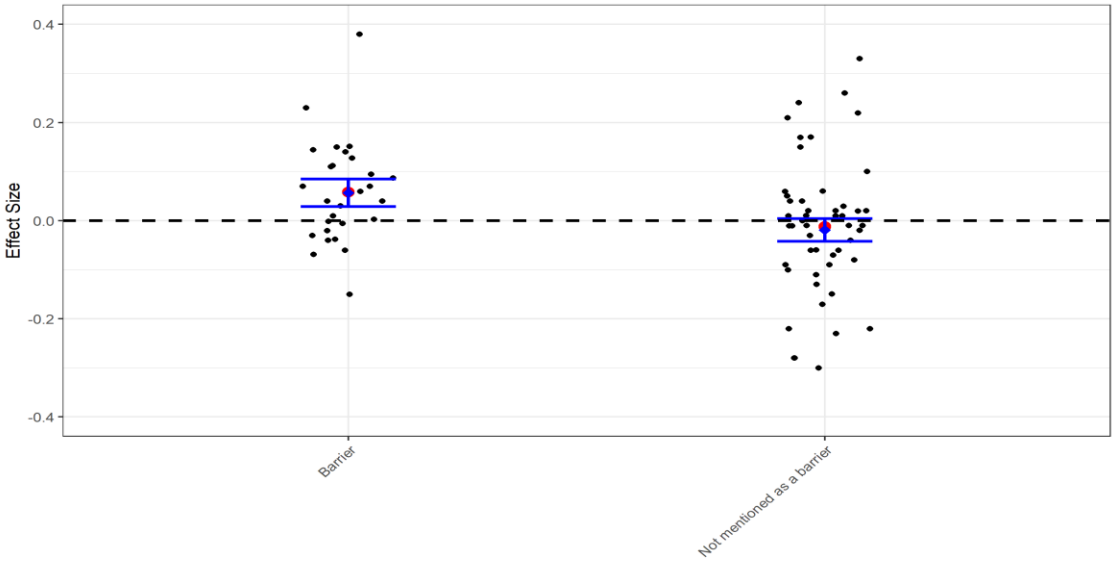
The meta-analyses found the association between specialist facilities and space and FSM effect size to be statistically significant at the 1% level ($p < 0.01$). Echoing what was observed with primary ITT effect sizes (and subject to the same caution), reports that did **not** mention specialist facilities and space to be a barrier had a lower mean effect size (−0.02 SD) compared with reports that did mention this as a barrier (+0.06 SD).

Table 66: Effect size by specialist facilities and space secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	28	+0.06	0.014	+0.03	+0.09
Not mentioned as a barrier	50	−0.02	0.012	−0.04	0.00

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Figure 46: Effect size by specialist facilities and space secondary ITT attainment outcomes



FSM

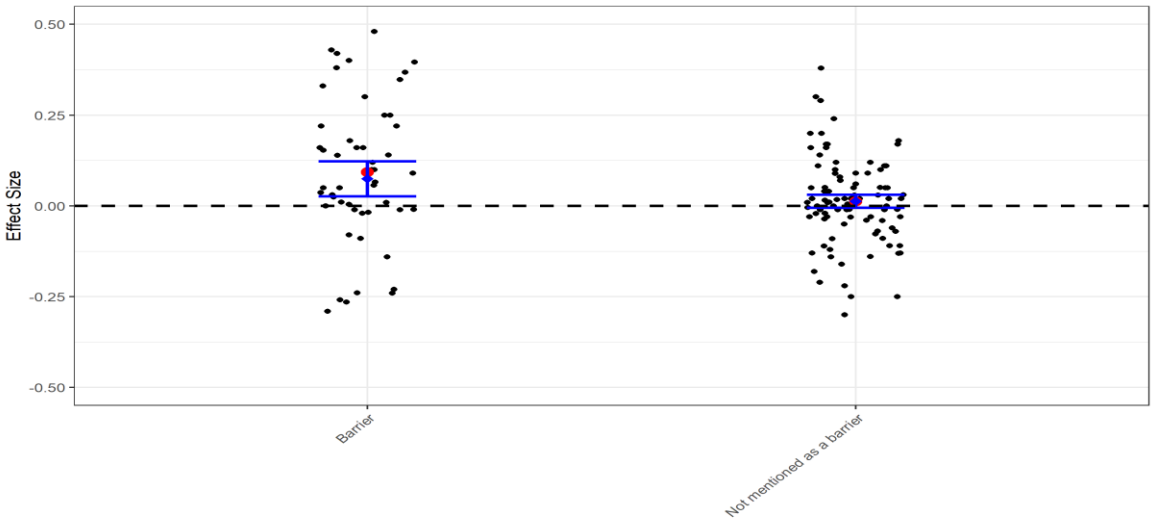
The meta-analyses found the association between specialist facilities and space and FSM effect size to be statistically significant at the 5% level ($p < 0.05$). Echoing what was observed with primary ITT effect sizes (and, again, subject to caution), reports that did **not** mention specialist facilities and space to be a barrier had a lower mean effect size (+0.01 SD) compared with reports that did mention this as a barrier (+0.08 SD).

Table 67: Effect size by specialist facilities and space FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	53	+0.08	0.025	+0.03	+0.12
Not mentioned as a barrier	96	+0.01	0.009	−0.01	+0.03

Meta *p*-value < 0.05**. Overall weighted mean = +0.03 SD.

Figure 47: Effect size by specialist facilities and space FSM attainment outcomes



Staff time and availability

No significant association between effect size and staff time and availability was observed. This was found for primary ITT, secondary ITT and FSM effect sizes.

Workforce capacity

Primary ITT

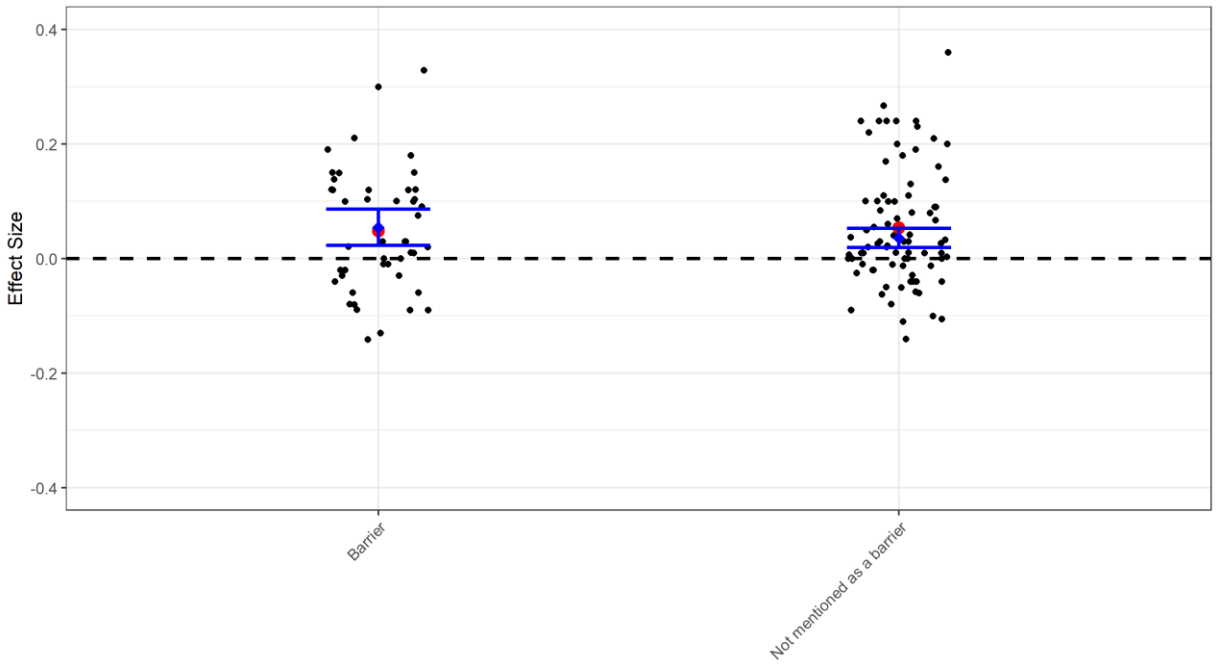
No significant association between effect size and workforce capacity was observed.

Table 68: Effect size by workforce capacity primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	47	+0.05	0.016	+0.02	+0.09
Not mentioned as a barrier	86	+0.04	0.009	+0.02	+0.05

Meta *p*-value > 0.10(NS). Overall weighted mean = +0.04 SD.

Figure 48: Effect size by workforce capacity primary ITT attainment outcomes



Secondary ITT

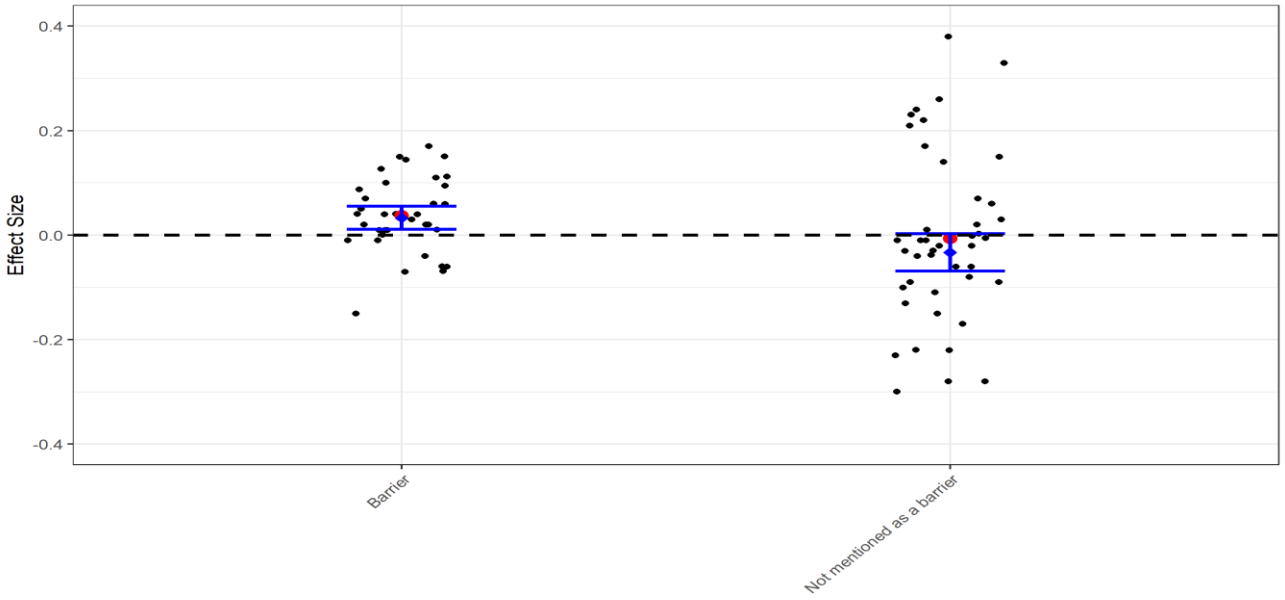
The meta-analyses found the association between perceptions on workforce capacity and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). Reports that mentioned workforce capacity to be a barrier had a higher mean effect size (+0.03 SD) compared with reports that did not mention this as a barrier (−0.03 SD). There is no clear reason for this finding, given it was not found for the primary ITT.

Table 69: Effect size by workforce capacity secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	35	+0.03	0.011	+0.01	+0.06
Not mentioned as a barrier	43	−0.03	0.018	−0.07	+0.00

Meta *p*-value <0.01***. Overall weighted mean = +0.01 SD.

Figure 49: Effect size by workforce capacity secondary ITT attainment outcomes



FSM

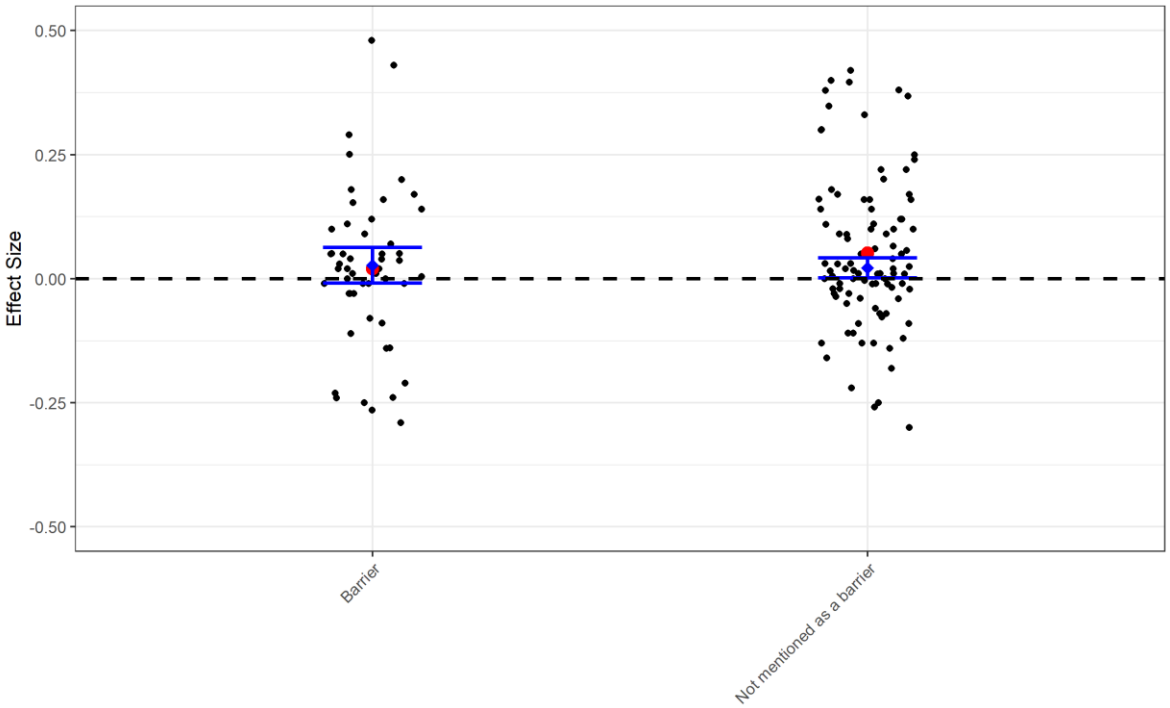
No significant association between FSM effect size and workforce capacity was observed.

Table 70: Effect size by workforce capacity FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	51	+0.03	0.018	−0.01	+0.06
Not mentioned as a barrier	98	+0.02	0.010	0.00	+0.04

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 50: Effect size by workforce capacity FSM attainment outcomes



Alignment between intervention and existing practice

Primary ITT

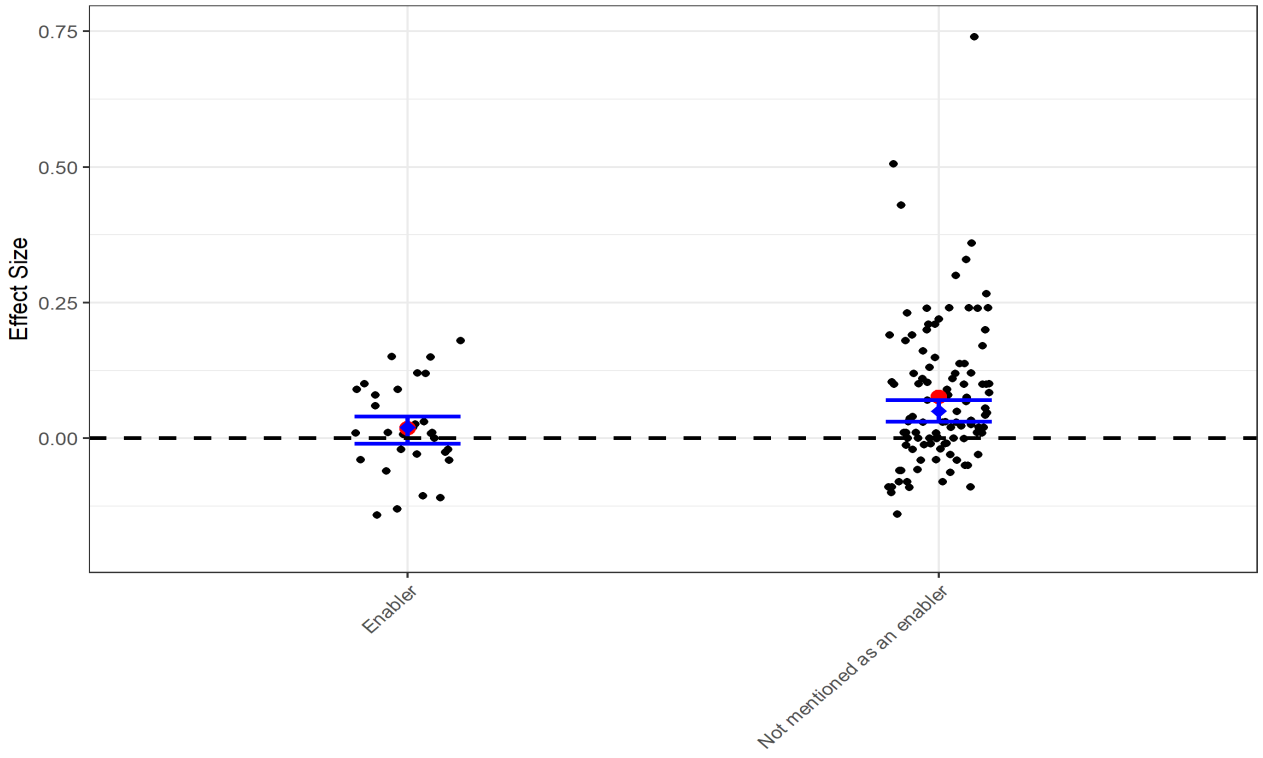
Evaluations that mentioned the alignment of the intervention and existing practice as an enabler were associated with lower average effect sizes (weighted mean = +0.02 SD; 95% CI: −0.01 to +0.04) compared with evaluations that did not mention such alignment as an enabling factor (weighted mean = +0.05 SD; 95% CI: +0.03 to +0.07). The meta-analyses found the association between alignment with existing practice and effect size to be statistically significant at the 10% level ($p < 0.10$). Although the implementation process is likely to be easier when the new intervention is more closely aligned, this finding suggests that the primary outcome effect sizes may be higher when the intervention is more of a departure from existing practice. While caution is needed, as this variable is based on perceptual data and is not consistently reported across all evaluations, it does highlight the importance of recognising that conditions that support effective implementation do not necessarily also lead to the strongest effect on pupil outcomes.

Table 71: Effect size by alignment between intervention and existing practice primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Enabler	31	+0.02	0.014	−0.01	+0.05
Not mentioned as an enabler	103	+0.05	0.010	+0.03	+0.07

Meta *p*-value < 0.10*. Overall weighted mean = +0.04 SD.

Figure 51: Effect size by alignment between intervention and existing practice primary ITT attainment outcomes



Secondary ITT

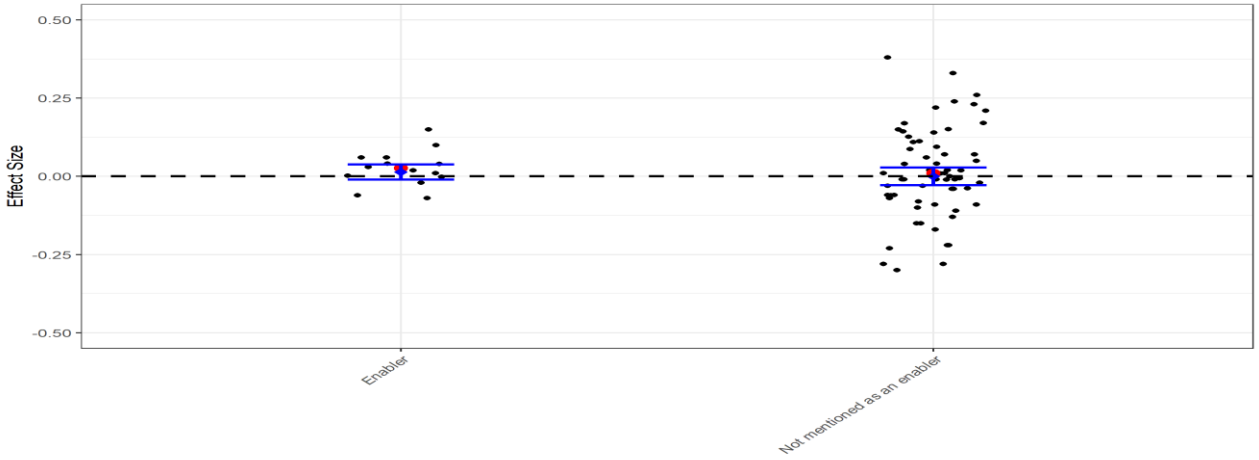
No evidence of an association between perceived alignment of the intervention and existing practice and secondary ITT effect size was observed.

Table 72: Effect size by alignment between intervention and existing practice secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Enabler	15	+0.01	0.012	−0.01	+0.04
Not mentioned as an enabler	63	0.00	0.014	−0.03	+0.03

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.01 SD.

Figure 52: Effect size by alignment between intervention and existing practice secondary ITT attainment outcomes



FSM

No evidence of an association between perceived alignment of the intervention and existing practice with FSM effect size was observed.

Table 73: Effect size by alignment between intervention and existing practice FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Enabler	35	+0.03	0.014	0.00	+0.06
Not mentioned as an enabler	114	+0.03	0.013	0.00	+0.05

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 53: Effect size by alignment between intervention and existing practice FSM attainment outcomes

Staff teamwork

Primary ITT

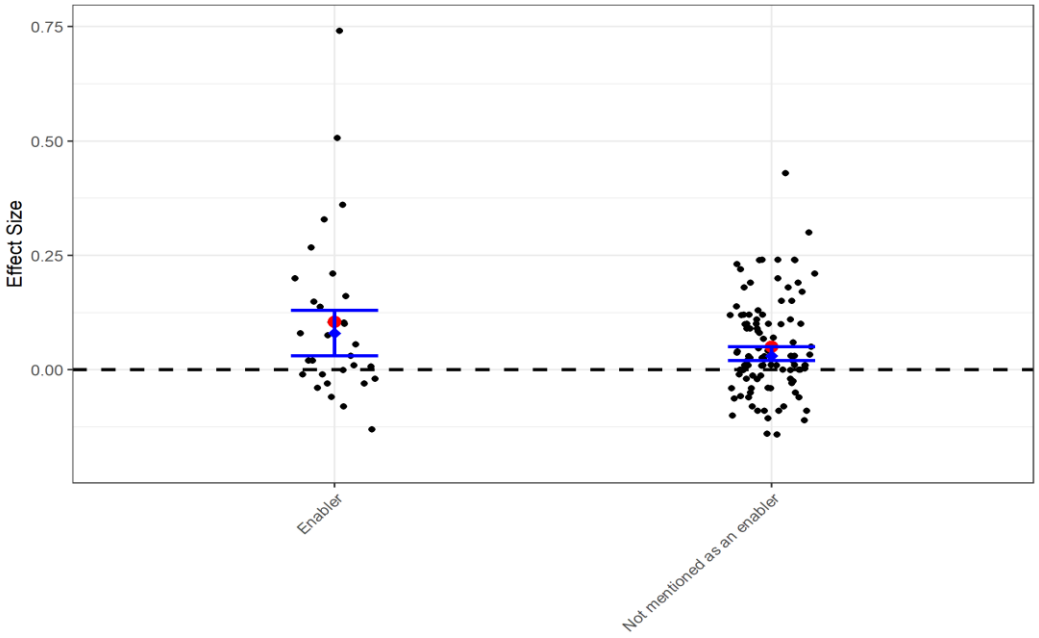
On average, evaluations that mentioned staff teamwork as an enabler were associated with higher effect sizes (weighted mean = +0.08 SD; 95% CI: +0.03 to +0.13) compared with evaluations that did not cite staff teamwork as an enabler (weighted mean = +0.03 SD; 95% CI: +0.02 to +0.05). The meta-analyses found the association between staff teamwork and effect size to be statistically significant at the 10% level ($p < 0.10$).

Table 74: Effect size by staff teamwork primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Enabler	31	+0.08	0.024	+0.03	+0.13
Not mentioned as an enabler	102	+0.03	0.008	+0.02	+0.05

Meta *p*-value < 0.10*. Overall weighted mean = +0.04 SD.

Figure 54: Effect size by staff teamwork primary ITT attainment outcomes



Secondary ITT

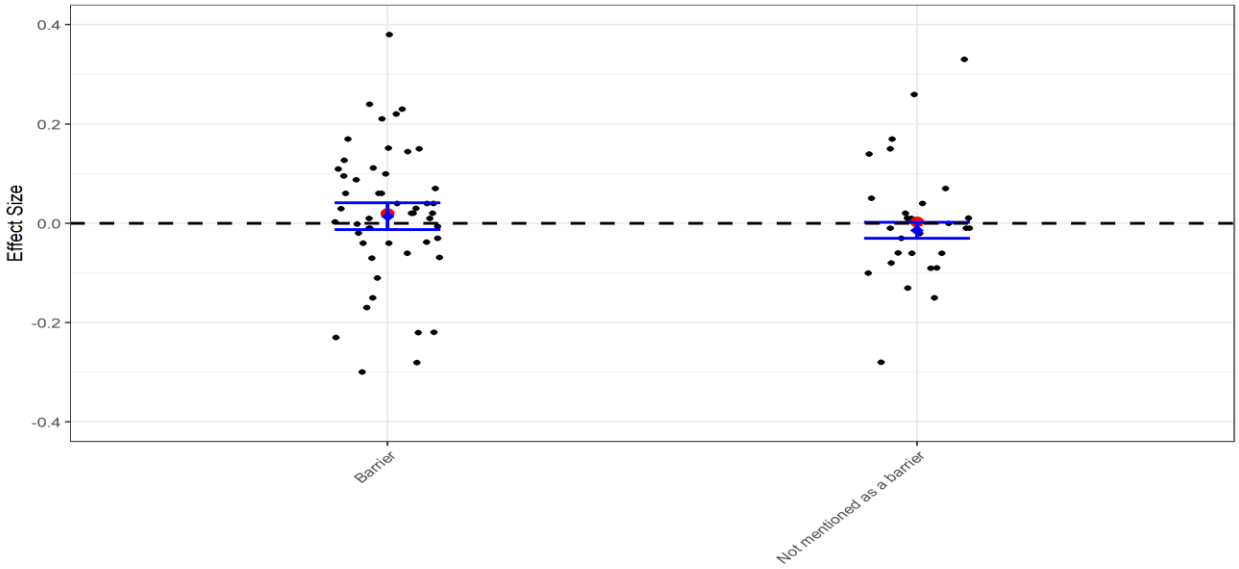
The meta-analyses found the association between perceptions on staff teamwork and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). Reports that mentioned staff teamwork to be an enabler had a higher mean effect size (+0.04 SD) compared with reports that did not mention this as an enabler (−0.02 SD).

Table 75: Effect size by staff teamwork secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Enabler	24	+0.04	0.013	+0.01	+0.06
Not mentioned as an enabler	54	−0.02	0.014	−0.05	+0.01

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Figure 55: Effect size by staff teamwork secondary ITT attainment outcomes



FSM

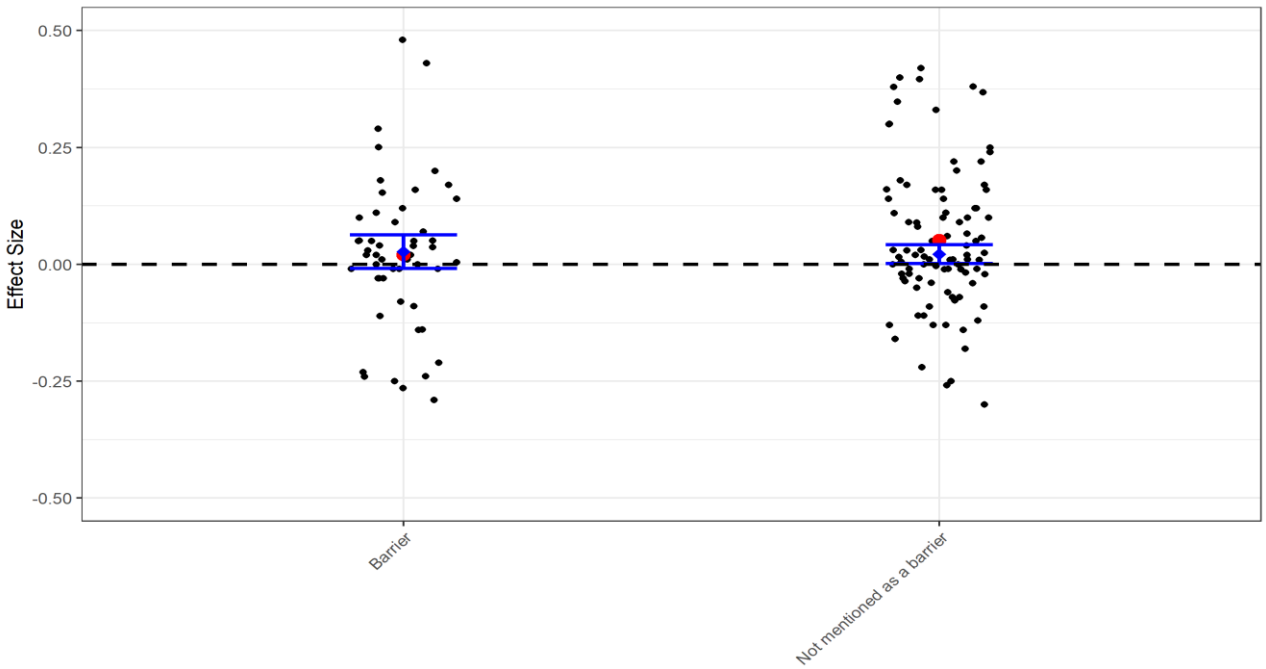
No evidence of an association between perceptions on staff teamwork and FSM effect size was observed.

Table 76: Effect size by staff teamwork FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Enabler	35	+0.03	0.014	0.00	+0.05
Not mentioned as an enabler	114	+0.03	0.012	+0.01	+0.05

Meta *p*-value > 0.10. Overall weighted mean = +0.03 SD.

Figure 56: Effect size by staff teamwork FSM attainment outcomes



Characteristics of participating individuals

Pupil behaviour

Primary ITT

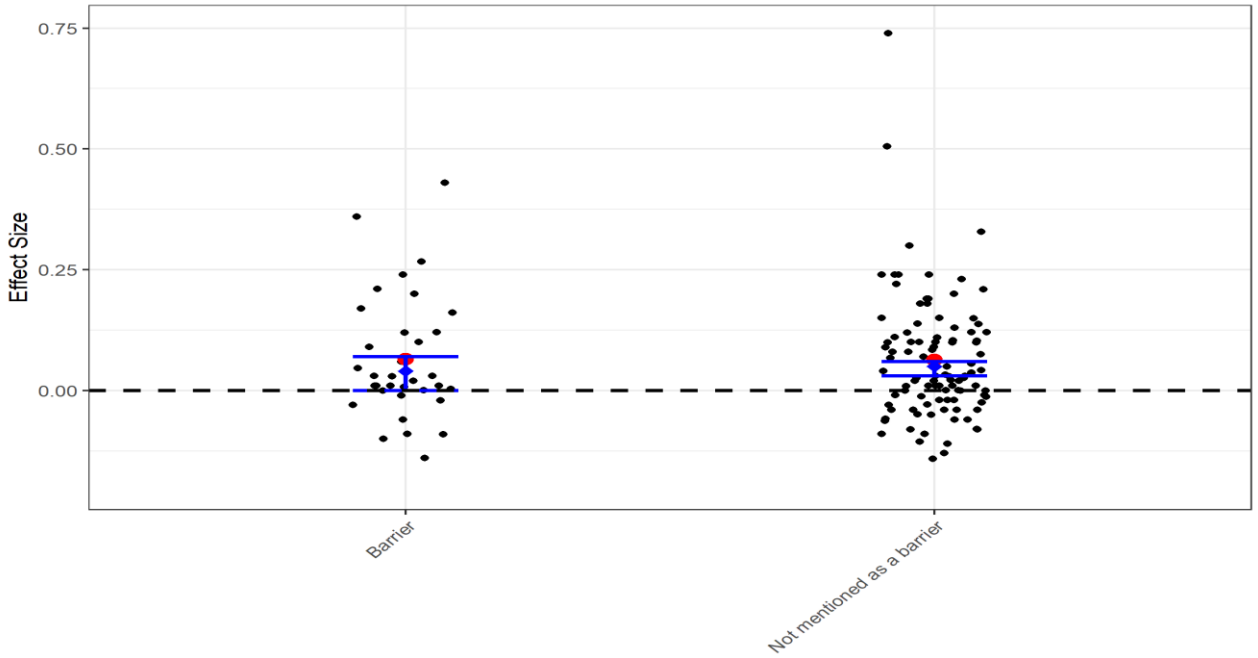
No association between effect sizes and perceptions on pupil behaviour was observed.

Table 77: Effect size by pupil behaviour primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	34	+0.04	0.018	+0.00	+0.07
Not mentioned as a barrier	99	+0.05	0.009	+0.03	+0.06

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Figure 57: Effect size by pupil behaviour primary ITT attainment outcomes



Secondary ITT

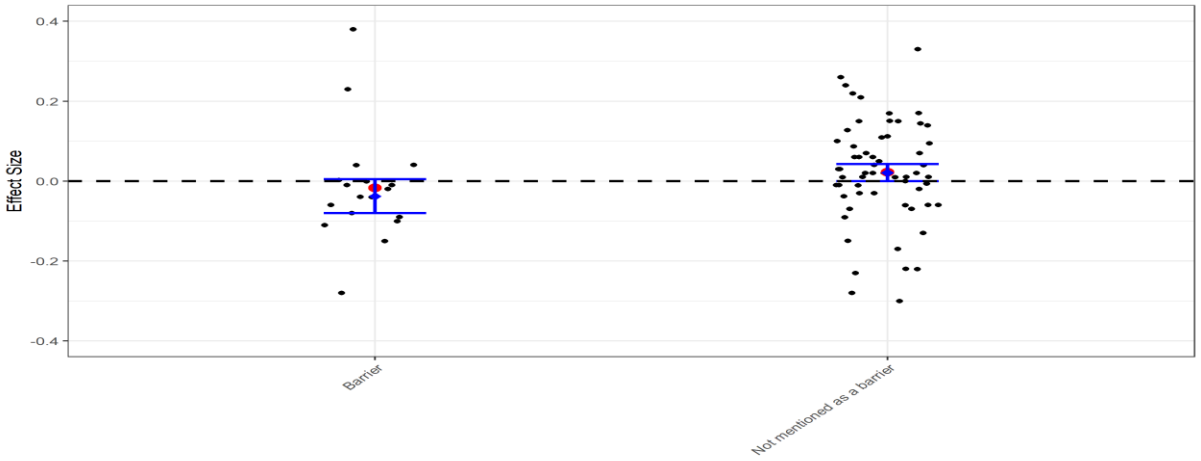
The meta-analyses found the association between perceptions on pupil behaviour and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). Reports that did **not** mention pupil behaviour to be a barrier had a higher mean effect size (+0.02 SD) compared with reports that did mention this as a barrier (−0.04 SD).

Table 78: Effect size by pupil behaviour secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	18	−0.04	0.022	−0.08	+0.01
Not mentioned as a barrier	60	+0.02	0.011	0.00	+0.04

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Figure 58: Effect size by pupil behaviour secondary ITT attainment outcomes



FSM

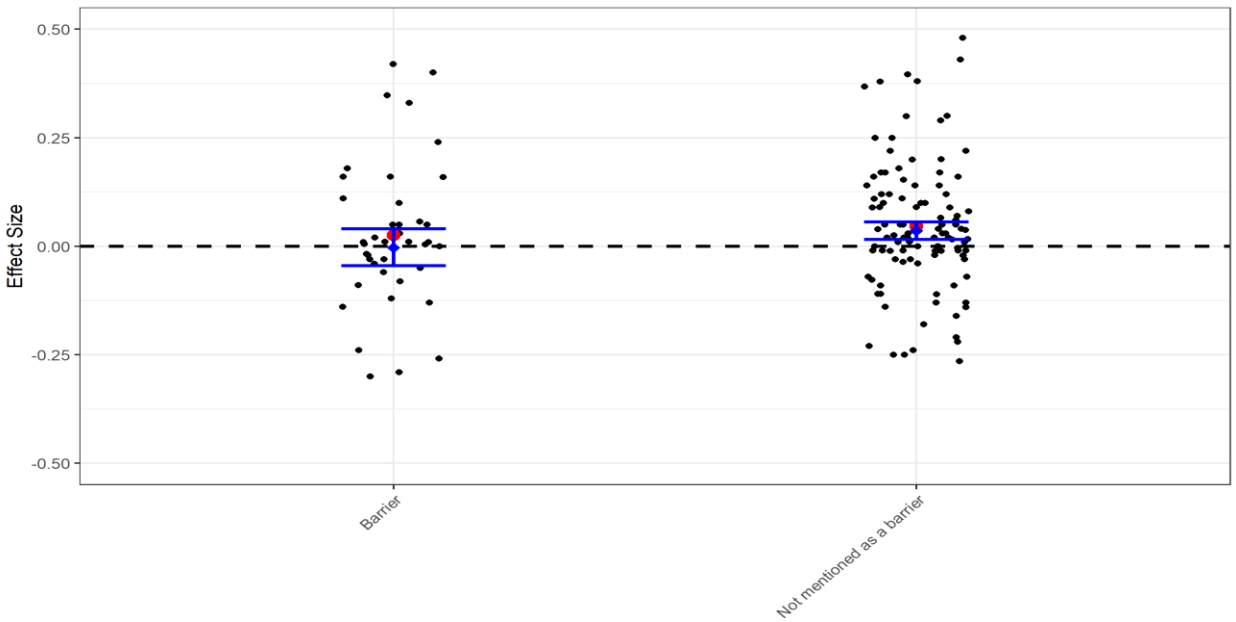
The meta-analyses found the association between perceptions on pupil behaviour and FSM effect size to be statistically significant at the 10% level ($p < 0.10$). Reports that did **not** mention pupil behaviour to be a barrier had a higher mean effect size (+0.04 SD) compared with reports that did mention this as a barrier (0.00 SD).

Table 79: Effect size by pupil behaviour FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	42	0.00	0.022	−0.05	+0.04
Not mentioned as a barrier	107	+0.04	0.010	+0.02	+0.06

Meta *p*-value < 0.10*. Overall weighted mean = +0.03 SD.

Figure 59: Effect size by pupil behaviour FSM attainment outcomes



Staff expectations and motivations

Primary ITT

No association between effect sizes and staff expectations and motivations was observed.

Table 80: Effect size by staff expectations and motivations primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	18	+0.04	0.021	0.00	+0.08
Both barrier and enabler	20	+0.01	0.018	−0.02	+0.05
Enabler	27	+0.06	0.017	+0.03	+0.09
Not mentioned / unclear	68	+0.05	0.012	+0.03	+0.07

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Secondary ITT

The meta-analyses found the association between perceptions on staff expectations and motivations and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). Reports that mentioned staff expectations and motivations to be an enabler had a higher mean effect size (+0.05 SD) compared with reports that did not mention this as an enabler (+0.01 SD or lower). It is difficult to interpret this finding given the lack of association relation to the primary ITT.

Table 81: Effect size by staff expectations and motivations secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	3	–	–	–	–
Both barrier and enabler	11	+0.01	0.014	−0.02	+0.04
Enabler	26	+0.05	0.015	+0.03	+0.08
Not mentioned / unclear	38	−0.05	0.020	−0.09	−0.01

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

FSM

No association between FSM effect sizes and perceptions on staff expectations and motivations was observed.

Table 82: Effect size by staff expectations and motivations FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Barrier	18	0.00	0.014	−0.02	+0.03
Both barrier and enabler	23	−0.02	0.030	−0.08	+0.04
Enabler	30	+0.04	0.016	+0.01	+0.07
Not mentioned / unclear	78	+0.04	0.015	+0.01	+0.07

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Effect sizes and implementation & fidelity

Summary

Table 83: Summary of meta-analyses of effect sizes of attainment outcomes for explanatory variables included in the **implementation & fidelity** theme

Subtheme	Explanatory variable	Primary ITT	Secondary ITT	FSM
Developer characteristics	Charity / university / private company / school or academy or MAT / council or LA / mixed	✓**	✓**	✓***
Implementation planning and time	Clarity of implementation plan	✓	✓***	✓
	Lead-in time for implementation	✓	✓	✓
Professional development	Whether implementation uses CPD	✓#	✓***	✓
	Sequencing of CPD	✓#	✓***	✓
	Whether CPD is subject-specific or generic	✓#	✓***	✓
	Who delivers CPD?	✓	✓***	✓
	Types of CPD	✓	✓	✓
Support & monitoring	Whether developer provided support other than CPD	✓	✓***	✓
	Monitoring of implementation	✓***	✓***	✓**
	SLT support	✓	✓	✓
Fidelity	CPD fidelity	✓**	✓***	✓
	Intended fidelity (by direct implementers)	✓	✓	✓
	Implementation fidelity (by direct implementers)	✓#	✓***	✓

All of the explanatory variables included in the implementation theme were included in the meta-analyses of secondary ITT and FSM attainment outcomes.

Developer characteristics

Primary ITT

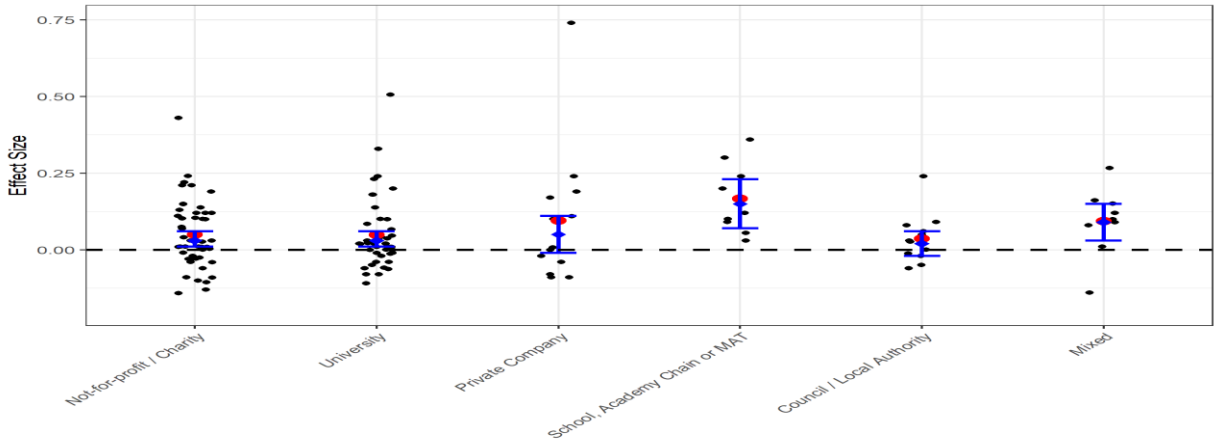
The nine programmes that were from school or academy trust developers had the highest weighted mean effect size (+0.15 SD; 95% CI: +0.07 to +0.23), compared with other types of developers (weighted mean = +0.09 or lower). The association between type of developer and effect size is statistically significant at the 5% level ($p < 0.05$). This aligns with earlier review findings and may occur as school and academy trust developers are able to draw on their knowledge of school context to design interventions that are relatively easy to implement and well matched to 'the problem' they are seeking to address. In addition, they are likely to have gone through a process of piloting and refining the intervention in their own schools prior to the trial.

Table 84: Effect size by developer characteristics primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Not-for-profit / charity	48	+0.04	0.013	+0.01	+0.06
University	42	+0.04	0.013	+0.01	+0.06
Private company	13	+0.05	0.030	−0.01	+0.11
School, academy chain or MAT	9	+0.15	0.042	+0.07	+0.23
Council / local authority	12	+0.02	0.020	−0.02	+0.06
Mixed	9	+0.09	0.031	+0.03	+0.15

* Meta *p*-value < 0.05**. Overall weighted mean = +0.04 SD.

Figure 60: Effect size by developer characteristics primary ITT attainment outcomes



Secondary ITT

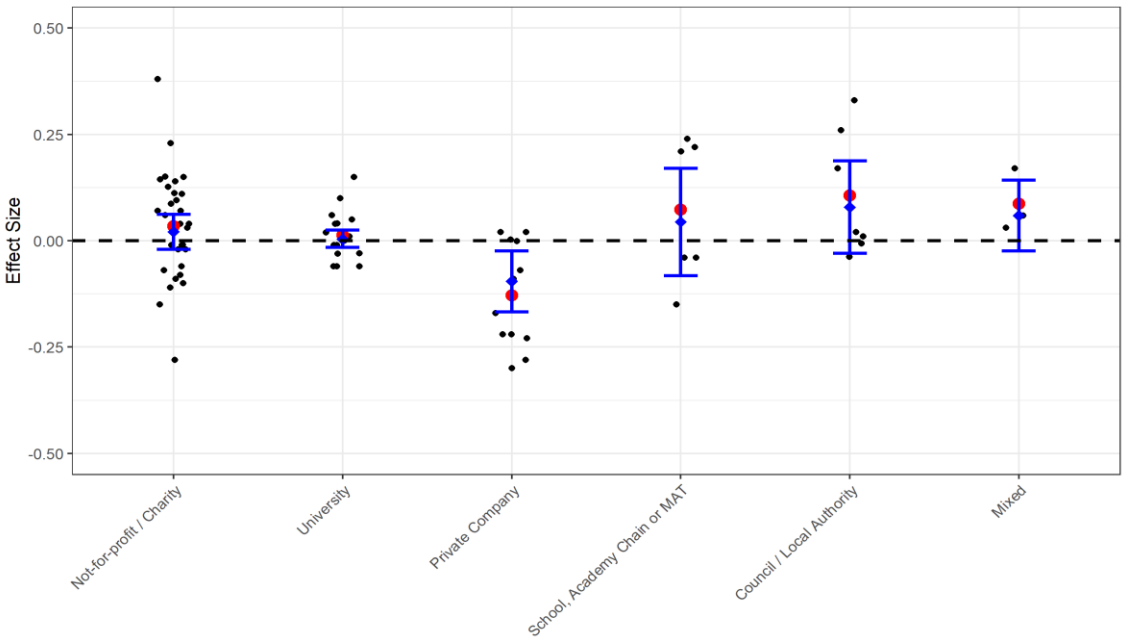
The meta−analyses found the association between type of developer and secondary ITT effect size to be statistically significant at the 5% level ($p < 0.05$). The highest weighted mean effect size was also observed for programmes that were from local authority developers (+0.08 SD; 95% CI: −0.03 to +0.19) compared with other types of developers (weighted mean = +0.04 SD or lower). It is unclear why this differs from the primary ITT analyses.

Table 85: Effect size by developer characteristics secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Not-for-profit / charity	30	+0.02	0.021	−0.02	+0.06
University	19	0.00	0.011	−0.02	+0.03
Private company	13	−0.10	0.036	−0.17	−0.02
School, academy chain or MAT	6	+0.04	0.064	−0.08	+0.17
Council / local authority	7	+0.08	0.055	−0.03	+0.19
Mixed	3	–	–	–	–

* Meta *p*-value < 0.05**. Overall weighted mean = +0.01 SD.

Figure 61: Effect size by developer characteristics secondary ITT attainment outcomes



FSM

The meta-analyses found the association between type of developers and FSM effect size to be statistically significant at the 1% level ($p < 0.01$). In line with the findings for the primary ITT analysis, the highest weighted mean FSM effect size was also observed for programmes that were from schools or academy trust developers (+0.14 SD; 95% CI: +0.04 to +0.24) compared with other types of developers (weighted mean = +0.06 SD or lower).

Table 86: Effect size by developer characteristics FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Not-for-profit / charity	47	+0.03	0.020	−0.01	+0.07
University	41	+0.03	0.013	0.00	+0.05
Private company	24	−0.02	0.025	−0.07	+0.03
School, academy chain or MAT	10	+0.14	0.051	+0.04	+0.24
Council / local authority	15	−0.01	0.026	−0.06	+0.04
Mixed	12	+0.06	0.038	−0.01	+0.13

* Meta *p*-value < 0.01***. Overall weighted mean = +0.03 SD.

FSM attainment outcomes (meta *p*-value < 0.01***): The R 'metafor' package was unable to complete analyses with this variable[30] and so the weighted mean estimates were done by hand (see Appendix D). This means that a plot for FSM effect sizes is not available.

Implementation planning and time

Primary ITT

No evidence was found for an association between effect size and reported perceptions on clarity of implementation planning or the preparation lead-in time. It is important to note here that we have less confidence in the data for this category, as it was not routinely or systematically reported.

Table 87: Effect size by perceived clarity of implementation plan; primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Clearly understood	49	+0.04	0.011	+0.02	+0.06
Variation in perceptions	37	+0.04	0.019	0.00	+0.08
Unclear or not mentioned	47	+0.05	0.014	+0.02	+0.08

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Secondary ITT

The meta-analyses found the association between reported perceptions on clarity of implementation planning and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). The highest weighted mean effect size was observed where the implementation plan was unclear or not mentioned (+0.10 SD; 95% CI: +0.07 to +0.14) compared with when the plan was clearly understood (0.00 SD; 95% CI: −0.02; +0.02) or when a variation in understanding was reported (−0.03 SD; 95% CI: −0.07; +0.01). The finding contradicts the implementation literature, for

[30] Error in RMA …Fischer scoring algorithm did not converge (R 'metafor' error message). On investigation, this relates to the tau^2 (T^2) estimate for the council/local authority grouping, which was close to zero but negative (−0.003). Whilst T^2 cannot be negative, methods to estimate this with T^2 can result in negative values (Borenstein et al., 2009). The manual approach to resolve this is to set the T^2 estimate to zero (Borenstein et al., 2009 and see Appendix B). This seems to have presented a problem for the R metafor package and so no meta-analysis output (including the error bars) was generated.

example as reviewed in the EEF implementation guidance (Sharples et al., 2019). Beyond limitations in the data, it is unclear why this is the case.

No evidence was found for an association between secondary ITT effect size and reported perceptions on preparation lead-in time as was found for the primary outcome.

Table 88: Effect size by perceived clarity of implementation plan; secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Clearly understood	32	0.00	0.010	−0.02	+0.02
Variation in perceptions	30	−0.03	0.019	−0.07	+0.01
Unclear or not mentioned	16	+0.10	0.019	+0.07	+0.14

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

FSM

No evidence was found for an association between FSM effect size and reported perceptions on clarity of implementation planning or the preparation lead-in time.

Table 89: Effect size by perceived clarity of implementation plan; FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Clearly understood	54	+0.04	0.018	+0.01	+0.07
Variation in perceptions	55	0.00	0.012	−0.03	+0.02
Unclear or not mentioned	40	+0.05	0.022	+0.01	+0.09

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Professional development

Use of CPD

Primary ITT

No evidence was found for an association between effect size and whether or not CPD was provided to support implementation of the intervention (Table 90). Whilst CPD was very common (119 of the 133 effect sizes) and had a higher weighted mean effect size (+0.05 SD; 95% CI: +0.03 : +0.06) than interventions with no CPD (+0.02 SD; 95% CI: −0.03 : +0.08), this difference was not statistically significant. Amongst interventions where CPD was provided, no evidence was found for an association between effect size and type of CPD (Table 91). The lack of association between effect size and whether or not CPD was provided may have occurred as there were few trials that did not include CPD. The lack of association between effect size and type of CPD is surprising, given the growing body of literature on the effectiveness of different forms of CPD; this may have occurred given the small number of trials that did not include CPD. Furthermore, the inconsistencies in reporting undermine the reliability of this finding.

Table 90: Effect size by CPD provision; primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
No CPD / unclear	14	+0.02	0.027	−0.03	+0.08
CPD provided in intervention	119	+0.05	0.008	+0.03	+0.06

* Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Table 91: Effect size by types of CPD primary ITT attainment outcomes

Factor		n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Face-to-face training (meta *p*-value >0.10)	Yes	115	+0.04	0.008	+0.03	+0.06
	Not mentioned or unclear	4	+0.11	0.105	−0.10	+0.31
Online training (meta *p*-value >0.10)	Yes	15	+0.03	0.022	−0.02	+0.07
	No	104	+0.05	0.009	+0.03	+0.07
Coaching or mentoring (meta *p*-value >0.10)	Yes	18	+0.05	0.023	+0.01	+0.10
	No	101	+0.05	0.009	+0.03	+0.06
Cascade 'train the trainer' training (meta *p*-value >0.10)	Yes	24	+0.05	0.020	+0.01	+0.08
	No	95	+0.05	0.009	+0.03	+0.06

Overall weighted mean = +0.04 SD.

Secondary ITT

The meta-analyses found the association between use of CPD and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$) (Table 92). The highest weighted mean effect size was observed where no CPD was reported (+0.08 SD; 95% CI: +0.03 to +0.13), compared with when CPD was reported (−0.01 SD; 95% CI: −0.03 to +0.01). No associations were observed between secondary ITT effect size and types of CPD (Table 93).

Table 92: Effect size by CPD provision; secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
No CPD / unclear	10	+0.08	0.023	+0.03	+0.13
CPD provided in intervention	68	−0.01	0.010	−0.03	+0.01

* Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Table 933: Effect size by types of CPD secondary ITT attainment outcomes

Factor		n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Face-to-face training (meta *p*-value < 0.01***)	Yes	68	−0.01	0.010	−0.03	+0.01
	Not mentioned or unclear	0	–	–	–	–
Online training (meta *p*-value > 0.10)	Yes	5	+0.06	0.035	0.00	+0.13
	No	63	−0.01	0.010	−0.03	+0.01
Coaching or mentoring (meta *p*-value > 0.10)	Yes	9	+0.03	0.053	−0.08	+0.13
	No	59	−0.01	0.008	−0.02	+0.01
Cascade 'train the trainer' training (meta *p*-value > 0.10)	Yes	13	−0.06	0.045	−0.14	+0.03
	No	55	−0.01	0.009	−0.02	+0.01

Overall weighted mean = +0.01 SD.

FSM

As for the primary ITT analysis, no evidence was found for an association between FSM effect size and whether or not CPD was provided to support implementation of the intervention (Table 94) or across different types of CPD (Table 95).

Table 94: Effect size by CPD provision; FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
No CPD / Unclear	11	−0.01	0.040	−0.08	+0.07
CPD provided in intervention	138	+0.03	0.009	+0.01	+0.05

* Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Table 95: Effect size by types of CPD FSM attainment outcomes

Factor		n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Face-to-face training (meta *p*-value > 0.10)	Yes	135	+0.03	0.009	+0.01	+0.04
	Not mentioned or unclear	3	–	–	–	–
Online training (meta *p*-value > 0.10)	Yes	17	+0.04	0.028	−0.02	+0.09
	No	121	+0.03	0.010	+0.01	+0.05
Coaching or mentoring (meta *p*-value > 0.10)	Yes	23	+0.03	0.020	−0.01	+0.07
	No	115	+0.03	0.011	+0.01	+0.05
Cascade 'train the trainer' training (meta *p*-value > 0.10)	Yes	37	+0.02	0.013	0.00	+0.05
	No	101	+0.04	0.012	+0.01	+0.06

Overall weighted mean = +0.03 SD**.**

Sequencing of CPD

Primary ITT

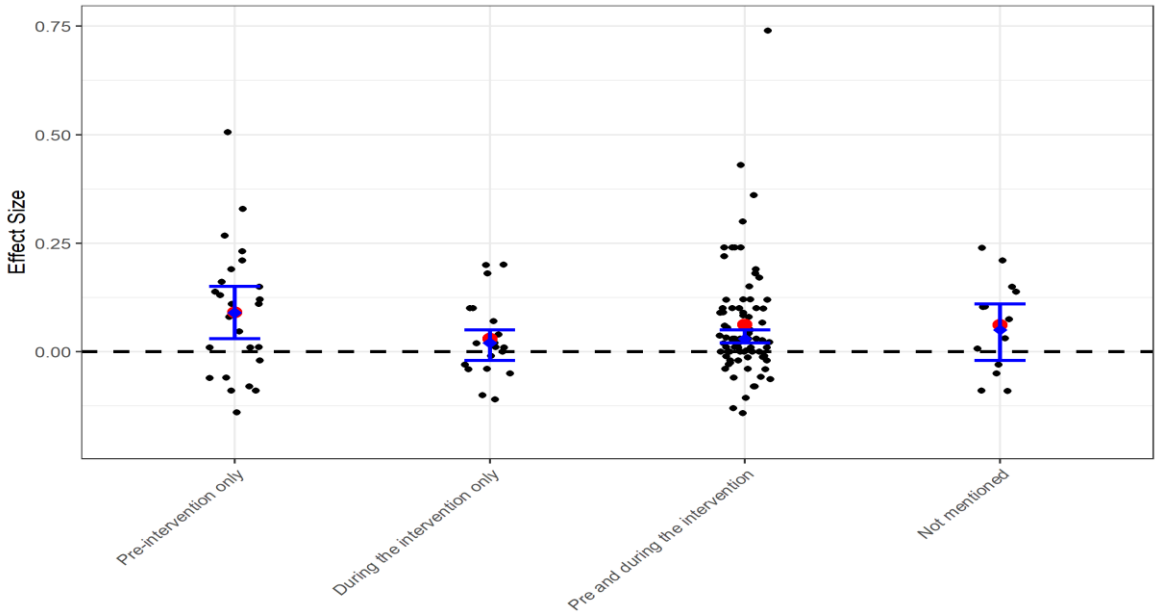
Whilst not statistically significant, it was notable that the highest average effect sizes were observed for trials that delivered CPD pre-intervention only (weighted mean = +0.09 SD; 95% CI: +0.03 to +0.15) compared with programmes where CPD was also delivered during the intervention period (weighted mean = +0.03 SD or lower). This may indicate the importance of early CPD for building implementers' confidence, knowledge, skills and capacities and for ensuring fidelity to the intervention, but this does explain the notably lower effect size for CPD delivered pre-intervention and during the intervention.

Table 96: Effect size by sequencing of CPD primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Pre-intervention only	25	+0.09	0.029	+0.03	+0.15
During the intervention only	19	+0.02	0.016	−0.02	+0.05
Pre-intervention and during the intervention	72	+0.04	0.008	+0.02	+0.05
Not mentioned	3	–	–	–	–

Meta *p*-value > 0.10. Overall weighted mean = +0.04 SD.

Figure 62: Effect size by sequencing of CPD primary ITT attainment outcomes



Secondary ITT

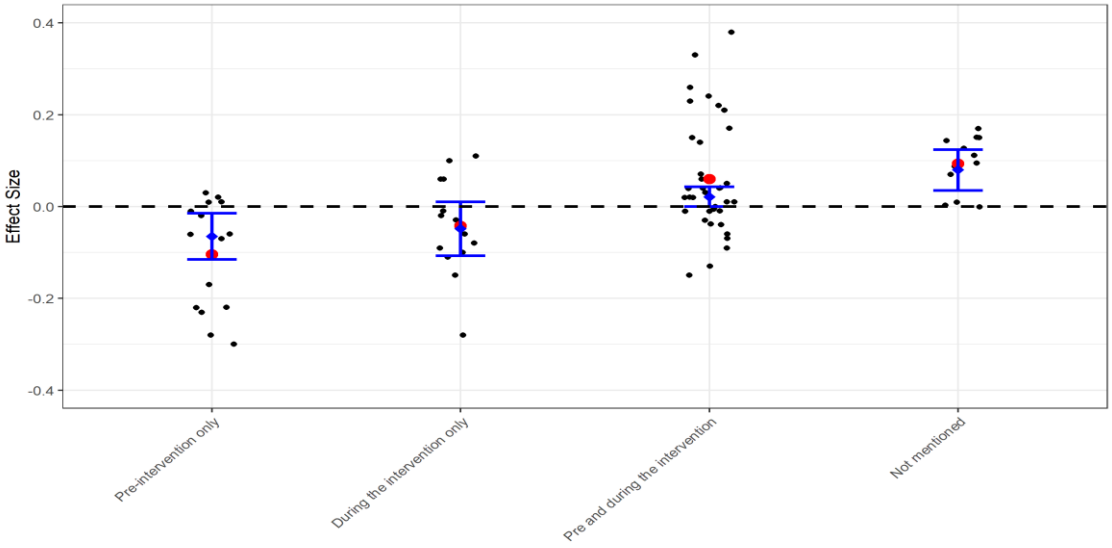
The meta-analyses found the association between the sequencing of CPD and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). The highest weighted mean when scheduling of CPD reported was for pre-intervention and during the intervention (+0.02 SD), compared with pre-intervention only (−0.07 SD) or during intervention only (−0.05 SD).

Table 97: Effect size by sequencing of CPD secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Pre-intervention only	15	−0.07	0.025	−0.12	−0.02
During the intervention only	15	−0.05	0.030	−0.11	+0.01
Pre-intervention and during the intervention	36	+0.02	0.011	0.00	+0.04
Not mentioned	2	–	–	–	–

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Figure 63: Effect size by sequencing of CPD secondary ITT attainment outcomes



FSM

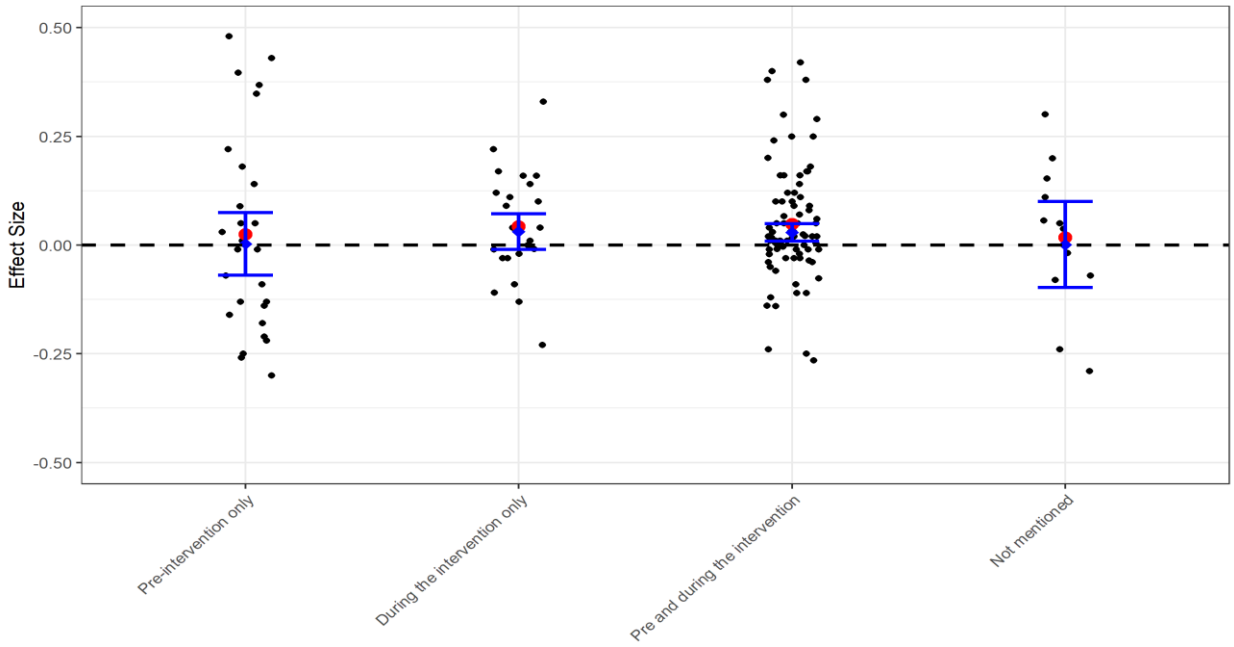
No evidence was found for an association between FSM effect size and sequencing of CPD.

Table 98: Effect size by sequencing of CPD FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Pre-intervention only	30	0.00	0.037	−0.07	+0.08
During the intervention only	24	+0.03	0.021	−0.01	+0.07
Pre-intervention and during the intervention	79	+0.03	0.011	+0.01	+0.05
Not mentioned	5	0.10	0.089	−0.07	+0.28

Meta *p*-value > 0.10. Overall weighted mean = +0.03 SD.

Figure 64: Effect size by sequencing of CPD FSM attainment outcomes



Whether CPD is subject-specific or generic

Primary ITT

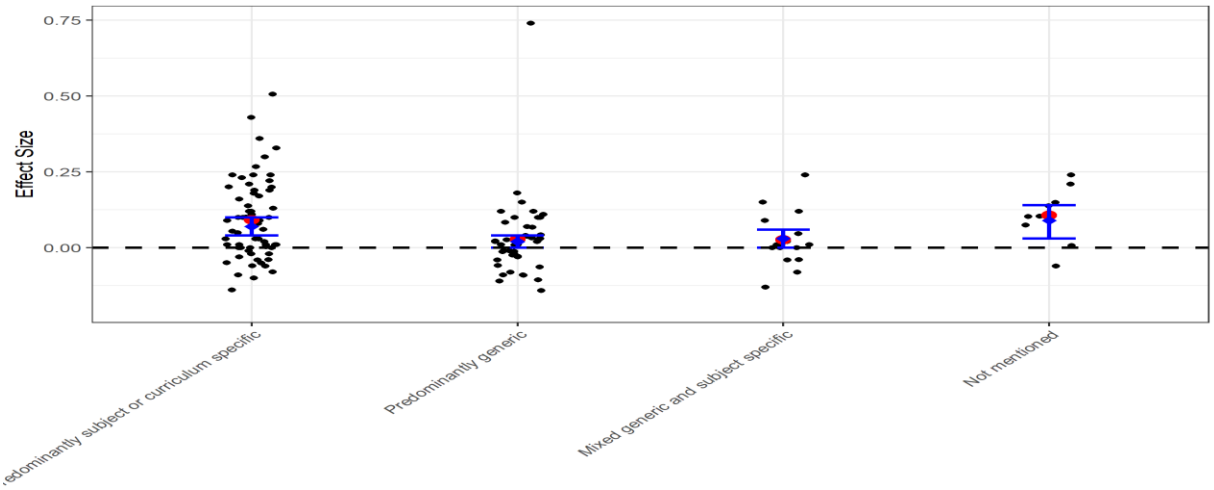
Programmes with CPD that is subject-specific or curriculum-specific are associated with higher average effect sizes (weighted mean = +0.07 SD; 95% CI: +0.04 to +0.10) than programmes with more generic CPD (weighted mean = +0.03 SD) but this was not found to be statistically significant. This finding resonates with existing research which indicates that subject-specific CPD is more likely to be effective in changing teachers' practices than generic CPD.

Table 99: Effect size by whether CPD was subject-specific/generic primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Predominantly subject- or curriculum-specific	63	+0.07	0.015	+0.04	+0.11
Predominantly generic	40	+0.03	0.010	0.00	+0.04
Mixed generic and subject-specific	15	+0.03	0.018	−0.01	+0.06

Meta *p*-value > 0.10. Overall weighted mean = +0.04 SD.

Figure 65: Effect size by whether CPD is subject-specific/generic primary ITT attainment outcomes



Secondary ITT

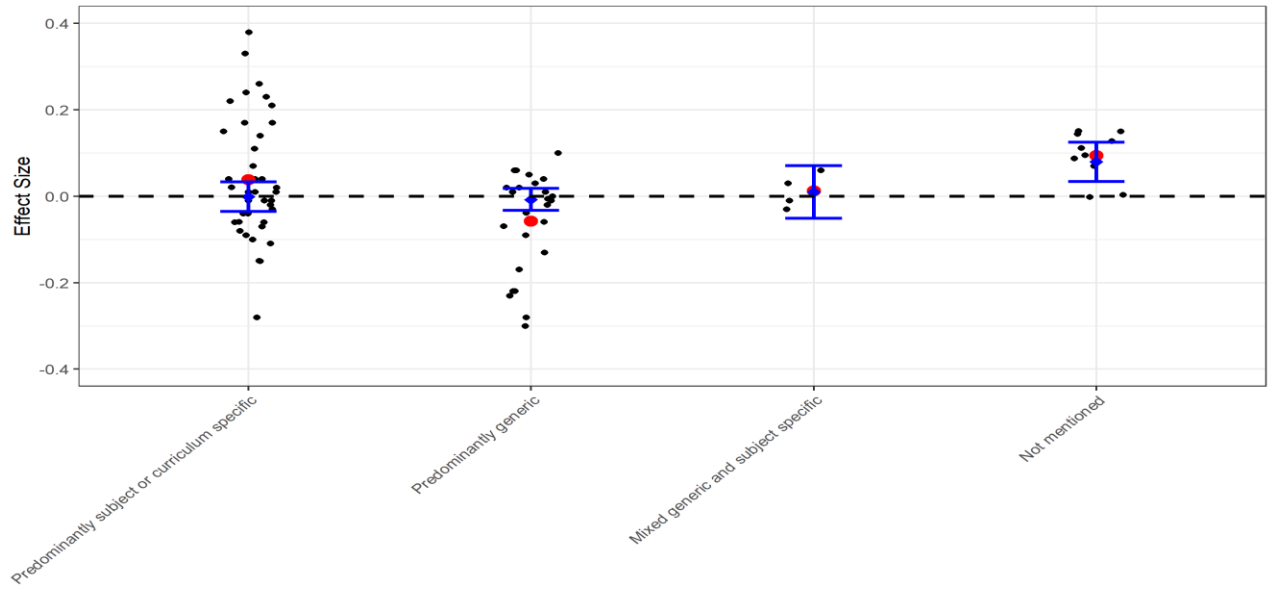
No evidence for an association between Secondary ITT effect size and whether the CPD was subject-specific, generic or mixed was observed.

Table 100: Effect size by whether CPD was subject-specific/generic secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Predominantly subject- or curriculum-specific	39	0.00	0.017	−0.04	+0.03
Predominantly generic	25	−0.01	0.013	−0.03	+0.02
Mixed generic and subject-specific	4	+0.01	0.031	−0.05	+0.07

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.01 SD.

Figure 66: Effect size by whether CPD is subject-specific/generic secondary ITT attainment outcomes



FSM

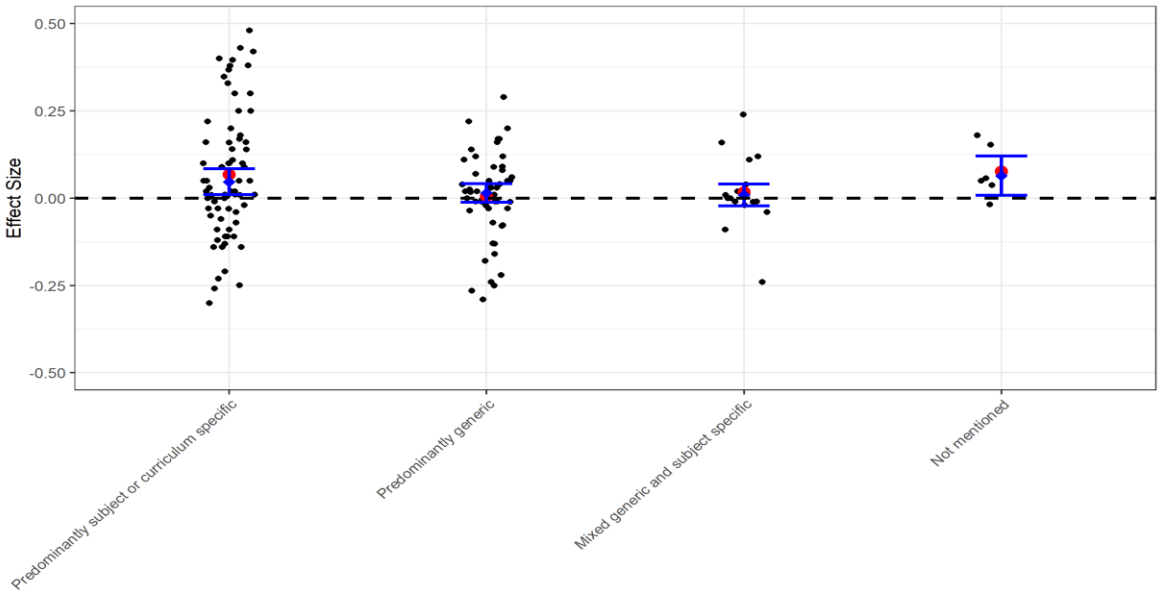
There was no evidence for an association between FSM effect size and whether the CPD was subject-specific, generic or mixed was observed, although the highest weighted mean when the focus of CPD was mentioned was for subject- or curriculum-specific CPD (weighted mean = +0.05 SD; 95% CI: +0.01 to +0.08).

Table 101: Effect size by whether CPD was subject-specific/generic FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Predominantly subject- or curriculum-specific	74	+0.05	0.019	+0.01	+0.08
Predominantly generic	47	+0.03	0.011	0.00	+0.05
Mixed generic and subject-specific	16	+0.01	0.016	−0.02	+0.04
Not mentioned	1	–	–	–	–

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 67: Effect size by whether CPD is subject-specific/generic FSM attainment outcomes



Who delivers CPD

Primary ITT

No evidence was found for an association between primary ITT effect size and who delivered the CPD.

Secondary ITT

No evidence was found for an association between secondary ITT effect size and who delivered the CPD.

FSM

No evidence was found for an association between FSM effect size and who delivered the CPD.

Support and monitoring of intervention

Provision of support other than CPD

Primary ITT

No evidence was found for an association between effect size and whether the developer provided any informal support, by telephone or email for the intervention beyond the formal CPD constituted by training sessions or structured mentoring and coaching, etc.

Secondary ITT

The meta-analyses found the association between whether the developer provided any informal support beyond CPD and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). The highest weighted mean effect size was observed where support was provided just during the intervention (+0.10 SD; 95% CI: +0.07 to +0.14) compared with other categories (0.00 SD or lower).

FSM

No evidence was found for an association between FSM effect size and whether the developer provided any informal support beyond CPD.

Monitoring of implementation

Primary ITT

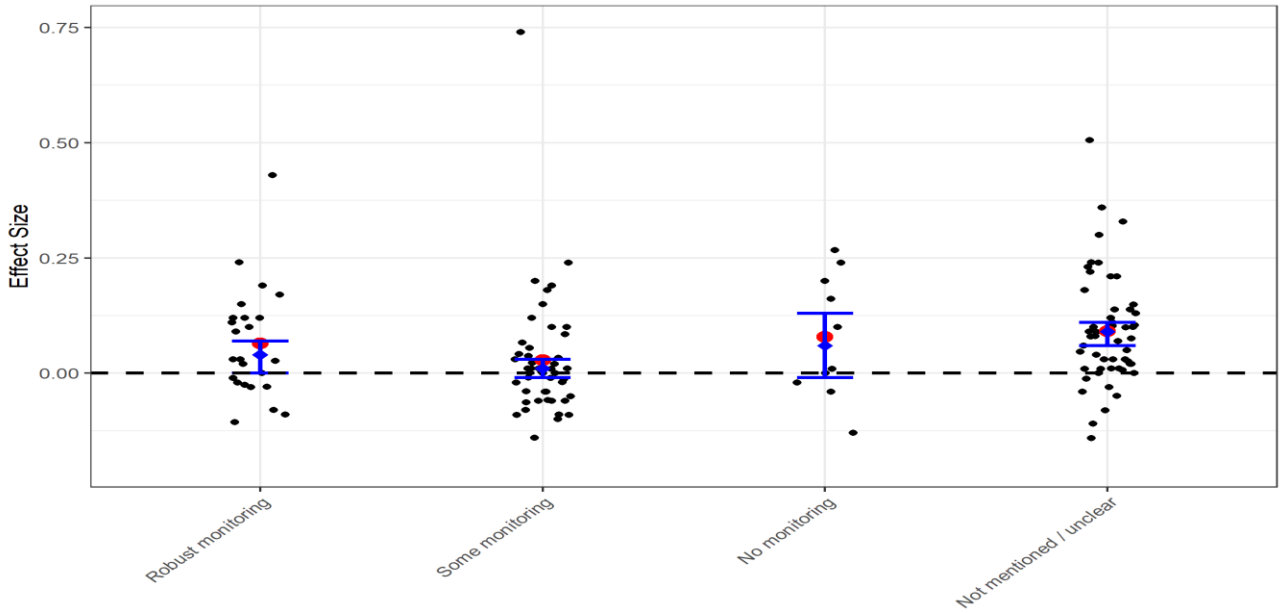
The meta-analyses found the association between monitoring of interventions and primary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). Higher effect sizes were associated with evaluations that reported no monitoring of interventions (weighted mean = +0.06 SD; 95% CI: −0.01 to +0.13) or when monitoring was not mentioned (weighted mean = +0.09 SD; 95% CI: +0.06 to +0.11) compared with programmes where robust monitoring was reported (weighted mean = +0.04 SD; 95% CI: 0.00 to +0.07). This is a particularly surprising finding, as it runs counter to the review finding of Anders et al. (2017) and the intervention evaluation literature more widely. It may reflect inconsistent reporting in relation to this category.

Table 102: Effect size by monitoring of intervention primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Robust monitoring	24	+0.04	0.018	0.00	+0.07
Some monitoring	47	+0.01	0.010	−0.01	+0.03
No monitoring	10	+0.06	0.036	−0.01	+0.13
Not mentioned / unclear	52	+0.09	0.015	+0.06	+0.12

Meta *p*-value < 0.01***. Overall weighted mean = +0.04 SD.

Figure 68: Effect size by monitoring of intervention primary ITT attainment outcomes



Secondary ITT

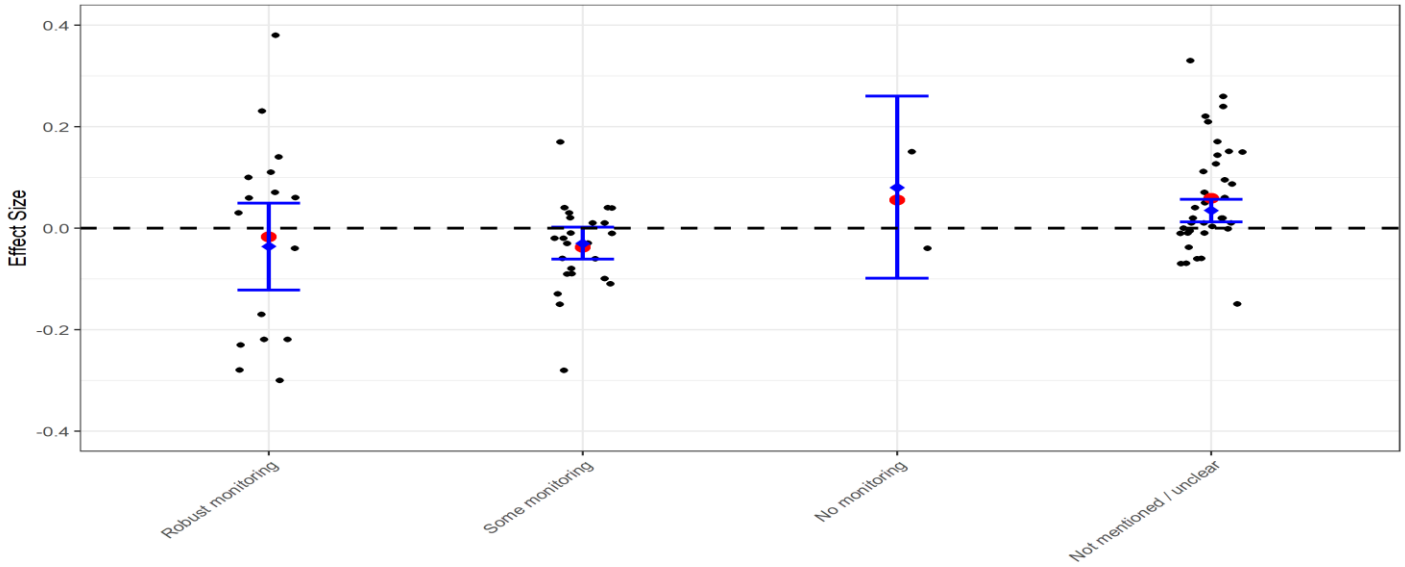
The meta-analyses found the association between monitoring of interventions and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). Higher effect sizes were associated with evaluations that made no mention of monitoring (weighted mean = +0.04 SD; 95% CI: +0.01 to +0.06) compared with when some monitoring was reported (weighted mean = −0.03 SD; 95% CI: −0.06 to 0.00) or where robust monitoring was reported (weighted mean = −0.04 SD; 95% CI: −0.12 to +0.05).

Table 103: Effect size by monitoring of intervention secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Robust monitoring	16	−0.04	0.044	−0.12	+0.05
Some monitoring	24	−0.03	0.016	−0.06	0.00
No monitoring	2	–	–	–	–
Not mentioned / unclear	36	+0.04	0.012	+0.01	+0.06

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Figure 69: Effect size by monitoring of intervention secondary ITT attainment outcomes



FSM

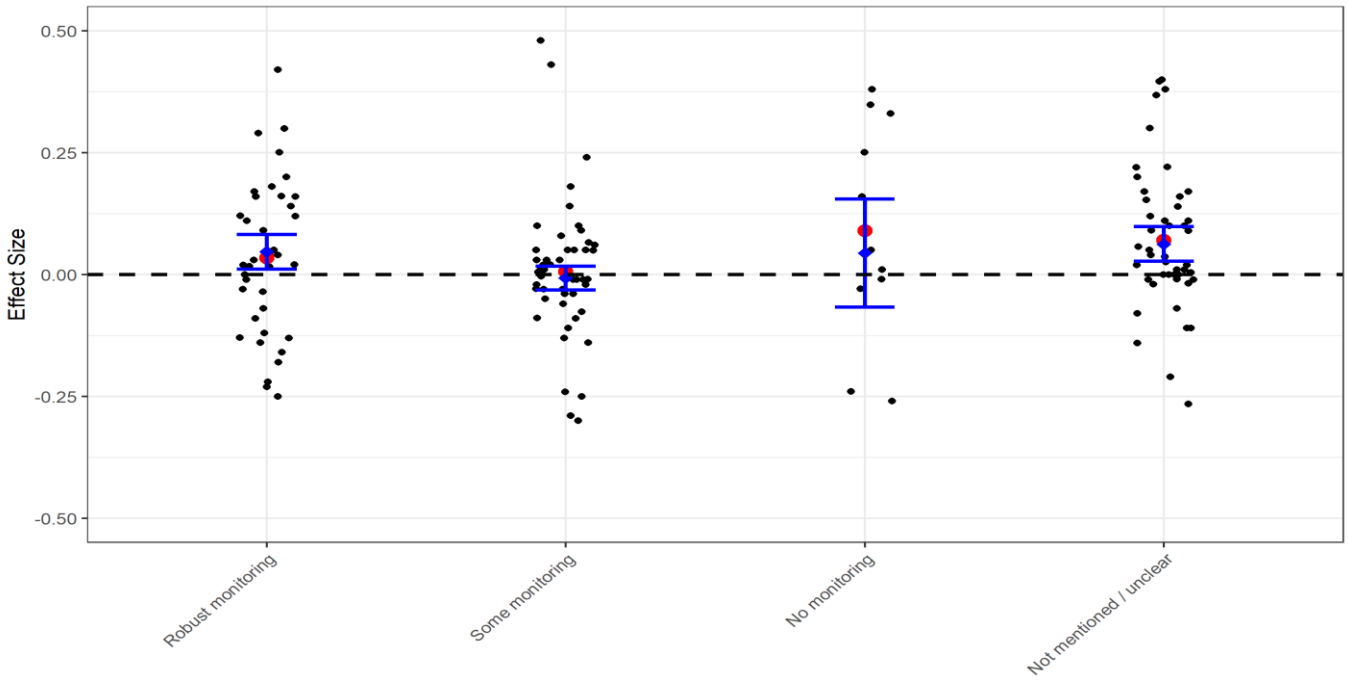
The meta-analyses found the association between monitoring of interventions and FSM effect size to be statistically significant at the 5% level ($p < 0.05$). Higher FSM effect sizes were associated with evaluations that made no mention of monitoring (weighted mean = +0.06 SD; 95% CI: +0.03 to +0.10) or when monitoring was robust (weighted mean = +0.05 SD; 95% CI: +0.01 to +0.08) compared with programmes where no monitoring was reported (weighted mean = +0.04 SD; 95% CI: −0.07 to +0.16) or some monitoring (weighted mean = −0.01 SD; 95% CI: −0.03 to +0.02).

Table 104: Effect size by monitoring of intervention FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Robust monitoring	38	+0.05	0.018	+0.01	+0.08
Some monitoring	51	−0.01	0.012	−0.03	+0.02
No monitoring	13	+0.04	0.057	−0.07	+0.16
Not mentioned / unclear	47	+0.06	0.018	+0.03	+0.10

Meta *p*-value < 0.05**. Overall weighted mean = +0.03 SD.

Figure 70: Effect size by monitoring of intervention FSM attainment outcomes



Senior leadership team (SLT) support

No evidence was found for an association between effect size and SLT support. This was observed for primary ITT, secondary ITT and FSM effect sizes. This is a particularly surprising finding, as it runs counter to the review finding of Anders et al. (2017) and the intervention evaluation literature more widely, and again may reflect limitations in the data.

Fidelity

Fidelity to CPD and fidelity of implementation by the direct implementer (who in most but not all trials would have taken part in the CPD) were examined separately.

Fidelity related to CPD

Primary ITT

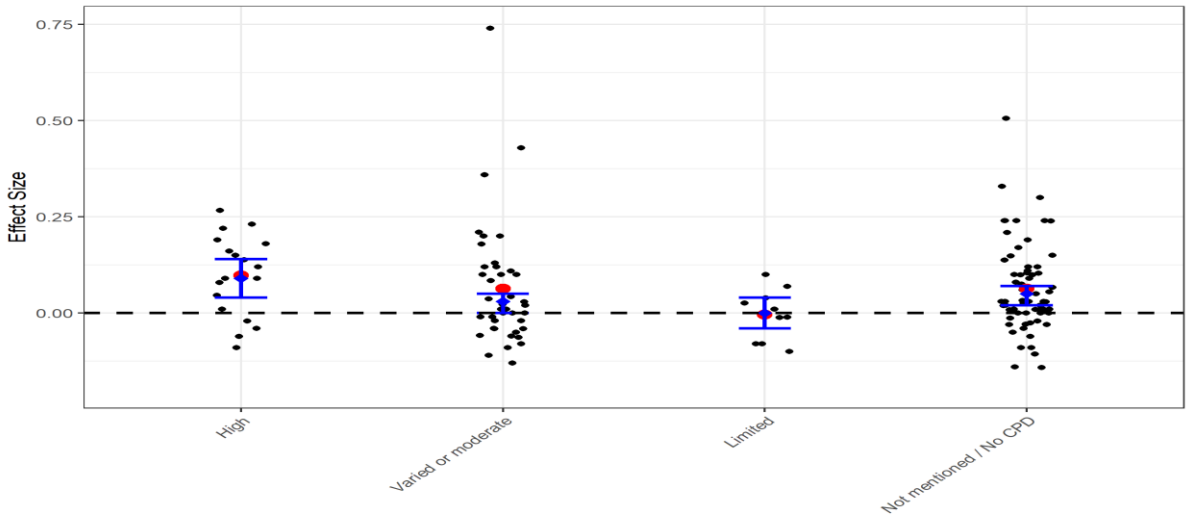
Fidelity relating to CPD was observed to be statistically significantly ($p < 0.05$) associated with effect size. The 12 evaluations that reported high CPD fidelity are observed to have a higher average effect size (weighted mean = +0.09 SD; 95% CI: +0.05 to +0.14) compared with the 38 evaluations that did not mention CPD fidelity (weighted mean = +0.05 SD; 95% CI: +0.03 to +0.21) or reported lower levels of CPD fidelity (weighted mean = +0.03 or lower).

Table 105: Effect size by fidelity related to CPD primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
High	18	+0.09	0.025	+0.05	+0.14
Varied or moderate	40	+0.03	0.013	0.00	+0.05
Limited	10	0.00	0.020	−0.04	+0.04
Not mentioned / no CPD	65	+0.05	0.012	+0.03	+0.07

Meta *p*-value < 0.05**. Overall weighted mean = +0.04 SD.

Figure 71: Effect size by fidelity related to CPD primary ITT attainment outcomes



Secondary ITT

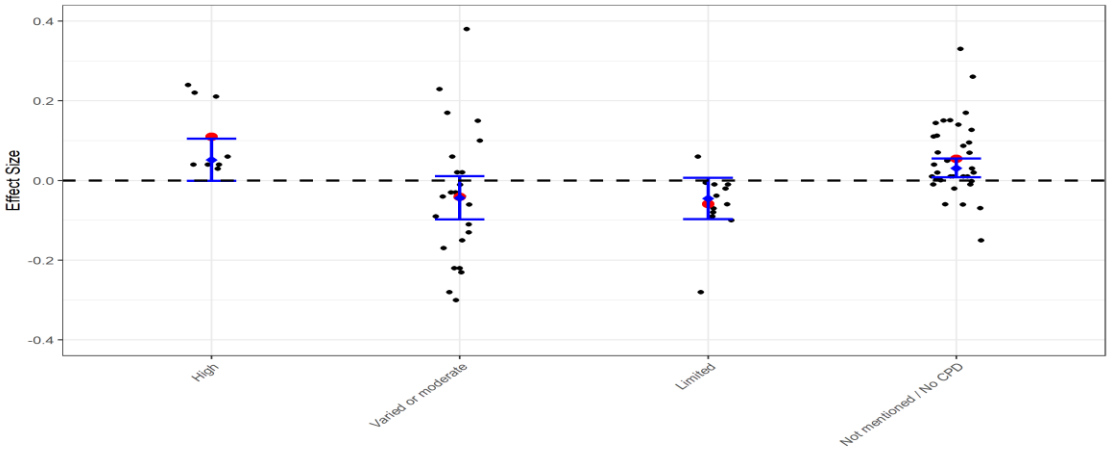
Similarly, the meta-analyses found the association between fidelity relating to CPD and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). The highest weighted mean effect size was also observed where reported CPD fidelity was high (+0.05 SD; 95% CI: 0.00 to +0.11) compared with other categories (+0.03 SD or lower). This finding for both the primary and secondary outcomes underscores the importance of delivery partners paying attention to ensuring high fidelity of CPD delivery, both in terms of content of the CPD and the attendance by target participants.

Table 106: Effect size by fidelity related to CPD secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
High	8	+0.05	0.027	0.00	+0.11
Varied or moderate	24	−0.04	0.028	−0.10	+0.01
Limited	12	−0.05	0.027	−0.10	+0.01
Not mentioned / no CPD	34	+0.03	0.012	+0.01	+0.06

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Figure 72: Effect size by fidelity related to CPD secondary ITT attainment outcomes



FSM

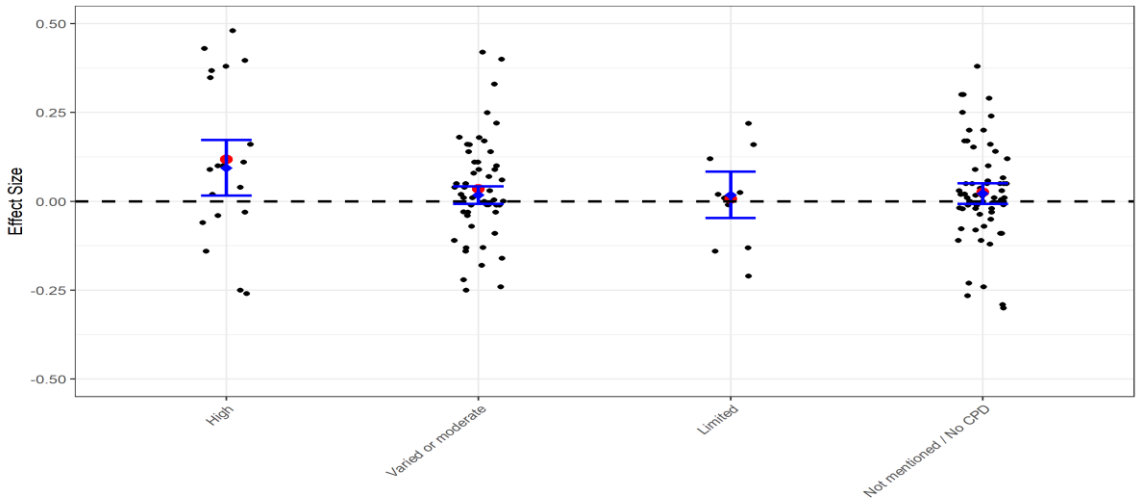
No evidence for an association between FSM effect size and fidelity relating to CPD was observed. However, evaluations that reported high CPD fidelity are observed to have a higher average FSM effect size (weighted mean = +0.09 SD; 95% CI: +0.02 to +0.17) compared with other categories (weighted mean in all = +0.02 SD).

Table 107: Effect size by fidelity related to CPD FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
High	22	+0.09	0.040	+0.02	+0.17
Varied or moderate	53	+0.02	0.013	−0.01	+0.04
Limited	11	+0.02	0.034	−0.05	+0.08
Not mentioned / no CPD	63	+0.02	0.015	−0.01	+0.05

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 73: Effect size by fidelity related to CPD FSM attainment outcomes



Intended fidelity

No evidence was found for an association between effect size and intended approach to fidelity (i.e., faithful adoption or adaptation to context). This was found for primary ITT, secondary ITT and FSM effect sizes. The lack of association suggests that interventions that less tightly codified interventions that are intended to be adapted to context may be equally as likely to lead to positive effects as more strictly codified interventions that are designed to be faithfully adopted.

The lack of association between intended fidelity and effect size is interesting. Providing an intervention does actually impact the primary outcome, it might be assumed that interventions that are intended to be adopted faithfully are more likely to lead to positive effect than interventions where there the direct implementer has more flexibility to adapt the intervention. Tentatively, this is an important finding, indicating that interventions that are intended to be adapted to context, and therefore less tightly codified, may be equally as likely to lead to positive effect sizes as interventions that are more strictly codified and designed to be faithfully adopted. However, some caution is needed, due to inconsistency in reporting and the notable number of 'not mentioned/unclear' (see Appendix A for details on how this variable was coded). In addition, it is important to note that this finding may appear contradictory to the finding that TA-led interventions that are highly codified are associated with higher effect sizes. This contradiction may at least in part be explained, as teachers may be more able to draw on their professional knowledge and expertise to adapt interventions appropriately to context to maximise effect, whereas research (Sharples et al., 2015) indicates that TA effectiveness is increased when they are trained to use highly structured programmes. In addition, there is association between positive impact and 1:1 and small group tuition, which is also a key feature of the TA-led interventions in this study.

Table 108: Effect size by intended fidelity primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Faithful adoption	52	+0.02	0.010	0.00	+0.04
Adaptation to context	57	+0.05	0.011	+0.02	+0.07
Not mentioned / unclear	24	+0.09	0.030	+0.04	+0.15

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Figure 74: Effect size by intended fidelity primary ITT attainment outcomes

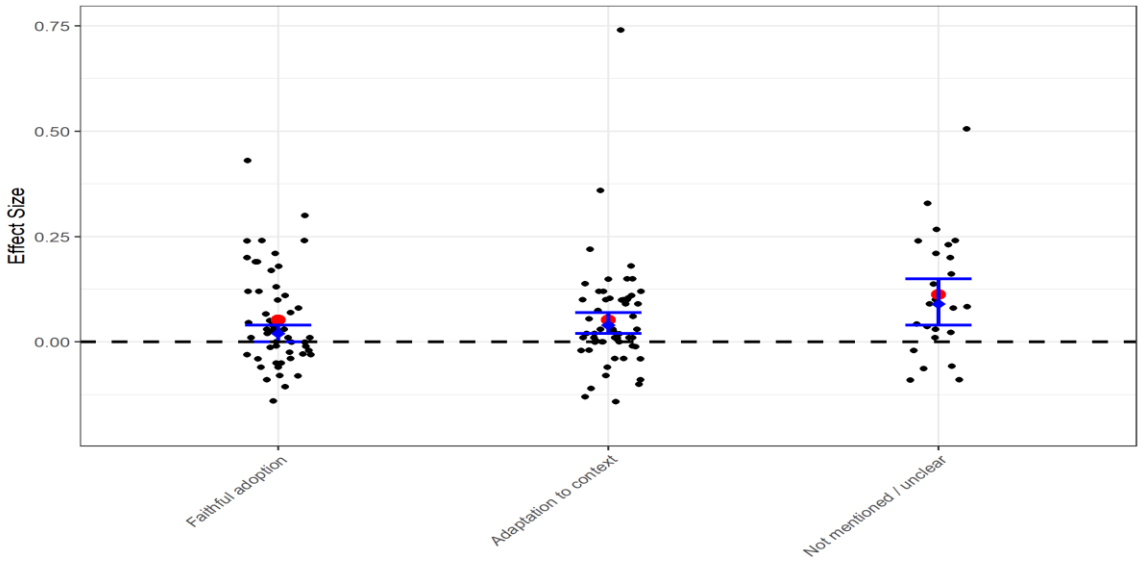


Table 109: Effect size by intended fidelity secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Faithful adoption	36	+0.01	0.012	−0.02	+0.03
Adaptation to context	35	−0.01	0.021	−0.05	+0.03
Not mentioned / unclear	7	+0.04	0.032	−0.03	+0.10

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.01 SD.

Figure 75: Effect size by intended fidelity secondary ITT attainment outcomes

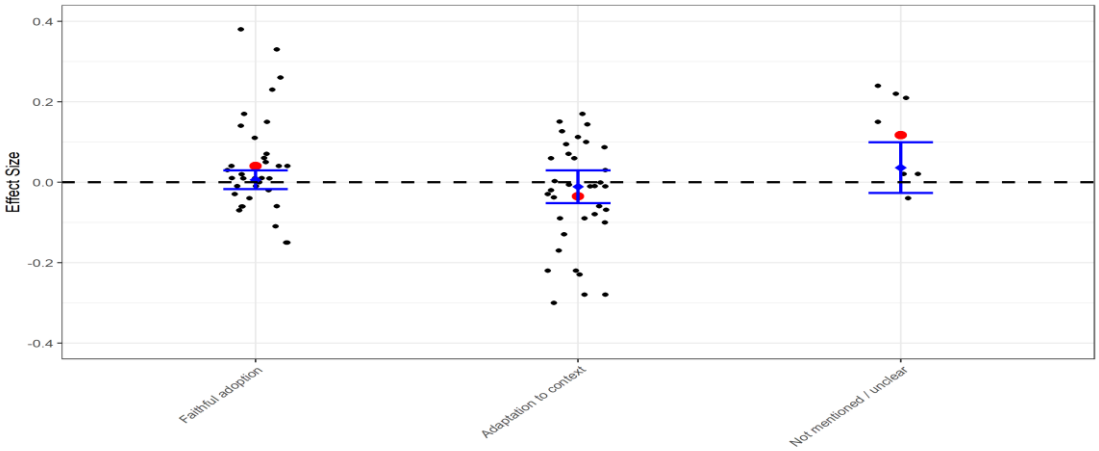
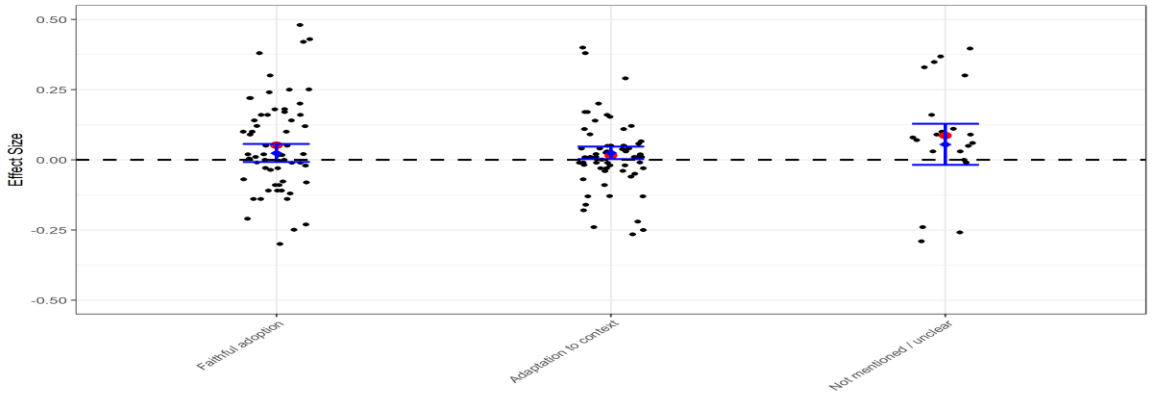


Table 110: Effect size by intended fidelity FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Faithful adoption	60	+0.02	0.016	−0.01	+0.06
Adaptation to context	66	+0.03	0.012	0.00	+0.05
Not mentioned / unclear	23	+0.06	0.037	−0.02	+0.13

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 76: Effect size by intended fidelity FSM attainment outcomes



Actual fidelity

In terms of actual fidelity, similarly little difference was found in terms of effect size according to whether fidelity was high, medium or low and no statistically significant association was observed. The findings are similar for interventions intended to be faithfully adopted and interventions intended to be adapted to context. This is somewhat surprising for the reasons stated above. However, this finding should again be treated with caution because of the issues in coding this variable (please see Appendix A). It should also be noted here that fidelity is being observed in relation to effect size rather than successful implementation; little difference has been found in terms of effect size in relation to fidelity of implementation. However, this does not mean that the same would be true in terms of fidelity relating to successful implementation, which we do not measure here.

Actual implementation fidelity

Primary ITT

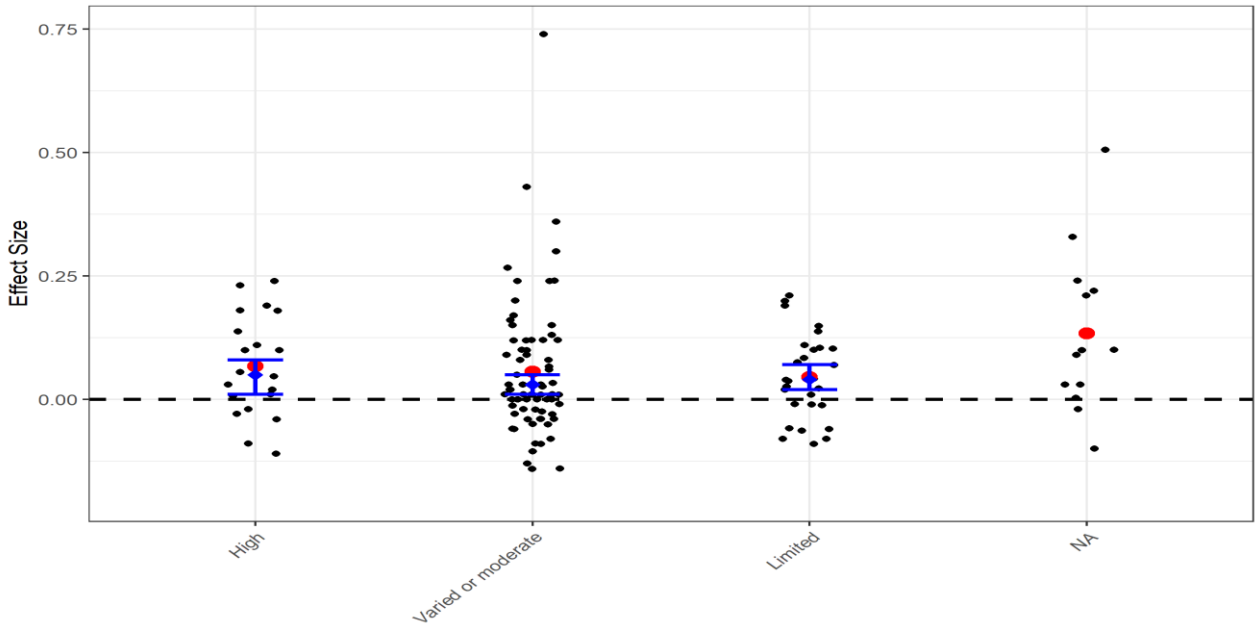
In terms of actual fidelity, similarly little difference was found in terms of effect size according to whether fidelity was high, medium or low, and no statistically significant association was observed. The findings are similar for interventions intended to be faithfully adopted and interventions intended to be adapted to context. This is somewhat surprising for the reasons for the reasons stated above. However, this finding should again be treated with caution because of the issues in coding this variable (please see Appendix A). It should also be noted here that fidelity is being observed in relation to effect size rather than successful implementation; little difference has been found in terms of effect size in relation to fidelity of implementation. However, this does not mean that the same would be true in terms of fidelity relating to successful implementation which we do not measure here.

Table 111: Effect size by actual fidelity primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
High	20	+0.05	0.019	+0.01	+0.09
Varied or moderate	72	+0.03	0.011	+0.01	+0.05
Limited	28	+0.04	0.012	+0.02	+0.07
Not mentioned	13	+0.12	0.046	+0.03	+0.21

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Figure 77: Effect size by actual fidelity primary ITT attainment outcomes



Secondary ITT

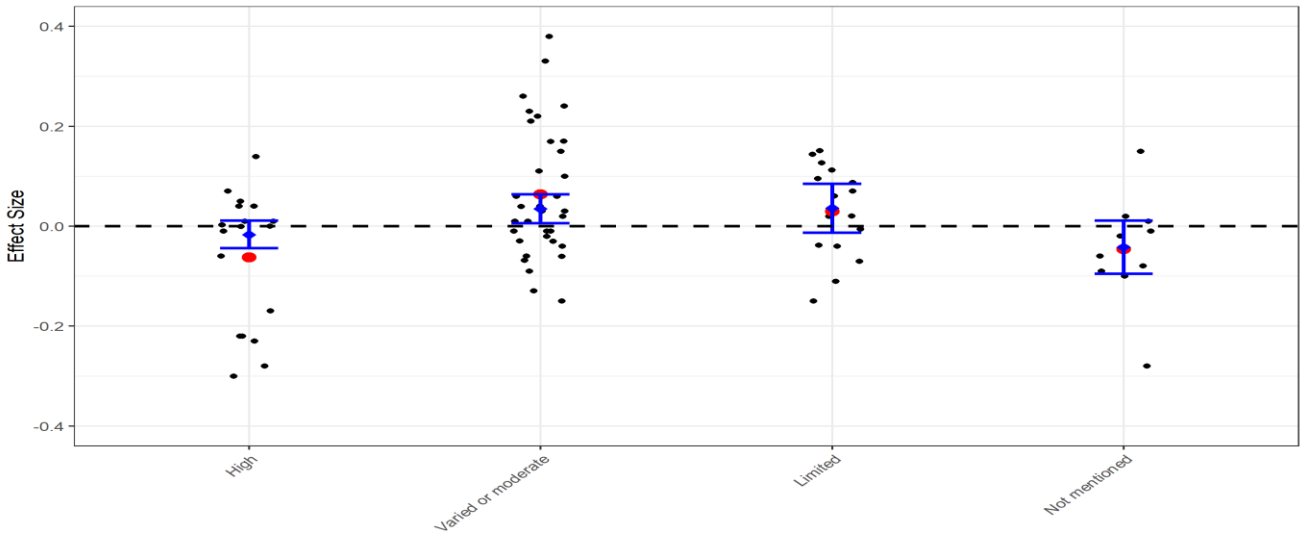
The meta-analyses found the association between actual fidelity and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). The highest weighted mean effect size was also observed where reported actual fidelity was moderate or varied (+0.04 SD; 95% CI: +0.01 to +0.06) or limited (+0.04 SD; 95% CI: −0.01 to +0.09) compared with when actual fidelity was high (−0.02 SD; 95% CI: −0.04 to +0.01) or not mentioned (−0.04 SD; 95% CI: −0.10 to +0.01). Again, this is an unexpected finding with no obvious explanation.

Table 112: Effect size by actual fidelity secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
High	18	−0.02	0.014	−0.04	+0.01
Varied or moderate	34	+0.04	0.015	+0.01	+0.06
Limited	16	+0.04	0.025	−0.01	+0.09
Not mentioned	10	−0.04	0.027	−0.10	+0.01

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.01 SD.

Figure 78: Effect size by actual fidelity secondary ITT attainment outcomes



FSM

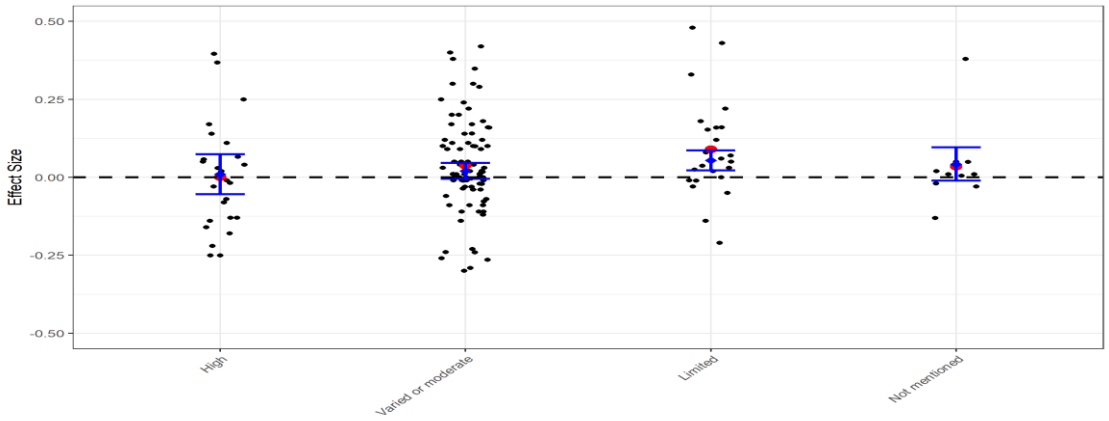
No evidence for an association between FSM effect size and actual fidelity was observed.

Table 113: Effect size by actual fidelity FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
High	25	+0.01	0.033	−0.05	+0.07
Varied or moderate	89	+0.02	0.013	−0.01	+0.05
Limited	25	+0.05	0.016	+0.02	+0.09
Not mentioned	10	+0.04	0.027	−0.01	+0.10

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 79: Effect size by actual fidelity FSM attainment outcomes



Effect sizes and evaluation design

Summary

Table 114: Summary of meta-analyses of ITT effect sizes and evaluation design

Subtheme	Explanatory variable	Primary ITT	Secondary ITT	FSM
Trial design description	Type of trial (RCT/CRT)	✓***	✓#	✓#
	Level of randomisation	✓**	✓**	✓#
	Efficacy/effectiveness	✓	✓**	✓#
	Type of evaluator	✓	✓***	✓#
Size and length of intervention	Intervention length (weeks)	✓#	✓***	✓***
	Number of schools	✓#	✓***	✓*
	Number of pupils	✓***	✓#	✓#
Statistical sensitivity, attrition and trial quality	Statistical sensitivity (MDES estimate)	✓***		
	Pupil level % attrition	✓		
	Trial quality (EEF padlocks)	✓#	✓***	✓***
Evaluation burden	Testing burden	✓	✓	✓
	IPE data collection burden	✓#	✓	✓
Primary outcome	Type (commercial / statutory / other)	✓#	✓**	✓
	Outcome curriculum area	✓#		
	Number of primary outcomes	✓		
	Alignment of intervention and primary outcome	✓***		

11 of the 16 explanatory variables included in the evaluation design theme were included in the meta-analyses of secondary ITT and FSM attainment outcomes. Five variables were dropped because they focused specifically on a measure of the primary outcome: MDES estimate; pupil level attrition; outcome curriculum area, number of primary outcomes and the alignment of intervention and primary outcome.

Trial design description

Primary ITT

On average, RCTs were associated with a statistically significantly ($p < 0.01$) higher effect size (weighted mean = +0.10 SD; 95% CI: +0.05 to +0.14) compared with CRTs (weighted mean = +0.03 SD; 95% CI: +0.01 to +0.04) (Table 115 and Figure 80). RCTs with pupil-level randomisation had an even higher weighted mean effect size (weighted mean = +0.11 SD; 95% CI: +0.06 to +0.17) compared with CRTs with school-level randomisation (weighted mean = +0.03 SD; 95% CI: +0.01 to +0.05) (Table 116 and Figure 81).

Table 115: Effect size by trial design primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
RCT	41	+0.10	0.023	+0.05	+0.14
Clustered RCT (CRT)	92	+0.03	0.008	+0.01	+0.05

Meta *p*-value < 0.01***. Overall weighted mean = +0.04 SD.

Figure 80: Effect size by trial design primary ITT attainment outcomes

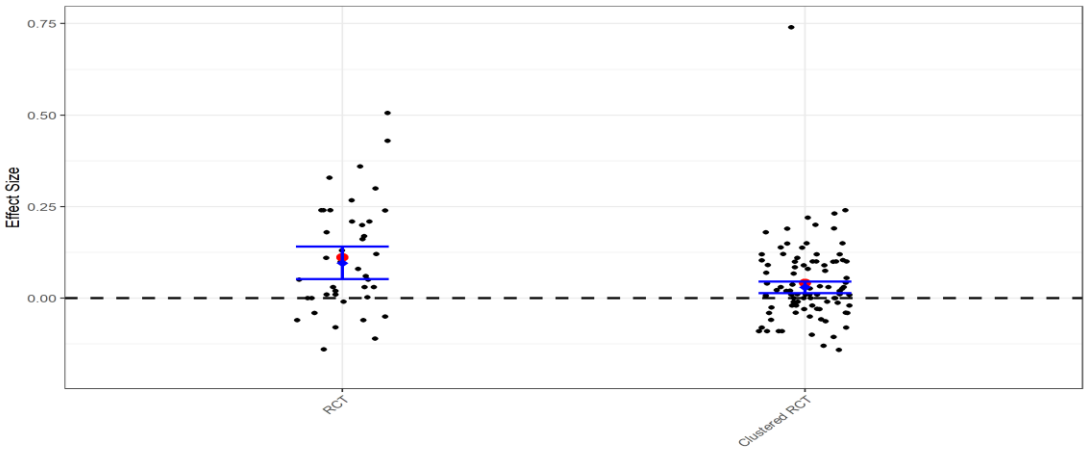
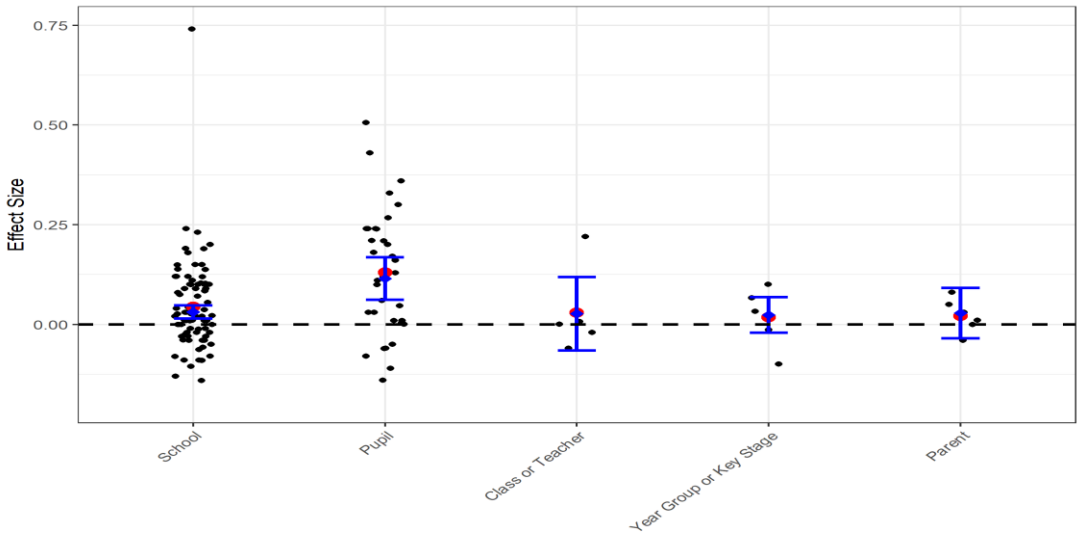


Table 116: Effect size by level of randomisation primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
School	82	+0.04	0.009	+0.02	+0.05
Pupil	34	+0.12	0.027	+0.06	+0.17
Class or teacher	4	−0.02	0.022	−0.06	+0.03
Key Stage or year group	5	+0.02	0.023	−0.02	+0.07
Parent	7	+0.03	0.032	−0.04	+0.09
Other/complex	0	–	–	–	–

Meta *p*-value < 0.05**. Overall weighted mean = +0.04 SD.

Figure 81: Effect size by level of randomisation primary ITT attainment outcomes



Secondary ITT

RCTs also had higher secondary ITT effect size (weighted mean = +0.07 SD; 95% CI: −0.01 to +0.15) compared with CRTs (weighted mean = 0.00 SD; 95% CI: −0.02 to +0.02) but this difference was not statistically significant (Table 117 and Figure 82). RCTs with pupil-level randomisation had a slightly higher weighted mean secondary ITT effect size (weighted mean = +0.02 SD; 95% CI: −0.07 to +0.11) compared with other levels of randomisation (+0.01 SD or lower) and these differences were statistically significant at the 5% level ($p < 0.05$) (Table 118 and Figure 83).

Table 117: Effect size by trial design secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
RCT	13	+0.07	0.042	−0.01	+0.15
Clustered RCT (CRT)	65	0.00	0.011	−0.02	+0.02

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.01 SD.

Figure 82: Effect size by trial design secondary ITT attainment outcomes

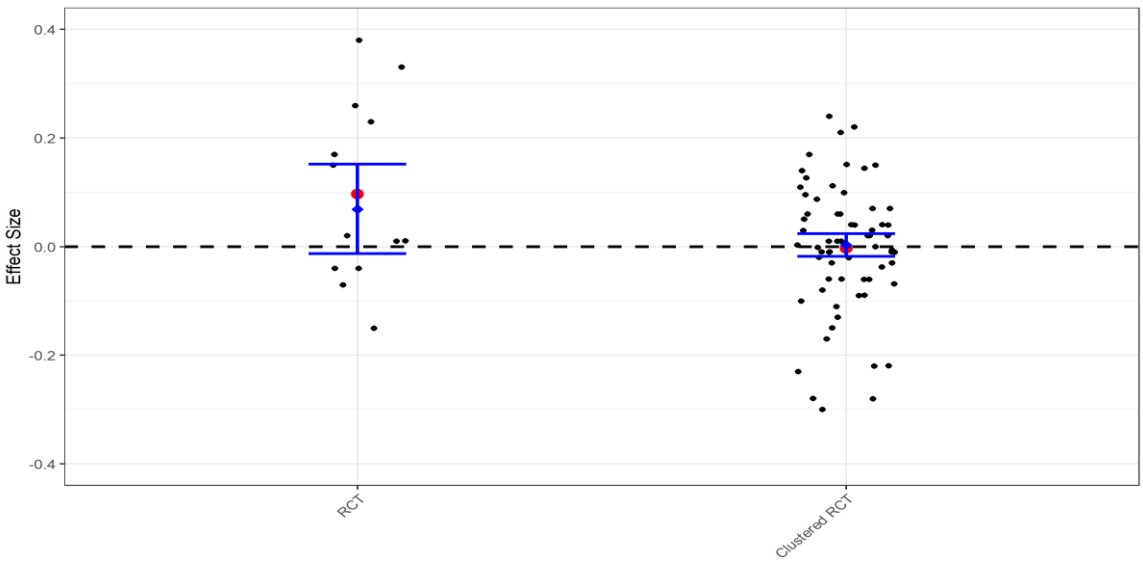
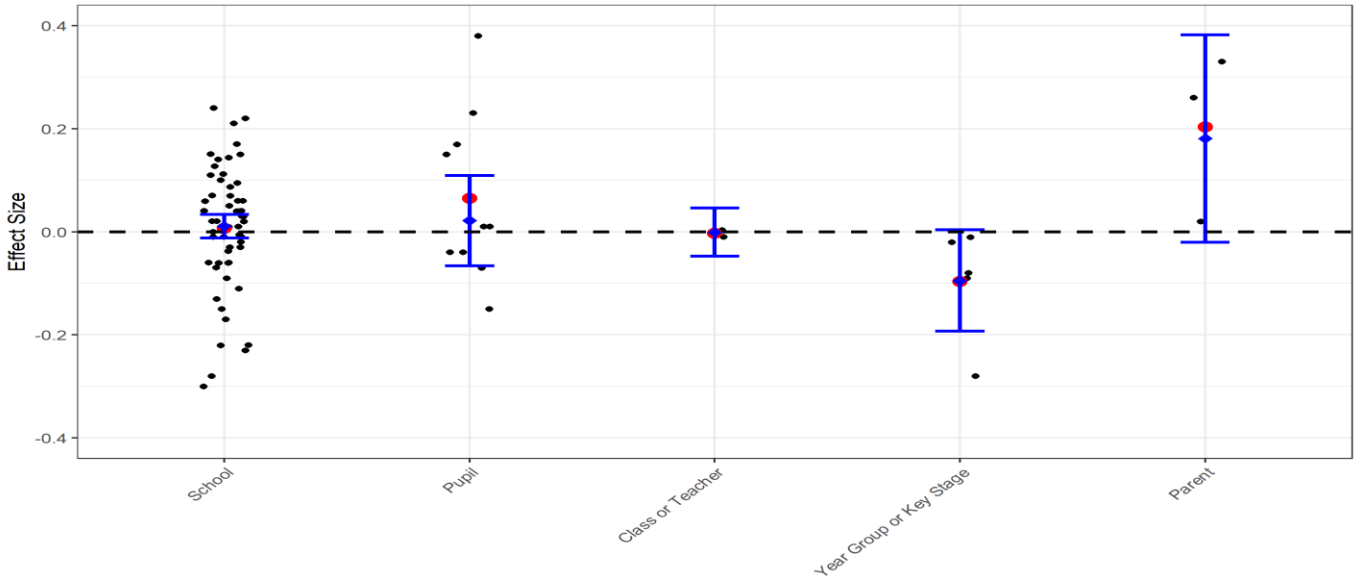


Table 118: Effect size by level of randomisation secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
School	56	+0.01	0.012	−0.01	+0.03
Pupil	10	+0.02	0.045	−0.07	+0.11
Class or teacher	3	–	–	–	–
Key Stage or year group	6	−0.10	0.050	−0.19	0.00
Parent	3	–	–	–	–
Other/complex	0	–	–	–	–

Meta *p*-value < 0.05**. Overall weighted mean = +0.01 SD.

Figure 83: Effect size by level of randomisation secondary ITT attainment outcomes



FSM

RCTs also had higher FSM effect size (weighted mean = +0.04 SD; 95% CI: 0.00 to +0.09) compared with CRTs (weighted mean = +0.02 SD; 95% CI: 0.00 to +0.05) but this difference was not statistically significant (Table 119 and Figure 84). RCTs with pupil-level randomisation had a slightly higher weighted mean FSM effect size (weighted mean = +0.06 SD; 95% CI: 0.00 to +0.12) compared with other levels of randomisation (+0.03 SD or lower) but these differences were also not statistically significant (Table 120 and Figure 85).

Table 119: Effect size by trial design FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
RCT	50	+0.04	0.022	0.00	+0.09
Clustered RCT (CRT)	99	+0.02	0.011	0.00	+0.05

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 84: Effect size by trial design FSM attainment outcomes

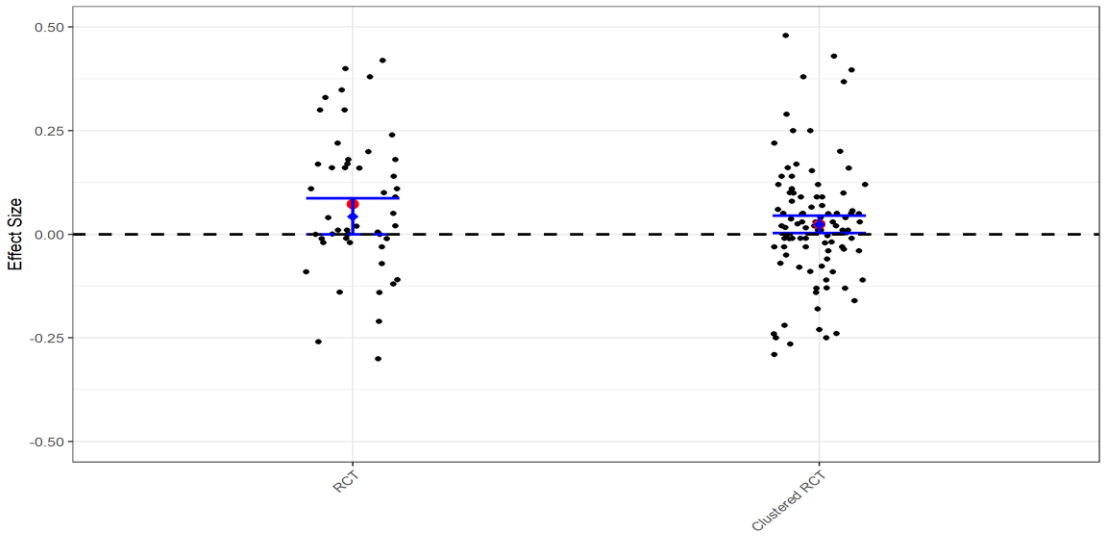
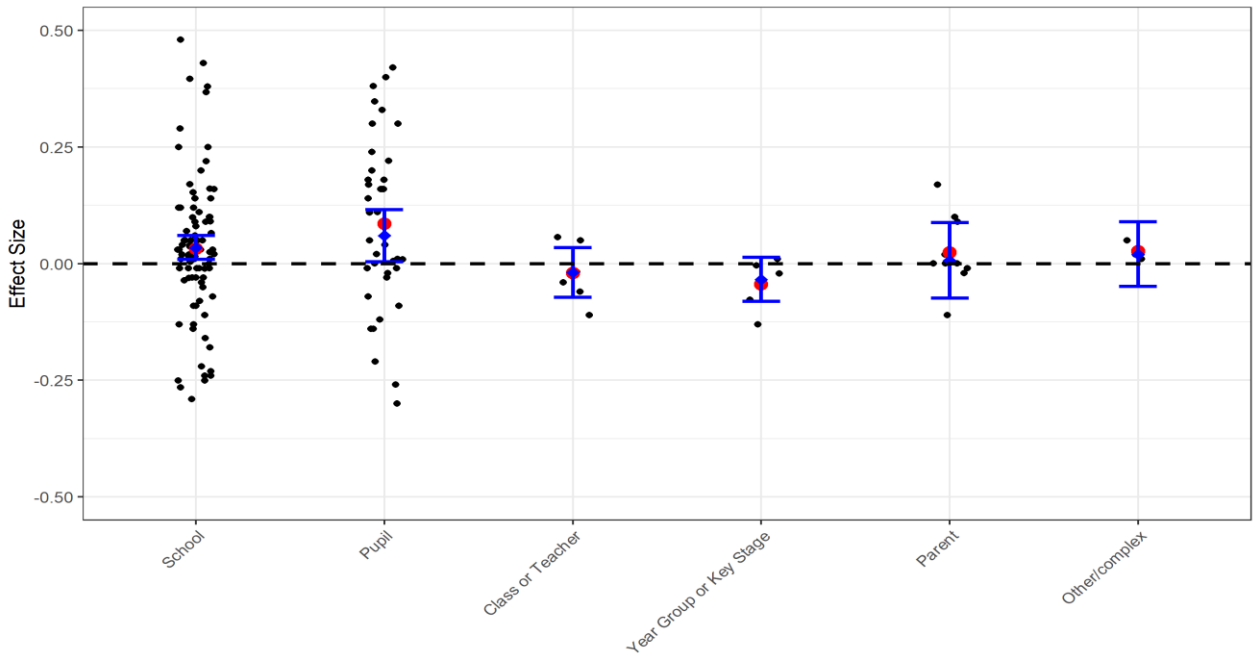


Table 120: Effect size by level of randomisation FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
School	85	+0.03	0.013	+0.01	+0.06
Pupil	40	+0.06	0.028	0.00	+0.12
Class or teacher	6	−0.02	0.027	−0.07	+0.03
Year group or Key Stage	5	−0.03	0.024	−0.08	+0.01
Parent	10	+0.01	0.041	−0.07	+0.09
Other/complex	3	–	–	–	–

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 85: Effect size by level of randomisation FSM attainment outcomes



Efficacy and effectiveness trials

Primary ITT

No evidence of an association between type of trial (efficacy vs effectiveness) and effect size was observed.[31]

Secondary ITT

The meta-analyses found the association between type of trial and secondary ITT effect size to be statistically significant at the 5% level ($p < 0.05$). Efficacy trials were observed to have a higher weighted mean effect size (weighted mean = +0.04 SD; 95% CI: +0.01 to +0.07) compared with effectiveness trials (weighted mean = −0.01 SD; 95% CI: −0.04 to +0.02).

FSM

No evidence of an association between type of trial and FSM effect sizes was observed.

Type of evaluator

Primary ITT

No evidence of an association between the type of evaluator (University or non-University) and effect size was observed.

Secondary ITT

The meta-analyses found the association between type of evaluator and secondary ITT effect size to be statistically significant at the 1% level ($p < 0.01$). Non-university evaluators were observed to have a higher weighted mean effect size (weighted mean = +0.07 SD; 95% CI: +0.04 to +0.10) compared with university evaluators (weighted mean = −0.01 SD; 95% CI: −0.03 to +0.01).

FSM

No evidence of an association between type of evaluator and FSM effect sizes was observed.

Length and size of the intervention

Intervention length

Primary ITT

The association between effect size and intervention length is complex but not statistically significant. On average, lower effect sizes are observed for evaluations of relatively long or short interventions. The smallest weighted mean effect size is observed for evaluations of interventions lasting for more than one year (= +0.01 SD; 95% CI: −0.01 to +0.03) and the largest is observed for trials lasting for between 16 and up to 30 weeks (weighted mean = +0.08 SD; 95% CI: +0.03 to +0.13).

Table 121: Effect size by intervention length primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Up to 15 weeks (one term)	37	+0.04	0.014	+0.01	+0.07
16–30 weeks (two terms)	31	+0.08	0.025	+0.03	+0.13
31–45 weeks (three terms / one year)	39	+0.05	0.014	+0.02	+0.08
More than 45 weeks	26	+0.01	0.011	−0.01	+0.03

[31] Note: the definition of efficacy and effectiveness trials was provided by EEF. Some inconsistency in this classification is noted in the *Presenting the explanatory variables* section.

Meta *p*-value > 0.10. Overall weighted mean = +0.04 SD.

Secondary ITT

The association between secondary ITT effect size and intervention length is complex and statistically significant at the 1% level ($p < 0.01$). On average, the largest mean effect sizes are observed for interventions that lasted 31–45 weeks (+0.06 SD; 95% CI: +0.02 to +0.10). Interventions that were shorter had a mean effect size of +0.03 SD or less whilst those lasting longer than a year had a mean effect size of −0.02 SD.

Table 122: Effect size by intervention length secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Up to 15 weeks (one term)	17	+0.03	0.020	−0.01	+0.07
16–30 weeks (two terms)	16	−0.04	0.020	−0.08	0.00
31–45 weeks (three terms / one year)	18	+0.06	0.020	+0.02	+0.10
More than 45 weeks	27	−0.02	0.016	−0.05	+0.01

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

FSM

The association between FSM effect size and intervention length is complex and statistically significant at the 1% level ($p < 0.01$). On average, the largest FSM mean effect sizes are observed for interventions that lasted around one year (+0.07 SD; 95% CI: +0.03 to +0.11). Interventions that were shorter had a mean effect size of +0.03 SD or less, whilst those lasting longer than a year had a mean effect size of 0.00 SD.

Table 123: Effect size by intervention length FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Up to 15 weeks (one term)	45	+0.03	0.014	0.00	+0.06
16–30 weeks (two terms)	31	−0.01	0.027	−0.06	+0.05
31–45 weeks (three terms / one year)	34	+0.07	0.020	+0.03	+0.11
More than 45 weeks	39	0.00	0.014	−0.03	+0.03

Meta *p*-value < 0.01***. Overall weighted mean = +0.03 SD.

Size of trial (number of schools)

Primary ITT

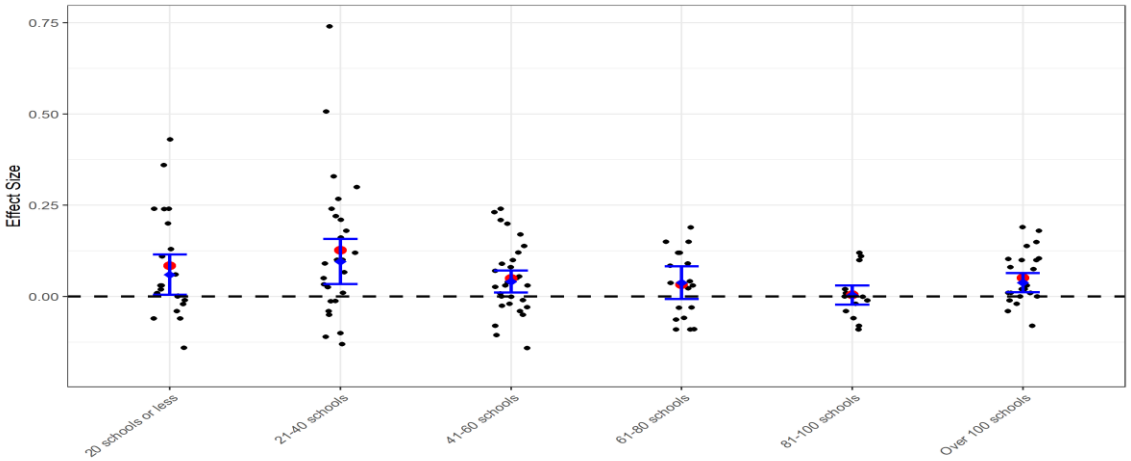
No evidence of an association between number of schools and primary ITT effect size was observed.

Table 124: Effect size by number of schools primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
20 or less	21	+0.06	0.028	+0.01	+0.12
21–40	26	+0.10	0.032	+0.03	+0.16
41–60	30	+0.04	0.015	+0.01	+0.07
61–80	18	+0.04	0.023	−0.01	+0.08
81–100	15	0.00	0.013	−0.02	+0.03
101 or more	23	+0.04	0.013	+0.01	+0.06

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Figure 86: Effect size by number of schools primary ITT attainment outcomes



Secondary ITT

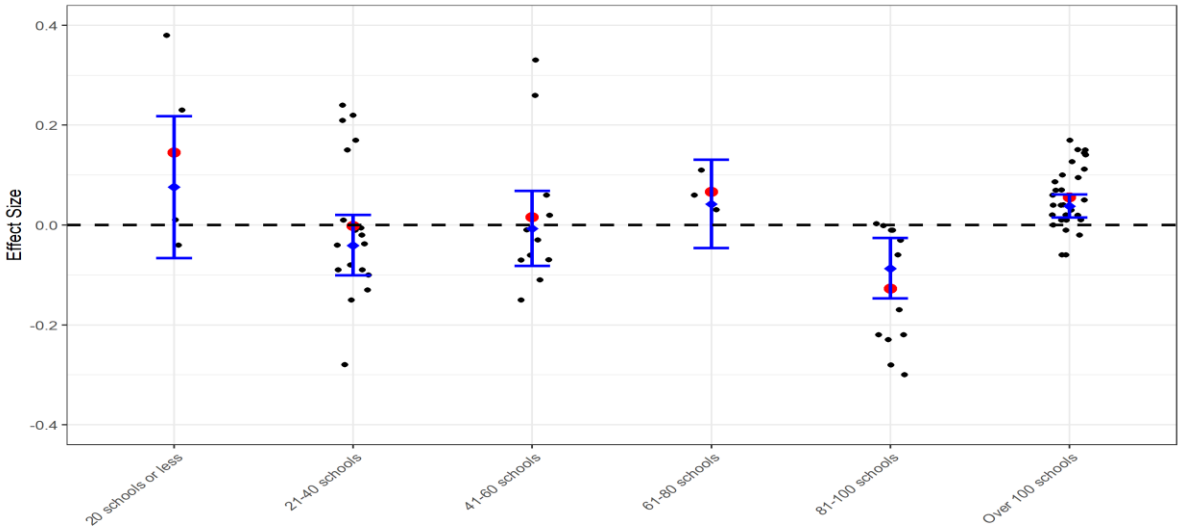
The association between secondary ITT effect size and number of schools is complex and statistically significant at the 1% level ($p < 0.01$). On average, the largest mean effect sizes are observed for interventions that involved 20 or fewer schools[32] (weighted mean = +0.08 SD; 95% CI: −0.07 to +0.22) or over 100 schools (weighted mean = +0.04 SD; 95% CI: +0.02 to +0.06). Between these extremes, the weighted mean effect sizes were negative (−0.01 to −0.09 SD).

Table 125: Effect size by number of schools secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
20 or less	4	+0.08	0.072	−0.07	+0.22
21–40	18	−0.04	0.031	−0.10	+0.02
41–60	11	−0.01	0.038	−0.08	+0.07
61–80	3	–	–	–	–
81–100	12	−0.09	0.031	−0.15	−0.03
101 or more	30	+0.04	0.012	+0.02	+0.06

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

Figure 87: Effect size by number of schools secondary ITT attainment outcomes



[32] Note that this only related to four of the 78 secondary ITT effect sizes.

FSM

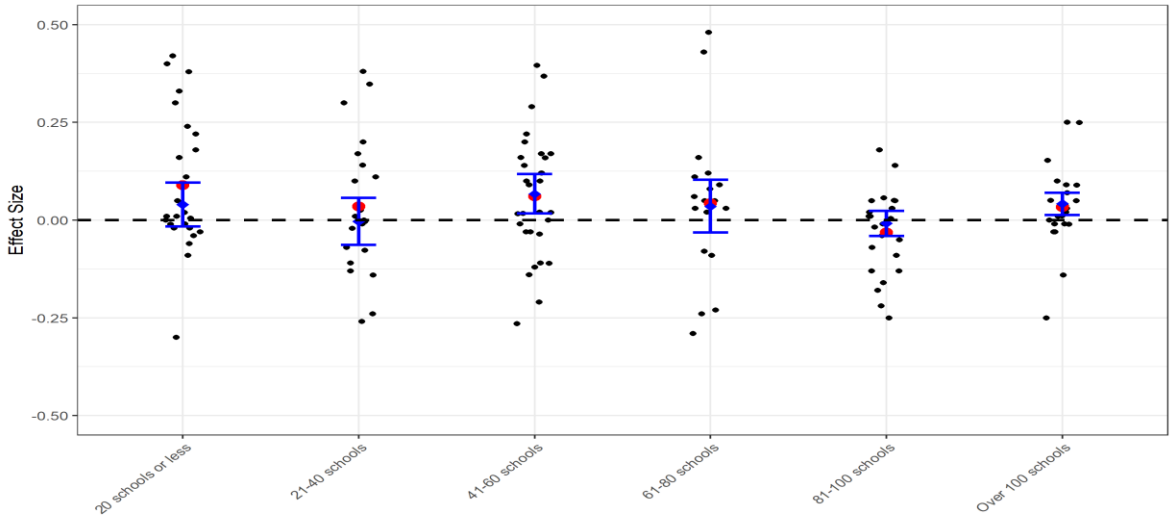
The association between FSM effect size and number of schools is complex and statistically significant at the 10% level ($p < 0.10$). On average, the largest mean FSM effect sizes are observed for interventions that involved 41–60 schools (+0.07 SD; 95% CI: +0.02 to +0.12). Evaluations that involved a fewer or greater number of schools had a mean FSM effect size of +0.04 SD or less.

Table 126: Effect size by number of schools FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
20 or less	27	+0.04	0.029	−0.02	+0.10
21–40	25	0.00	0.031	−0.06	+0.06
41–60	31	+0.07	0.026	+0.02	+0.12
61–80	19	+0.04	0.034	−0.03	+0.10
81–100	24	−0.01	0.017	−0.04	+0.02
101 or more	23	+0.04	0.014	+0.01	+0.07

Meta *p*-value < 0.10*. Overall weighted mean = +0.03 SD.

Figure 88: Effect size by number of schools FSM attainment outcomes



Size of trial (number of pupils)

Primary ITT

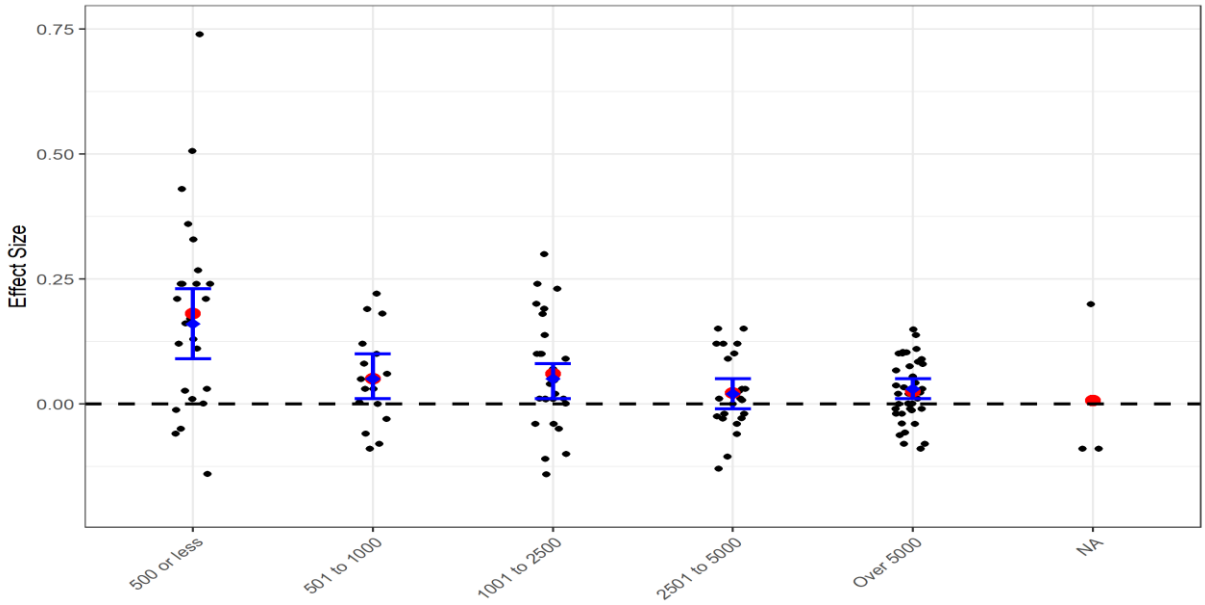
The association between primary ITT effect size and number of pupils was statistically significant at the 1% level ($p < 0.01$). On average, the largest mean effect sizes are observed for trials that involved 500 or fewer pupils (weighted mean = +0.16 SD; 95% CI: +0.09 to +0.23). Smaller weighted mean effect sizes were observed for larger trials (+0.06 SD or lower).

Table 127: Effect size by number of pupils primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
500 or less	25	+0.16	0.036	+0.09	+0.23
501–1000	16	+0.06	0.023	+0.01	+0.10
1001–2500	27	+0.05	0.019	+0.02	+0.09
2501–5000	23	+0.02	0.015	−0.01	+0.05
Over 5000	39	+0.03	0.009	+0.01	+0.05

Meta *p*-value < 0.01***. Overall weighted mean = +0.04 SD.

Figure 89: Effect size by number of pupils primary ITT attainment outcomes



Secondary ITT

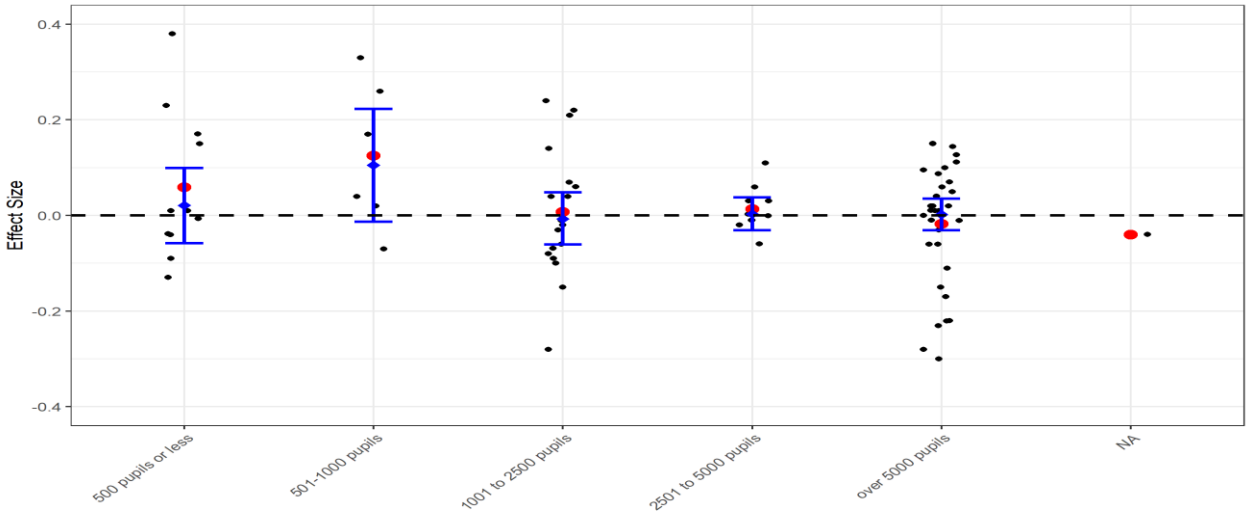
The association between secondary ITT effect size and number of pupils was not statistically significant.

Table 128: Effect size by number of pupils secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
500 or less	11	+0.02	0.040	−0.06	+0.10
501–1000	6	+0.11	0.060	−0.01	+0.22
1001–2500	18	−0.01	0.028	−0.06	+0.05
2501–5000	10	0.00	0.018	−0.03	+0.04
Over 5000	32	0.00	0.017	−0.03	+0.04

Meta *p*-value >0.10 (NS). Overall weighted mean = +0.01 SD.

Figure 90: Effect size by number of pupils secondary ITT attainment outcomes



FSM

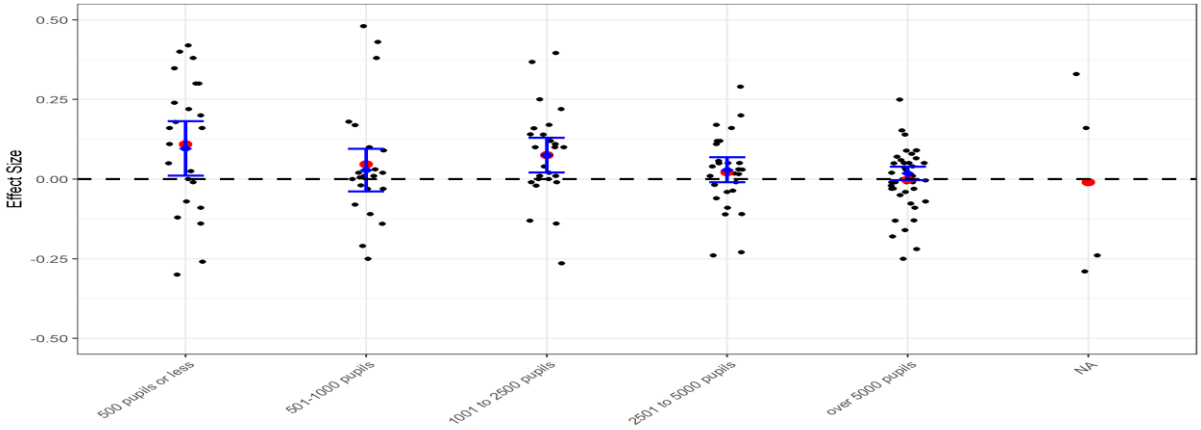
The association between FSM effect size and number of pupils was not statistically significant.

Table 129: Effect size by number of pupils FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
500 or less	26	+0.10	0.044	+0.01	+0.18
501–1000	23	+0.03	0.034	−0.04	+0.10
1001–2500	25	+0.08	0.028	+0.02	+0.13
2501–5000	28	+0.03	0.020	−0.01	+0.07
Over 5000	43	+0.02	0.011	0.00	+0.04

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.03 SD.

Figure 91: Effect size by number of pupils FSM attainment outcomes



Statistical sensitivity, attrition and trial quality

Statistical sensitivity (MDES)

The association between primary effect size and statistical sensitivity (MDES) is complex and statistically significant ($p < 0.01$). On average, higher effect sizes are observed for trials with an MDES between +0.25 and less than +0.30 (weighted mean effect size = +0.12 SD; 95% CI: +0.08 to +0.16) compared with trials with higher MDES estimates (weighted mean = +0.08 SD; 95% CI: −0.01 to +0.18) or lower MDES estimates (weighted mean = +0.03 SD or lower). This variable was not included in the meta-analyses of secondary or FSM attainment effect sizes.

Figure 92: Statistical sensitivity of trial design (MDES estimates)

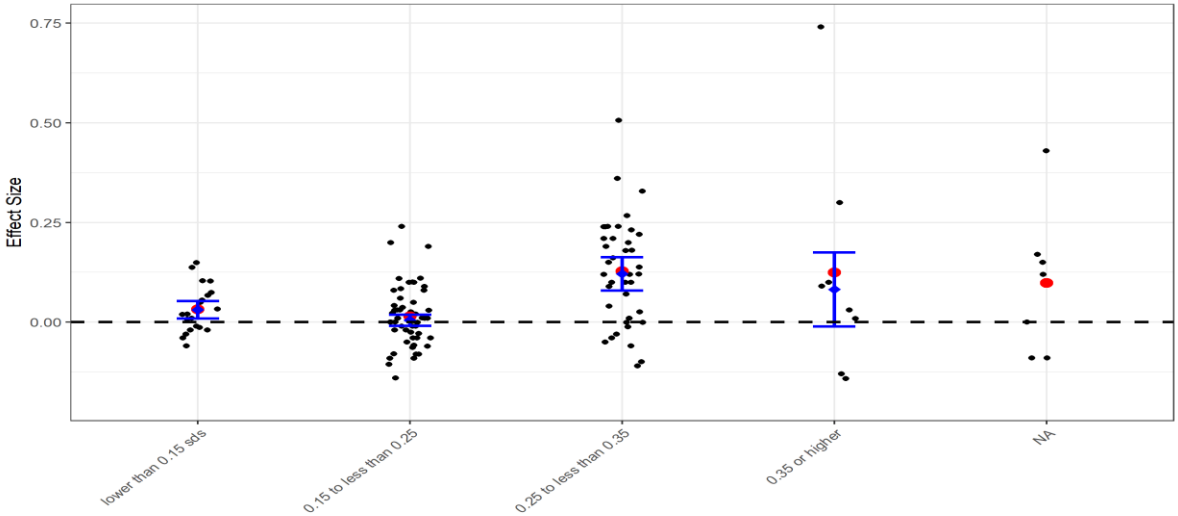


Table 130: Statistical sensitivity of trial design (MDES estimates) primary ITT effect sizes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
<0.15 SD	21	+0.03	0.01	+0.01	+0.05
0.15 to < 0.25 SD	56	0.00	0.01	−0.01	+0.02
0.25 to < 0.35 SD	40	+0.12	0.02	+0.08	+0.16
0.35 SD or higher	9	+0.08	0.05	−0.01	+0.18

Meta *p*-value < 0.01.

No evidence of an association between pupil-level attrition and primary ITT effect size was observed.

Trial quality (EEF padlocks)

Primary ITT

Other than a notably high weighted mean effect size for the four effect sizes reported by three trials with a zero EEF padlock rating (= +0.20 SD; 95% CI: +0.06 to +0.35), a relationship between effect size and EEF padlocks is not evident.

Table 131: Effect size by EEF padlocks primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Zero padlocks	4	+0.21	0.074	+0.06	+0.35
One padlock	9	−0.01	0.030	−0.06	+0.05
Two padlocks	30	+0.06	0.023	+0.01	+0.10
Three padlocks	44	+0.06	0.014	+0.03	+0.08
Four padlocks	35	+0.03	0.013	+0.01	+0.06
Five padlocks	11	+0.02	0.013	−0.01	+0.04

Meta *p*-value > 0.10 (NS). Overall weighted mean = +0.04 SD.

Secondary ITT

The association between secondary ITT effect size and trial quality was statistically significant at the 1% level ($p < 0.01$). On average, the largest mean effect sizes are observed for trials with two padlocks (weighted mean = +0.08 SD; 95% CI: +0.05 to +0.12). Smaller weighted mean effect sizes were observed for trials with more padlocks (+0.01 SD or lower).

Table 132: Effect size by EEF padlocks secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Zero padlocks	3	–	–	–	–
One padlock	3	–	–	–	–
Two padlocks	15	+0.08	0.018	+0.05	+0.12
Three padlocks	26	−0.01	0.019	−0.05	+0.03
Four padlocks	14	−0.09	0.040	−0.17	−0.01
Five padlocks	17	+0.01	0.012	−0.01	+0.04

Meta *p*-value < 0.01***. Overall weighted mean = +0.01 SD.

FSM

The association between FSM effect size and trial quality was statistically significant at the 5% level ($p < 0.05$). On average, the largest mean effect sizes are observed for trials with three padlocks (weighted mean = +0.05 SD; 95% CI: +0.02 to +0.09). Smaller weighted mean effect sizes were observed for trials with more padlocks (+0.03 SD or lower).

Table 133: Effect size by EEF padlocks FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Zero padlocks	3	–	–	–	–
One padlock	8	−0.12	0.037	−0.19	−0.05
Two padlocks	22	+0.02	0.029	−0.04	+0.08
Three padlocks	49	+0.05	0.018	+0.02	+0.09
Four padlocks	50	+0.01	0.015	−0.02	+0.04
Five padlocks	17	+0.03	0.013	0.00	+0.05

Meta *p*-value < 0.05**. Overall weighted mean = +0.03 SD.

Testing burden

Primary ITT

No evidence of an association between testing burden and primary ITT effect size was observed.

On average, evaluations with low IPE data collection burdens were associated with a higher effect size (weighted mean = +0.10 SD; 95% CI: +0.02 to +0.18) than evaluations with medium or high IPE data collection burdens (weighted mean = +0.04 SD for both), but this difference was not statistically significant.

Secondary ITT

No evidence of an association between testing or IPE burden and secondary ITT effect size was observed.

FSM

No evidence of an association between testing or IPE burden and FSM effect size was observed.

Alignment of primary outcome(s) and intervention

No evidence of an association between number of primary outcomes and primary ITT effect size was observed. This variable was not included in the meta-analyses of secondary or FSM attainment effect sizes.

On average, when there was a direct match between the intervention focus and primary outcome(s), the effect size was statistically significantly (p < 0.01) higher (weighted mean = +0.09 SD; 95% CI: +0.06 to +0.12) than where the match was not direct (weighted mean = +0.03 SD or lower).

Figure 93: Effect size by alignment between intervention and primary outcome

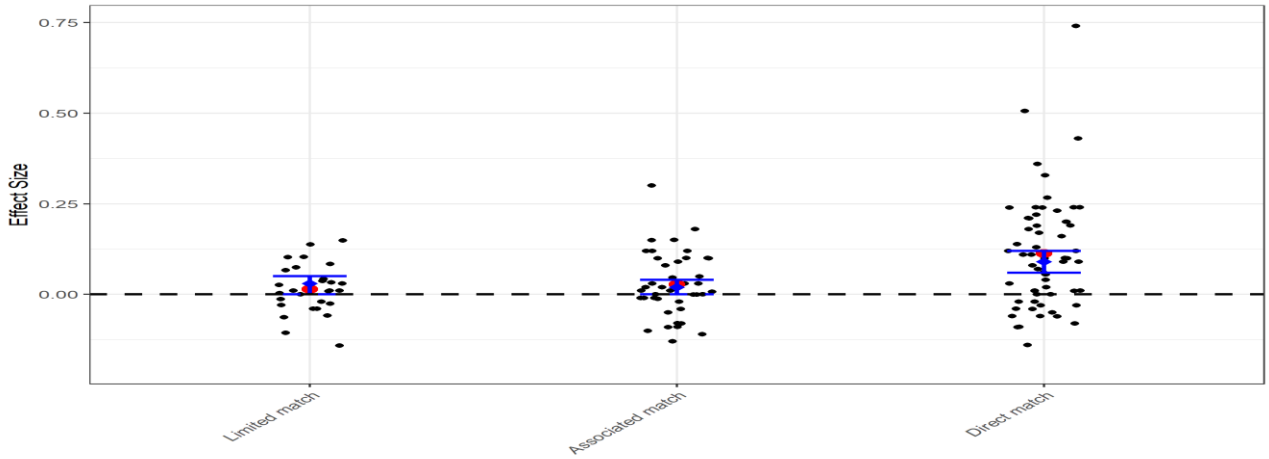


Table 134: Effect size by alignment between intervention and primary outcome

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Limited match	30	+0.03	0.01	0.00	+0.05
Associated match	43	+0.02	0.01	0.00	+0.04
Direct match	60	+0.09	0.02	+0.06	+0.12

Meta *p*-value < 0.01.

On average, commercial outcomes were associated with a higher effect size (weighted mean = +0.05 SD; 95% CI: +0.03 to +0.08) than statutory outcomes (weighted mean = +0.02 SD; 95% CI: 0.00 to +0.05), although this trend was not statistically significant.

Table 135: Effect size by type of primary outcome

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Commercial test	79	+0.05	0.01	+0.03	+0.08
Statutory	45	+0.02	0.01	0.00	+0.05
Other/mixed	9	+0.08	0.03	+0.03	+0.14

Meta *p*-value = 0.25.

Type of outcome

Primary ITT

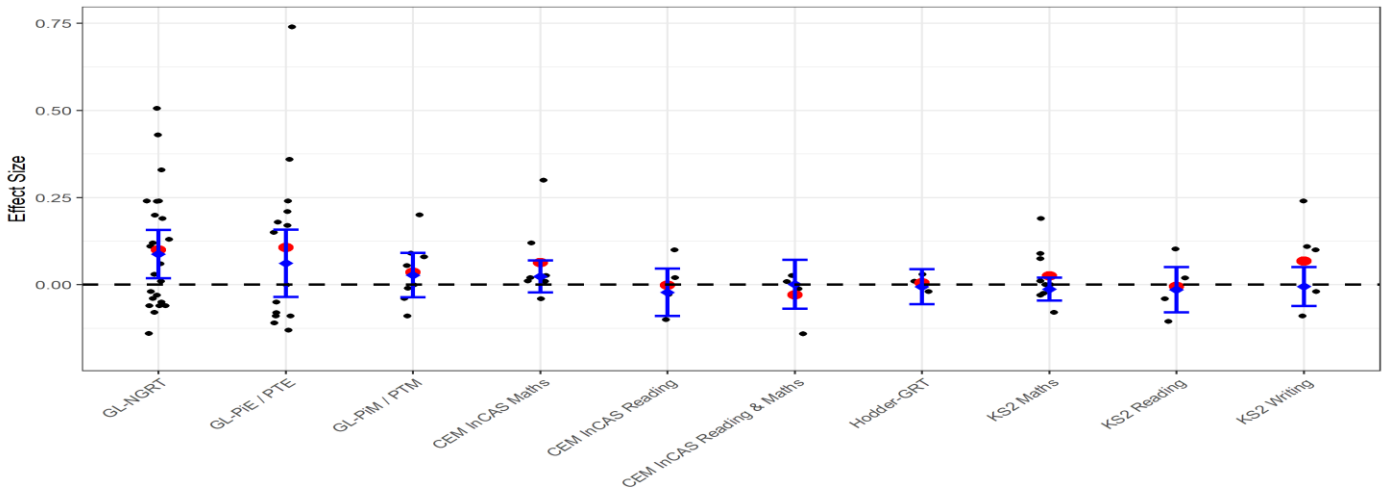
The association between primary ITT effect size and type of attainment outcome (commercial or statutory) was statistically significant at the 10% level ($p < 0.10$). On average, commercial outcomes were associated with a higher effect size (weighted mean = +0.06 SD; 95% CI: +0.03 to +0.08) compared with statutory or other outcomes (weighted mean = +0.02 SD; 95% CI: 0.00 to +0.05).

Table 136: Effect size by type of outcome/test primary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Commercial test	81	+0.06	0.012	+0.03	+0.08
Statutory	43	+0.02	0.011	0.00	+0.05
Other / mixed	9	+0.08	0.030	+0.03	+0.14

Meta *p*-value < 0.10*. Overall weighted mean = +0.04 SD.

Figure 94: Effect size by primary outcomes [10 most common primary outcomes]



Within the meta-analyses of primary ITT effect sizes, the types of outcome were examined more closely (Table 137). For commercial outcomes, on average, effect sizes range between a weighted mean of +0.01 SD (CEM) and +0.08 SD (GL Assessments). For statutory test outcomes, on average, effect sizes range between a weighted mean of +0.03 SD (KS4 attainment) and +0.04 SD (KS2 attainment).

Looking at the 10 most common specific outcomes[33] that accounted for 82 of the 133 effect sizes in the review (62%), GL NGRT was the most common outcome (23 effect sizes) and had the highest observed weighted mean effect size (+0.09 SD; ;95% CI: +0.02 to +0.16). Smaller weighted mean effect sizes were observed for GL PiE/PTE (+0.06 SD; 95% CI: −0.03 to +0.16) and GL PiM/PTM (+0.03 SD; 95% CI: −0.04 to +0.09). Across the other seven specific commercial and statutory test outcomes, the weighted mean effect size was +0.02 SD or lower. The smallest weighted mean effect size was −0.02 SD for CEM InCAS reading (5 effect sizes).

Table 137: Effect size by type of primary outcomes [10 most common primary outcomes]

		n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Commercial*	GL NGRT	23	+0.09	0.04	+0.02	+0.16
	GL PiE / PTE	14	+0.06	0.05	−0.03	+0.16
	GL PiM / PTM	8	+0.03	0.03	−0.04	+0.09
	CEM InCAS maths	7	+0.02	0.02	−0.02	+0.07
	CEM InCAS reading	5	−0.02	0.03	−0.09	+0.05
	CEM incas reading and maths	4	0.00	0.04	−0.07	+0.07
	Hodder GRT	4	−0.01	0.03	−0.06	+0.04
Statutory*	KS2 maths	9	−0.01	0.02	−0.05	+0.02
	KS2 reading	5	−0.01	0.03	−0.08	+0.05
	KS2 writing	5	−0.01	0.03	−0.06	+0.05

***Notes**: Meta *p*-value n/a (data too sparse); 'other commercial' (*n* = 16); GCSE maths, GCSE English, GCSE overall (all *n* < 4) not shown**.**

Secondary ITT

The association between secondary ITT effect size and type of attainment outcome (commercial or statutory) was statistically significant at the 5% level ($p < 0.05$). On average, higher effect sizes are observed when a statutory secondary ITT attainment outcome was used (weighted mean effect size = +0.02 SD; 95% CI: 0.00; +0.04) compared with when a commercial outcome was used (weighted mean effect size = −0.04 SD; 95% CI: −0.08; +0.01).

[33] Effect sizes relating to KS4/GCSE maths, English or overall are not shown because data is too sparse (*n* < 4). Also, whilst these outcomes are 'specific' in terms of their name, the age range will vary. For example, GL PiM/PTM will be the GL maths attainment test for pupils in Y6, Y7, Y9 etc.

Table 138: Effect size by type of outcome/test secondary ITT attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Commercial test	26	−0.04	0.024	−0.08	+0.01
Statutory	51	+0.02	0.010	0.00	+0.04
Other/mixed	1	–	–	–	–

Meta *p*-value < 0.05**. Overall weighted mean = +0.01 SD.

FSM

No evidence of an association between type of attainment outcome (commercial or statutory) and FSM effect size was observed.

Table 139: Effect size by type of outcome/test FSM attainment outcomes

Factor	n	Weighted mean effect size	Weighted standard error	Weighted CI low	Weighted CI high
Commercial test	84	+0.04	0.014	+0.01	+0.07
Statutory	54	+0.02	0.014	−0.01	+0.04
Other/mixed	11	+0.05	0.055	−0.06	+0.16

Meta *p*-value > 0.10. Overall weighted mean = +0.03 SD.

Findings 2: Cost effectiveness

Please see above for detail on the cost effectiveness outcome variables. To summarise here, cost effectiveness was measured for the 40 evaluations that reported an effect size of above +0.05 SD for a majority of their ITT primary outcomes. In other words, for a trial to be included in the cost effectiveness outcome it had to report evidence that an intervention had a positive impact. The overall probability of one of the 82 trials being included in the cost effectiveness outcome was (p = 40/82) 0.49. Descriptive analyses examined the association between explanatory variables and the probability of a trial being included in the cost effectiveness outcome (p(CE)) and the actual cost effectiveness values for the subsample of 40 trials that reported a positive impact, specified in units of £ per pupil for an effect size of +0.10 SD. The probability of being included in the cost effectiveness outcome provided a trial-level perspective on positive 'impact' that we have used to supplement the more finely grained meta-analyses of ITT effect sizes.

This section succinctly summarises the key findings of these analyses; additional statistical tables can be found in the *Technical Annex*.

As with the *Findings 1* section, the analyses of primary and secondary outcomes are summarised at the start of each subsection using five tables, one for each of the overarching thematic areas. The following annotation applies to all tables.

Table 140: Cost effectiveness guide to tables

When a statistically significant association was observed…	
…$p \leq 0.01$	✓***
…$p \leq 0.05$	✓**
…$p \leq 0.10$	✓*
Interesting pattern observed but $p > 0.10$	✓#
Explanatory variable included in the analyses but no obvious association observed	✓

Cost effectiveness and the intervention

Summary

Table 141: Summary of descriptive analyses of cost effectiveness and intervention

Subtheme	**Explanatory variable**	**Probability of inclusion in cost effectiveness outcome** ***p*(CE)**	**Cost effectiveness of interventions reporting a positive effect** **(£/0.10 ES)**
Focus	School phase	✓#	✓*
	School Key Stage	✓#	✓*
	Curriculum focus of intervention	✓#	✓#
Intensity	Minutes per week	✓#	✓*
Who implements with direct target?	Direct implementer (teacher / TA / external)	✓#	✓*
Perceived quality of supporting resources	High / varied / low	✓	✓
Cost	Total cost	✓***	✓#
EEF intervention themes	Language and literacy	✓	✓
	Maths and numeracy	✓	✓
	Staff deployment and development	✓	✓
	Organising your school	✓	✓*
	Developing effective learners	✓	✓
	Feedback and monitoring pupil progress	✓	✓
	Behaviour	✓	✓
	Character and essential skills	✓	✓
	Parental engagement	✓	✓

	Science	✓	✓
	Enrichment	✓	✓
	Early years	✓	✓
	Special educational needs	✓	✓
Evidence of positive impact	Whether identified as promising on EEF website	✓***	✓***

Focus of the intervention

Interventions in secondary schools (p = 0.44, 11 evaluations) tended to be less cost effective (median = £69; mean = £109) compared with interventions in primary schools (p = 0.69, four evaluations; median = £43; mean = £130). Looking more closely at these patterns, cost effectiveness seems to decrease with Key Stage between KS1 (median = £11; mean = £25), KS2 (median = £60; mean = £188) and KS3 (median = £79; mean = £119). The association between cost effectiveness and both school phase and Key Stage was statistically significant at the 10% level (p < 0.10). This may be related to the greater efficiency in smaller organisations, such as working with fewer staff requiring less time overall.

Whilst relatively rare, primary to secondary school transition interventions were more likely to be included in the cost effectiveness outcome compared with interventions in primary school (p = 0.49, 25 evaluations) or secondary school. However, those that are included tend to be less cost effective (median = £376; mean = £385) than interventions at earlier or later Key Stages.

Interventions focusing on maths were more likely to be included in the cost effectiveness outcome (p = 0.64) than those focusing on English (p = 0.56), those having a cross-curriculum focus (p = 0.35) or the one science-focused intervention included. Among those interventions included, there was little evidence of difference in terms of cost effectiveness for English (median = £65; mean = £171), maths (median = £62; mean = £67) or cross-curriculum interventions (median = £24; mean = £196).

Intensity of intervention

Interventions with higher intensity (minutes per week) were more likely to be included in the cost effectiveness measure but on average were less cost effective. Interventions that had over two hours of delivery per week were the most likely to be included in the cost effectiveness outcome (p = 0.64) compared with less intense programmes (p = 0.47 or lower). Interventions that had over two hours of delivery per week were the least cost effective (median = £183; mean = £285) compared with less intense interventions (median = £69 or lower; mean = £162 or lower). The association between cost effectiveness and intensity was statistically significant at the 10% level (p < 0.10). In so far as greater intensity of delivery is likely to be more expensive than lower intensity delivery, this may in part be a trivial finding.

Direct implementers

TA-led interventions were much more likely to be included in the cost effectiveness outcome (p = 0.75) than teacher-led interventions (p = 0.51) or interventions led by someone external to a school (p = 0.33). However, TA-led interventions included were **not** the most cost effective (median = £62; mean = £139). Teacher-led interventions are seen to be the most cost effective (median = £33; mean = £50) and externally-led interventions the least cost effective (median = £257; mean = £364). The association between cost effectiveness and direct implementers was statistically significant at the 10% level (p < 0.10). TA-led interventions tend to involve working with small groups of pupils in an intense way, compared with the greater variability of teacher-led interventions which may partly explain these findings.

No significant association between cost effectiveness and perceived quality of supporting resources was observed.

Total cost

Twenty-two of the 40 evaluations (55%) included in the cost effectiveness outcome were for interventions that had a total cost of between £250k to less than £500k. Interventions in this cost band were the most likely to be included in the cost effectiveness outcome (p = 0.79) compared with cost bands above or below (p = 0.50 or lower). Ten of the 36 interventions that cost £500k or more (p = 0.28) and eight of the 26 interventions that cost less than £250k (p = 0.31) are included. Interventions that cost between £250k and less than £500k were also associated with greater cost effectiveness (median = £34; mean = £96) than interventions with higher or lower costs (median > £60; mean> £160).

The cost effectiveness outcome was constructed using the cost per pupil as a numerator (see above) and therefore cost per pupil is not included as an explanatory variable in the analyses of cost effectiveness.

EEF intervention themes

Interventions included in the theme *Feedback and monitoring pupil progress* were the most likely to be included in the cost effectiveness outcome (p = 0.70, seven evaluations) followed by *Maths and numeracy* (p = 0.69, 11 evaluations), *Developing effective learners* (p = 0.59, 10 evaluations), *Language and literacy* (p = 0.58, 22 evaluations) and *Organising your school* or *Staff deployment and development* (p = 0.50 for both).

Across the EEF intervention themes, median cost effectiveness ranged between £13 (Developing effective learners) and £163 (Organising your school). The lower cost effectiveness for the nine interventions included in the EEF theme Organising your school was the only one found to be statistically significant across the six themes with sufficient data. The low cost effectiveness is likely to relate to the higher cost of implementing whole-school interventions. Specifically, the 31 interventions **not** included in the EEF theme *Organising your school* had a notably lower average cost per 0.10 SD effect size (mean = £96; median = £43) compared with the nine programmes not in that theme (mean = £335; median = £163).[34]

EEF promising interventions

It is perhaps unsurprising that 16 of the 17 interventions identified by EEF as 'promising' (p = 0.94) were included in the cost effectiveness outcome. Inclusion in the outcome is an indication that a positive impact was found in an evaluation which would be a key consideration in whether a programme is classed as promising. The 16 promising interventions were more cost effective (median = £17; mean = £61) than the 24 interventions **not** classed as promising (median = £89; mean = £209).

Cost effectiveness and theory & evidence

Summary

Table 142: Summary of descriptive analyses of cost effectiveness and theory & evidence

Subtheme	Explanatory variable	Probability of inclusion in cost effectiveness outcome ***p*(CE)**	Cost effectiveness of interventions reporting a positive effect **(£/0.10 ES)**
Empirical evidence	Strength of prior evidence of impact	✓#	✓**
Theory	Level of theoretical detail	✓#	✓
Causal process	Focus of change (learning, teacher or wider outcomes)	✓	✓

Empirical evidence and theoretical detail

The 17 evaluations of interventions that drew on strong empirical evidence were slightly more likely to be included in the cost effectiveness outcome (p = 0.53, nine evaluations) than the evaluations with more limited or no evidence (p = 0.48 or lower). Also, on average, those interventions that were included were more cost effective (median = £43; mean = £74) than the 27 evaluations with limited evidence (median = £48; mean = £126) or the four evaluations with minimal or no evidence (median = £482; mean = £483). The association between cost effectiveness and empirical evidence was statistically significant at the 5% level (p < 0.05).

Programme evaluations with a high level of theoretical detail were less likely to be included in the cost effectiveness outcome (p = 0.35) than evaluations with more limited theoretical detail (p = 0.51 or higher). The relationship between theoretical detail and cost effectiveness is unclear (and quite possibly non-existent) but, on average, having minimal or

[34] For the mean difference, the ANOVA F-test was statistically significant at the 1% level (p < 0.01). For the medians, the non-parametric Mann-Whitney U test was statistically significant at the 10% level (p < 0.10).

no theoretical detail is associated with a higher cost per 0.10 SD effect size (median = £89, mean = £195) compared with evaluations with greater theoretical detail (median = £54 or lower; mean = £125 or lower).

Causal processes

Thirty-nine of the 40 evaluations included in the cost effectiveness outcome concerned interventions with a learning focus (98%), making scrutiny of variation in average cost effectiveness across this variable statistically unfeasible.[35]

Cost effectiveness and context

Summary

Table 143: Summary of descriptive analyses of cost effectiveness and context

Subtheme	Explanatory variable	Probability of inclusion in cost effectiveness outcome *p*(CE)	Cost effectiveness of interventions reporting a positive effect (£/0.10 ES)
External environment	Publication year	✓#	✓#
	Geography	✓	✓#

External environment

Evaluations of interventions published in 2014 and 2015 were more likely to be included in the cost effectiveness outcome ($p > 0.62$) compared with evaluations published between 2016 and 2018 ($p < 0.42$). This observed drop in the probability of EEF evaluations reporting a positive impact is a trial-level echo of the observed decline in the mean size of reported effect size. For the 40 trials included, cost effectiveness is seen to improve over the six years between 2014 (median = £100; mean = £238) and 2018 (median = £33; mean = £37), although this trend is not statistically significant.

The relationship between cost effectiveness and geographical context is not clear or statistically significant. The one thing to note is the relatively low cost effectiveness of interventions that took place in one geographical area (11 trials, median cost effectiveness = £183; mean = £254) compared with interventions with a wider geographical context (median < £59; mean < £153).

No other context variables were included in this analysis because there was not considered to be a theoretical reason for this outcome to be associated with the other context variables.

[35] This is because of the lack of variation across the explanatory variable (nearly all evaluations are found in a single category).

Cost effectiveness and implementation & fidelity

Summary

Table 144: Summary of descriptive analyses of cost effectiveness and implementation & fidelity

Subtheme	Explanatory variable	Probability of inclusion in cost effectiveness outcome *p*(CE)	Cost effectiveness of interventions reporting a positive effect (£/0.10 ES)
Developer characteristics	Charity / university / private company / school or academy or MAT / council or LA / mixed	✓**	✓
Professional development	Whether implementation uses CPD	✓	✓
	Types of CPD	✓	✓
Fidelity	Intended fidelity	✓	✓
	CPD fidelity	✓#	✓
	Implementation fidelity	✓	✓

Developers

Nine of the developers of the 82 interventions in the review were schools and/or academy trusts, and eight of these were included in the cost effectiveness outcome (p = 0.89). Four out of five interventions with developers across more than one category (mixed) were also included (p = 0.80). Other types of developers were less likely to be included: private companies (p = 0.67, six included); charities (p = 0.41; 13 included); local authorities (p = 0.38, three included) and universities (p = 0.32, six included). The association between type of developer and being included in the cost effectiveness outcome was statistically significant at the 5% level ($p < 0.05$).

Of those included, the four interventions with a mixture of developers were the most cost effective (median = £25; mean = £26) followed by universities (median = £28; mean = £74) and other types of developer (median = £41 or higher; mean = £107 or higher). There is no obvious explanation for this finding.

Professional development

The vast majority of interventions in the review involved CPD (93%, 77 evaluations) and these were more likely to be included in the cost effectiveness outcome (p = 0.49; 38 evaluations) than the five non-CPD interventions (p = 0.40, two evaluations). Whilst 38 of the 40 interventions included in the cost effectiveness outcome involved CPD, 36 of these involved face-to-face CPD.

Fidelity

No evidence was found for an association between cost effectiveness and either the intended fidelity or the actual fidelity of implementation.

For CPD fidelity, evaluations of interventions that mentioned high CPD fidelity were more likely to be included in the cost effectiveness outcome (p = 0.67, 8 evaluations) compared with evaluations that reported moderate or varied CPD fidelity (p = 0.50, 13 evaluations) or low CPD fidelity (p = 0.33, 2 evaluations). Of those included, no association between cost effectiveness and CPD fidelity was observed.

Cost effectiveness and evaluation design

Summary

Table 145: Summary of descriptive analyses of cost effectiveness and evaluation design

Subtheme	Explanatory variable	Probability of inclusion in cost effectiveness outcome *p*(CE)	Cost effectiveness of interventions reporting a positive effect (£/0.10 ES)
Trial description	Type of trial (RCT/CRT)	✓*	✓**
	Level of randomisation	✓#	✓**
	Efficacy / effectiveness	✓	✓
Intervention length and size	Intervention length (weeks)	✓#	✓#
	Number of schools	✓#	✓#
	Number of pupils	✓**	✓#
Statistical sensitivity, attrition and trial quality	Trial quality (EEF padlocks)	✓	✓#
Primary outcome	Type (commercial / statutory)	✓#	✓*
	Outcome curriculum area	✓#	✓#

Trial design description

RCTs with pupil-level randomisation were more likely to be included in the cost effectiveness outcome (p = 0.68, 17 evaluations) compared with CRTs with school-level randomisation (p = 0.44, 21 evaluations). However, of those included, interventions evaluated using an RCT were less cost effective (median = £107; mean = £221) than those evaluated by CRTs (median = £34; mean = £71).

No evidence of an association between type of trial (efficacy vs effectiveness) and cost effectiveness was observed.

Intervention length and size

Shorter interventions were more likely to be included in the cost effectiveness outcome but were less cost effective compared with longer programmes. Fourteen of the 23 interventions that lasted for up to15 weeks were included in the cost effectiveness outcome (p = 0.61) but only four of the 17 trials that ran for over a year were included (p = 0.24). Cost effectiveness for trials up to 15 weeks was lower (median = £84; mean = £223) than for trials that ran for over a year (median = £33; mean = £39).

Smaller trials (in terms of numbers of schools or numbers of pupils) were more likely to be included in the cost effectiveness outcome, but those that were included were less cost-effective programmes than the interventions evaluated with larger trials, possibly related to economies of scale. The association between the number of pupils and inclusion in the cost effectiveness outcome was statistically significant, with probability of inclusion ranging between p = 0.74 (14 trials with 500 pupils or fewer) to p = 0.42 (10 trials with more than 2,500 pupils).

Statistical sensitivity, attrition and trial quality

Whilst no association between EEF padlocks and inclusion in the cost effectiveness outcome was observed, across the included trials the EEF padlock ratings were associated with increased cost effectiveness (i.e., positively correlated). Cost effectiveness improves with each padlock rating between two (median = £88; mean = £218; 6 evaluations) and four (median = £36; mean = £95; 12 evaluations). Evaluations with five padlocks or fewer than two padlocks that were included in the cost effectiveness outcome were too sparse (n < 4) to include in this analysis. This finding is difficult to interpret.

Primary outcome(s)

Trials that use commercial tests as primary outcome(s) were more likely to be included in the cost effectiveness outcome (p = 0.53, 27 evaluations) than trials that used statutory tes data (p = 0.32, 7 evaluations). However, among the trials

included, commercial tests were associated with interventions that had a less cost effective impact (median = £89; mean = £204) compared with trials that used statutory test data (median = £15; mean = £42).

Interventions with a maths/numeracy focus that used a maths attainment primary outcome[36] were the most likely to be included in the cost effectiveness outcome (p = 0.67, 10 evaluations), followed by interventions with an English/literacy focus that used an English/literacy attainment outcome (p = 0.53, 21 evaluations) or cross-curriculum interventions (p = 0.33, 8 evaluations). Of those included, cross-curriculum interventions are observed to be the most cost effective (median = £34; mean = £175) followed by maths/numeracy (median = £60; mean = £61) and English/literacy (median = £69; mean = £197).

[36] This relates to interventions with a maths focus that are evaluated using a specific maths attainment primary outcome(s) as opposed to interventions with a cross-curriculum focus that are evaluated by a maths attainment outcome alongside (or combined with) other measures of attainment.

Findings 3: Pupil-level attrition

Please see the *Review framework* section for detail on the pupil-level attrition outcome variables and approaches to analyses. To summarise here, the pupil-level attrition rate was obtained for 79 of the 82 trials in the review.

This section succinctly summarises the key findings of these analyses; additional statistical tables can be found in the *Technical Annex*.

As in the *Findings 1* and *Findings 2* sections, the analyses of attrition are summarised using tables at the start of each of the subsections on each of the five overarching thematic areas. Values in these tables are annotated as follows:

Table 146: Pupil-level attrition guide to tables

When a statistically significant association was observed…	
…$p \leq 0.01$	✓***
…$p \leq 0.05$	✓**
…$p \leq 0.10$	✓*
Interesting pattern observed but $p > 0.10$	✓#
Explanatory variable included in the analyses but no obvious association observed	✓

Pupil-level attrition and the intervention

Summary

Table 147: Summary of descriptive analyses of pupil-level attrition and intervention

Subtheme	Explanatory variable	% Pupil attrition
Focus	School phase	✓#
	School Key Stage	✓#
	Curriculum focus of intervention	✓#
Intensity	Minutes per week	✓
EEF intervention themes	Language and literacy	✓*
	Maths and numeracy	✓
	Staff deployment and development	✓
	Organising your school	✓
	Developing effective learners	✓
	Feedback and monitoring pupil progress	✓
	Behaviour	✓
	Character and essential skills	✓
	Parental engagement	✓
	Science	✓
	Enrichment	✓
	Early years	✓
	Special educational needs	✓
EEF promising intervention	Whether identified as promising on EEF website	✓*

Focus of the intervention

Attrition rates were higher for interventions located within the Y6–Y7 primary to secondary transition (median = 17%; mean = 33%) compared with interventions solely in secondary schools (median = 16%; mean = 19%) or solely in primary schools (median = 15%; mean = 17%). Across pupil Key Stages, the highest attrition rate is seen in KS3 (median = 21%; mean = 19%) and the lowest in KS4 (median = 6%; mean = 5%).

However, when the type of primary outcome (commercial test or statutory) is accounted for, a different picture emerges. First, five of the six primary to secondary school transition interventions used a commercial test for the primary outcome. Second, for evaluations using commercial tests, attrition rates for primary to secondary school transition interventions were actually lower (median = 11%) than for interventions in secondary (19%) or primary (18%) schools. The overall attrition rates for interventions in secondary or primary schools were smaller because of their use of an statutory primary outcome in seven secondary school interventions (median attrition = 9%) and 14 primary school interventions (median = 8%). In terms of school Key Stage, median attrition rates ranged between 6% for the four KS4 interventions (all of which used a statutory test outcome) to 27% for the 15 KS2 interventions that used a commercial test outcome. The vast majority of KS3 interventions used a commercial test (16 out of 20) as did the majority of KS2 interventions (15 out of 33). When comparisons were possible, the use of commercial tests is observed to result in higher attrition rates compared with the use of statutory test outcomes.

Whilst the overall median attrition rates for interventions that focused on maths (7%) were notably lower than those for English (16%) or cross-curriculum (16%) interventions, this pattern seems to relate primarily to the type of primary outcome used (i.e., commercial or statutory). The use of a statutory primary outcome was more common in maths (five out of 14, 36%) and cross-curriculum interventions (13 out of 27, 48%) compared with English (two out of 35, 6%) interventions. Among the evaluations that used a commercial test, attrition rates for maths (16%) were closely comparable to English (16%) but a higher rate was observed for cross-curriculum interventions (22%). Among the evaluations that used statutory test outcomes, attrition rates for maths (7%) were slightly lower than cross-curriculum interventions (9%).

Intensity of intervention

No evidence of an association between the intensity of an intervention and attrition was observed.

EEF intervention themes

Median attrition rates ranged between 8% (Feedback and monitoring pupil progress; nine interventions) and 28% (Parental engagement; six evaluations). The 16 interventions in the Maths and numeracy theme are again seen to have relatively low rates of attrition (median = 11%) compared with the 63 interventions not in the Maths and numeracy theme (median = 15%) or the 37 interventions in the Language and literacy theme (median = 18%). The differences in median and mean attrition rates for these 37 Language and literacy interventions and the 42 interventions **not** in this theme were both statistically significant at the 10% level ($p < 0.10$).

Among the trials that used a commercial test as their primary outcome, median attrition rates ranged between 14% (Parental engagement) and 26% (Organising your school). Among the trials that used a statutory primary outcome, median attrition rates ranged between 7% (Maths and numeracy; Organising your school) and 17% (Staff deployment and development).

EEF promising interventions

As would be expected, average pupil-level attrition for interventions identified as 'promising' by EEF was lower (median = 11%; mean = 13%) than for other interventions (median = 16%; mean = 21%).

Pupil-level attrition and theory & evidence

Summary

Table 148: Summary of descriptive analyses of pupil-level attrition and theory & evidence

Explanatory variable	% Pupil attrition
Focus of change (learning, teacher or wider outcomes)	✓

Causal processes

No evidence of an association between causal processes and attrition was observed.

Pupil-level attrition and context

Summary

Table 149: Summary of descriptive analyses of pupil-level attrition and context

Subtheme	Explanatory variable	% Pupil attrition
External environment	Publication year	✓#
	Geography	✓

External environment

Between 2014 and 2018, the average pupil-level attrition rate was seen to reduce from a median of 21% (mean = 27%) to a median of 10% (mean = 13%). However, earlier evaluations were much more likely to use a commercial test as a primary outcome and the use of statutory test outcome was observed to increase over time. On average, evaluations that used statutory data as primary outcome(s) had lower attrition rates which fluctuated from a median of 4% in 2016 up to 9% in 2017 and down to 8% in 2018. Median attrition rates for evaluations that used a commercial test fell from 21% in 2014 to 15% in 2018. This suggests that the observed drop in attrition rates is at least partly accounted for by a reduced use of commercial tests along with declining attrition rates for the commercial tests that were used.

The pupil-level attrition outcome was not analysed by characteristics associated with the intervention because the attrition measure here includes all intervention and control pupils involved in the trial; therefore, higher levels of attrition may be explained by control school drop-out. It is considered that an intervention-school-only measure of attrition may be more useful for considering attrition against variables describing the intervention.

Pupil-level attrition and implementation & fidelity

Summary

Table 150: Summary of descriptive analyses of pupil-level attrition and **implementation & fidelity**

Developer characteristics	Charity / university / private company / school or academy or MAT / council or LA / mixed	✓
Implementation planning and support	Clarity of implementation plan	✓
	Lead-in time for implementation	✓
Professional development	Whether implementation uses CPD	✓
	Sequencing of CPD	✓
Support and monitoring	Whether developer provided support other than CPD	✓
	Monitoring of implementation	✓
	SLT support	
Fidelity	Intended fidelity	✓
	CPD fidelity	✓
	Implementation fidelity	✓

No evidence was found for an association between pupil-level attrition and the explanatory variables under the implementation & fidelity theme. This may relate to the issue of alignment, as noted earlier: the attrition outcome has a broader focus (intervention and control samples) whilst these explanatory variables were specific to the intervention sample.

Pupil-level attrition and evaluation design

Summary

Table 151: Summary of descriptive analyses of pupil-level attrition and evaluation design

Trial description	Type of trial (RCT/CRT)	✓
	Level of randomisation	✓
	Efficacy/effectiveness	✓
	Type of evaluator	✓
Intervention length and size	Intervention length (weeks)	✓
	Number of schools	✓**
	Number of pupils	✓
Statistical sensitivity, attrition and trial quality	Trial quality (EEF padlocks)	✓***
Evaluation burden	Testing burden	✓**
	IPE data collection burden	✓
Primary outcome	Type (commercial / statutory)	✓***

Trial description

No significant association was observed between pupil-level attrition and any of the variables in this subtheme.

Intervention length and size

The number of schools in a trial has a complex but statistically significant association with attrition. In general, there is a weak negative correlation between attrition and the number of schools in a trial. However, looking at the categorised variable, attrition rates are seen to consistently fall into the '21–40' band (median = 26%; mean = 29%) and the '81–100' band (median = 9%; mean = 11%), but this trend is contradicted by the two extreme bands, showing relatively low attrition for trials with up to 20 schools (median = 11%; mean = 15%) and slightly higher attrition for trials with more than 101 schools (median = 12%; mean = 16%).

Statistical sensitivity, attrition and trial quality

A relatively strong association was observed between EEF padlocks and attrition. This is perhaps unsurprising given that pupil-level attrition is one of the key considerations for the awarding of EEF padlocks. Median attrition rates are seen to fall from 43% (seven evaluations with one EEF padlock) to 9% (nine evaluations with five EEF padlocks).

Evaluation burden

The association between testing burden and pupil attrition was statistically significant at the 1% level ($p < 0.01$) with much lower attrition found for evaluations with no external tests (median = 4%; mean = 8%) than for evaluations with one or more external test (median = 16% or higher; mean = 19% or higher).

Primary outcome(s)

Unsurprisingly, trials that use commercial tests as primary outcome(s) had higher attrition rates (median = 18%; mean = 22%) than trials that used statutory test data (median = 8%; mean = 11%).

Discussion and conclusion

Discussion

This review has broken new ground. As far as we are aware, there has been no previous attempt in education research to systematically build a dataset integrating the range of variables included here: from the intervention and underlying causal theory to its implementation in context, in order to undertake a systematic review of trials. The innovations extended from the development of the set of variables, through to the protocol for coding the trials and the approach to analysis. As may be expected with work that is new in the field, there has been significant learning and we have uncovered a number of issues to consider and address as the work begun by this study is taken forward by other researchers.

Throughout this report we have pointed to the high degree of variability in detail of trial reporting in relation to many variables, in particular:

- perceptions on resources used
- explanation of causal theory (although this is increasing over time)
- virtually all contextual variables
- implementation monitoring
- making an informed assessment of fidelity.

Nevertheless, the review has uncovered some new and valuable findings. These include findings drawn from univariate analyses of the outcome and explanatory variables (such as the data on the perceived quality of resources and all aspects of the theory & evidence, context and implementation themes) and how these changed over time, in addition to the meta-analyses of primary ITT attainment, secondary ITT attainment and FSM attainment effect sizes. Please note that the review was exploratory and descriptive and the selection of explanatory variables under the five overarching themes was purposefully broad. Consequently, all meta-analyses are bivariate and descriptive in nature, and it is therefore inappropriate to draw causal conclusions from the findings.

Turning first to the design of the intervention, evaluations that drew on strong empirical evidence were associated with a higher primary ITT and FSM effect size, aligned with what might be expected from evidence in the wider field on theory-based evaluation. A different pattern was seen with secondary ITT effect sizes, where higher effect sizes were found when empirical evidence was limited or not present.

Also relating to design, whilst just over half of the reports which mentioned intended fidelity indicated that the intervention (by the direct implementer) was intended to be adopted faithfully, the remainder reported that the intention was adaptation to context. In this study, no evidence of an association between effect size (all three) and intended fidelity was found. This finding may indicate that interventions that were intended to be adapted to context, and therefore less tightly codified, may be equally likely to lead to positive effect sizes as more strictly codified interventions designed to be implemented with fidelity. This appears to contradict the finding that TA interventions, which were highly codified, tend to have higher effect sizes. As discussed in the *Findings 1* section, an explanation may lie in the 1:1 and small group nature of TA interventions, which also appears to impact positively on effect. In addition, TAs may be more reliant on highly structured programmes, whereas teachers can more effectively adapt interventions to their context. Further research would be needed to investigate the potentially contradictory findings. No evidence was found for an association between primary ITT and FSM effect size and fidelity to the intervention during implementation, as judged against the intended approach to fidelity (faithful adoption or adaptation to context) but higher secondary ITT effect sizes were observed when implementation fidelity was reported as moderate or limited. However, primary ITT, secondary ITT and FSM effect sizes *were* found to be associated with fidelity to CPD.

Evaluations that mentioned the alignment of the intervention and existing practice as an enabler were – perhaps counter-intuitively – associated with lower average primary ITT effect sizes, indicating that while the implementation process is likely to be easier when the new intervention is more closely aligned with existing practice, primary outcome effect sizes were more likely to be higher when the intervention is less closely aligned to existing practice. One interpretation is that

interventions that were more ‘closely aligned’ to existing practice may not result in a substantially different experience (and hence outcome) as a ‘business as usual’ control group, as has been found more generally in implementation science literature. No evidence of association between reported alignment of the intervention and existing practice was observed for secondary ITT or FSM effect sizes.

The resources supporting interventions were perceived as variable in quality in over half of the reported projects, and high quality in just under 40% of cases. Lead-in time for implementation emerged as an important factor, mentioned as an issue in over half of the evaluations, with insufficient lead-in time being the most common finding.

CPD most commonly took place before and during the intervention (57% of cases). Around 60% of this CPD was curriculum-specific; 65% of CPD was delivered direct by intervention delivery organisations; 90% of interventions included face-to-face CPD; and fewer than a quarter of programmes employed either mentoring and coaching or online training. No evidence was found for an association between primary ITT or FSM effect size and whether or not CPD was provided to support implementation of the intervention. For secondary ITT outcomes, higher effect sizes were observed for the three interventions that did not involve CPD. Larger primary ITT effect sizes were observed for trials that delivered CPD pre-intervention only, compared with those where CPD also took place during the intervention period. This finding suggests the importance of early-stage CPD in building implementers' confidence, knowledge, skills and capacities, as well as ensuring fidelity to the intervention. However, for secondary ITT outcomes, higher effect sizes were observed when CPD was delivered pre-intervention and during the intervention, whilst no association was observed between the sequencing of CPD and FSM effect sizes. Programmes with CPD that is subject/curriculum-specific were associated with higher average primary ITT effect sizes, resonating with existing research on effective professional development that indicates that subject-specific CPD is more likely than generic CPD to be effective in changing teachers' practice. For secondary ITT and FSM effect sizes, no association between whether CPD was subject/curriculum specific or more generic was observed.

No evidence was found for an association between primary ITT or FSM effect size and whether the developer provided informal support for the intervention beyond formal CPD, whilst higher secondary ITT effect sizes were observed when support beyond CPD was provided during the intervention. SLT support for the intervention was not found to have an impact on primary ITT, secondary ITT or FSM effect sizes. However, both these variables pertaining to support had large amounts of missing data.

Regarding the focus of the intervention, interventions with an English curriculum focus were, on average, associated with higher primary ITT effect sizes compared with trials with a maths or cross-curriculum focus. The weaker primary ITT effect size for cross-curriculum interventions is potentially an area for future investigation. No association was observed between FSM effect size and curriculum focus, but higher secondary ITT effect sizes were observed for interventions that had a maths focus.

Addressing the contexts within which interventions took place, the meta-analyses found an association between geographical context and primary ITT effect size that showed higher effect sizes for interventions located in one or up to three geographical areas. This pattern was not observed for FSM or secondary ITT effect sizes. Previous reviews have not considered this variable. The impact of geographical context on primary ITT effect sizes may relate to greater ease of consistent implementation in smaller geographical areas. Related to this, although there is little evidence of an association between primary ITT effect size and the number of schools in a trial, a statistically significant negative correlation was observed between primary ITT effect size and number of pupils. On average, trials that involved 500 pupils or fewer were associated with higher weighted effect size. For secondary ITT effect sizes, smaller samples were also associated with higher effect sizes but were significant in terms of number of schools but **not** number of pupils. This may reflect greater ease of consistent implementation. However, FSM effect sizes were not observed to be associated with the number of schools or pupils.

Turning to organisational barriers and enablers, the most commonly mentioned factors were staff time and availability (66% of evaluations) followed by specialist facilities and space (43%) and workforce capacity (38%). Pupil behaviour was mentioned as a barrier to implementation in 32% of evaluations, and positive staff expectations and/or motivations were perceived as enabling features in 33 evaluations, although no association between effect sizes and either of these two variables was observed. These findings were unexpected, as these organisational factors were frequently cited in other research into contextual variation.

However, in the meta-analysis, evaluations that mentioned staff teamwork as an enabler were, on average, associated with higher primary ITT and secondary ITT effect sizes, but not associated with FSM effect sizes. One interpretation might be that staff teamwork is an indicator of positive school orientation to the intervention. No association was observed between primary ITT effect sizes and perceptions on pupil behaviour, but evaluations that mentioned pupil behaviour as a barrier were associated with lower secondary ITT and FSM effect sizes. No association was observed between primary ITT or FSM effect sizes and reported perceptions on staff expectations and motivations, but evaluations that mentioned staff expectations as an enabler were associated with higher secondary ITT effect sizes.

In relation to programme delivery, the nine programmes that were from schools or academy trust developers had the highest weighted mean primary ITT and FSM effect sizes, aligning with the review findings by Anders et al. (2017); the current review also found higher primary ITT and FSM effect sizes for TA-led interventions, and this impact is likely to be associated with the mode of delivery and potentially fidelity to the intervention.

Concerning the evaluation designs used, clustered RCT designs are becoming more common (90% of reports published from 2017 onwards, 46% before that) and padlock ratings have increased over time. However, clustered RCTs were associated with smaller mean primary ITT, secondary ITT and FSM effect sizes compared with RCTs. This may be explained to some extent by the larger size of clustered RCT trials (mean schools = 77; mean pupils = 5,861) compared with RCT trials (mean schools = 27; mean pupils = 796). No evidence of an association between primary ITT, secondary ITT and FSM effect sizes and pupil-level attrition or testing burden was observed. There were some interesting findings in relation to the type of outcome (commercial, statutory or other). On average, commercial outcomes were associated with a higher primary ITT but lower secondary ITT effect size than statutory test outcomes. FSM effect sizes were not observed to be associated with the type of outcome. Looking at the 10 most common specific primary outcomes that accounted for 82 of the 133 effect sizes in the review (62%), GL NGRT was the most common outcome (23 effect sizes) and had the highest observed weighted mean effect size (+0.09 SD).

Relatively few evaluations were included in the cost effectiveness analysis. Bearing this in mind, some interesting findings emerge. Interventions in secondary schools tended to be less cost effective than interventions in primary schools, but there was little evidence of any difference for cost effectiveness in relation to curriculum focus. Teacher-led interventions were the most cost effective and externally-led interventions the least cost effective. Furthermore, interventions that used commercial tests were less cost effective than interventions drawing on statutory data.

Turning now to pupil attrition, the first thing to note is that average pupil-level attrition rates for trials have fallen over time. Across pupil key stages, the highest attrition rate is seen in KS3 and the lowest in KS4. However, this may well be related to the form of outcome measure. The vast majority of KS3 interventions used a commercial test (as did the majority of KS2 interventions); where comparisons were possible, the use of commercial tests is linked to higher attrition rates compared to studies using statutory test outcomes. This may partly account for the overall fall in attrition rates, earlier evaluations were more likely to use a commercial test as a primary outcome and the use of a statutory test outcome was observed to increase over time. Finally in relation to testing, much lower attrition was found for evaluations involving no external tests than for evaluations using one or more external tests.

Interestingly, no evidence of an association was observed between attrition and the intensity of an intervention. Further, no evidence was found for an association between attrition and the explanatory variables included under the implementation & fidelity theme. This may relate to the issue of alignment noted earlier – the attrition outcome takes in the intervention and control samples while the explanatory variables were specific to the intervention sample.

Limitations

Given the ambition and novelty of key aspects of this review there are inevitably important limitations that need to be acknowledged. These limitations also highlight areas for further action by researchers seeking to build on this review or undertake a similar study.

Breadth of review

From inception, this review was exploratory and descriptive, and the selection of explanatory variables under the five overarching themes was purposefully broad. The thematic framework provided structure and boundaries for the analyses but the interpretation of findings needs to be done with care and awareness of limitations. For example, the 133 primary outcome ITT effect sizes relate to a diverse range of outcome measures, subject areas, pupil cohorts and ages. The 82

evaluations that reported the 133 effect sizes relate to a diverse range of interventions with variable levels of successful implementation. So, at the level of effect size/outcome and at the intervention/trial-level there is clearly a great deal of diversity. For the meta-analyses we adopted random effects models to reflect this diversity. However, the bivariate nature of the meta-analyses presented and the large number of explanatory variables examined means that that it is not appropriate to draw causal conclusions from our analyses. The bivariate meta-analyses provide an initial inspection of this effect size diversity and whether/how it is associated with reported primary outcome ITT effect size(s). It is likely that associations found between explanatory variables and effect size(s) will overlap, so future reviews might want to explore this through multivariate meta-analyses. For example, new reviews might draw on these descriptive findings to develop a deductive multivariate approach which might have a particular subject focus (e.g., English) and/or pupil key stage.

Reliability and validity of the new explanatory variables

The reliability and validity of the explanatory variables coded for the first time in this review have inevitably been affected by three key issues: the initial intention to produce a somewhat exhaustive list of possible explanatory variables within each theme; review time and resource constraints; and significant variation in reporting of implementation and process findings across the 82 trials. Although a systematic process was adopted to review the validity of the descriptors of explanatory variables and to check the reliability of coding, it is clear that further work is needed to refine some of the variable and coding descriptors and develop a more robust assessment of coding reliability. In hindsight, focusing on a smaller number of variables would have allowed for more time to develop definitions and codes.

In the reviewed trials, inconsistent collection of data and reporting was particularly apparent for variables within the theory & evidence, implementation and context themes. A number of variables in all themes had to be dropped from the analysis due to the amount of data missing from the reports, and this issue has led to some of the recommendations set out below, such as the requirement in all future reports for more clarity and consistency on key issues. In addition, a number of variables were difficult to code due to inevitable levels of subjectivity on the part of coders (notwithstanding the coding checks in place) and in evaluators' judgements in reporting.

Furthermore, the information required for the coding was not always reported consistently across all evaluation reports, so the data were derived from varying definitions. In future, this issue could potentially be overcome by requiring surveys of stakeholders to be conducted and reported in all process evaluations in a broadly consistent manner. These surveys would include some pre-defined question formats that can be applied across interventions and tabulated in the report. A further limitation to the coding process was that coders were obliged to code based on what was reported on, rather than what may have actually taken place. For example, there may have been some issues with SLT support during the course of an intervention, but this may not have been investigated in the process evaluation or deemed important enough to be reported. This could be overcome in future by applying a core set of variables to be explored in all evaluations, and where appropriate reporting findings in a standard table.

A number of codes were set up to assess whether a particular variable was perceived to be a barrier or enabler by those involved, rather than simply whether or not the variable was present. This was particularly difficult to assess since there was variation in perception among affected stakeholders as to whether the issue was a barrier or an enabler. Moreover, since the coding was often based on qualitative data gained in the process evaluation it was not always clear if reported barriers or enablers were based on perceptions from a large or smaller number of stakeholders.

In summary:

- There was a lot of missing data in relation to implementation and context variables. For future evaluation reports, guidance on reporting these variables is recommended.
- There was ambiguity in some variable and code definitions. This needs further development, together with the associated guidance for coders in future reviews of projects.
- There was particular difficulty in coding against theory & evidence variables, especially differentiating between 'strong evidence' and 'some evidence' in relation to 'Prior evidence of theory', and the changing guidance on level of detail in EEF reports on 'how and why the intervention will lead to the intended outcomes/impact' influenced the coding here. These variables need to be treated with caution and would need to be reconsidered in subsequent reviews of projects.

Use of statistical significance

We accept the limitations of using statistical significance in the context of data from all primary outcome ITT, secondary ITT and FSM subsample effect sizes reported in the first 82 EEF evaluations that employed an RCT/CRT design. The effect sizes and the evaluations were not random samples. Therefore, the inferential use of statistical significance is not appropriate. We use statistical significance to help illuminate the strength of statistical association observed between the explanatory variables under the five overarching themes and the quantitative outcomes for the review. Analyses are descriptive, exploratory and mostly bivariate in nature. Interpretation of the statistical analyses drew on descriptive statistics, discussion in the research team, critical judgement and statistical significance. It is likely that the statistical associations observed across explanatory variables will overlap. Where overlap was clearly apparent, we undertook some limited follow-on elaboration analyses. Specifically, for the pupil attrition outcome, higher attrition rates were observed for evaluations that had used a commercial test as a primary outcome compared with evaluations that had used a statutory test outcome, as might be expected. This association serves to obscure the interpretation of how attrition is associated with other explanatory variables; in other words, it is confounding. Therefore, within the attrition analyses, associations across a selection of explanatory variables were examined separately for trials using a commercial test and trials using statutory data. Other overlaps in how explanatory variables are associated with effect sizes, cost effectiveness and/or attrition are likely to exist. This phenomenon underlines the descriptive and exploratory nature of the quantitative analyses presented this review. It is therefore entirely inappropriate to link causality to the analyses presented here. Whilst these descriptive analyses have been structured using a theoretical framework of five overarching themes, the scope is purposely broad. A future review might draw upon these exploratory and descriptive details reported here, together with theory, to help formulate and specify a narrower/targeted multivariate hypothesis and analytical approach.

Secondary outcomes analysis

Further limitations relate to the secondary outcomes in the review: cost effectiveness and attrition. Both of these outcome variables were identified during the review and after the theoretical framework with five overarching themes had been developed. This meant that explanatory variables were included in the analyses of secondary outcomes in a post hoc way. Future reviews might want to focus more directly on cost effectiveness and/or attrition in the development of their theoretical framework. Further, the attrition outcome did not align well with explanatory variables that captured aspects of an intervention or how it was implemented. Overall pupil-level attrition included pupils in both intervention and control conditions whilst some variables related solely to the intervention condition. Future reviews might collect and analyse pupil-level attrition in intervention schools only in order to explore associations with the intervention and how it was implemented.

Measuring time

It became apparent that time periods and dates could be more clearly and consistently reported in EEF trials. For example, we extracted detail on the 'length of intervention' directly from EEF trial webpages – and this is reported using a number of different time metrics (e.g., days, weeks, terms, years). We suggest that it would be beneficial to standardise the metric used and for trials to report specific dates in a clearer way, for example, date(s) of randomisation, date(s) when CPD began (if appropriate) and ended, and testing dates.

References

Anders, J., Godfrey, D. & Nelson, R. (2017), *EEF Projects Review*. EEF Report: Unpublished.

Borenstein, M., Hedges, L.V., Higgins, J.P.T. & Rothstein, H.R. (2009) *Introduction to Meta-Analysis*. Wiley.

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

Coldwell, M. (2019) Reconsidering context**:** Six underlying features of context to improve learning from evaluation, *Evaluation*, 25:1, 99–117.

Coldwell, M., & Maxwell, B. (2018) Using evidence-informed logic models to bridge methods in educational evaluation. *Review of Education*, *6*(3), 267–300.

Demack, S., Maxwell, B., Coldwell, M., Stevens, A., Wolstenholme, C., Reaney-Wood, S., Stiell, B. & Lortie-Forges, H. (2021) Review of EEF projects: Summary of key *findings*, Available at TO ADD Hyperlink

Higgins JPT, & Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* [updated March 2011]. The Cochrane Collaboration, 2011. Available from **www.handbook.cochrane.org**

Kok, G., Gottlieb, N. H., Peters, G. J. Y., Mullen, P. D., Parcel, G. S., Ruiter, R. A., ... & Bartholomew, L. K. (2016). A taxonomy of behaviour change methods: an Intervention Mapping approach. *Health Psychology Review*, *10*(3), 297–312.

Lacouture, A., Breton, E., Guichard, A., & Ridde, V. (2015). The concept of mechanism from a realist approach: A scoping review to facilitate its operationalization in public health program evaluation, *Implementation Science, 10*, 153.

Lortie-Forgues, H., & Inglis, M. (2019) Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher* 48(3), 158–166.

Maxwell, B., Coldwell, M., Willis, B. & Culliney, M. (2019) *Teaching Assistants Regional Scale-up Campaigns: Lessons Learned.* Project Report, Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/public/files/Campaigns/TA_scale_up_lessons_learned.pdf

Maxwell, B., Stiell, B., Stevens, A., Demack, S., Coldwell, M., Wolstenholme, C., Reaney-Wood, S. & Lortie-Forges, H. (2021a) *EEF Review: Qualitative Analysis of Factors Influencing Scale-up from Efficiency to Effectiveness Trials* Available at https://educationendowmentfoundation.org.uk/public/files/Publications/qualitative_analysis_of_factors_influencing_scale_up_from_efficiency_to_effectiveness_trials.pdf

Maxwell, B., Stevens, A., Demack. S., Wolstenholme, C., Coldwell, M., Reaney-Wood, S. and Lortie-Forges, H (2021b) *EEF Review: IPE Quality Measure Pilot*. Available at https://educationendowmentfoundation.org.uk/public/files/Publications/ipe_quality_measure_pilot.pdf

Nilsen, P. (2015). Making sense of implementation theories, models and frameworks. *Implementation Science*, *10*(1), 53.

Pawson, R. & Tilley, N. (1997) *Realistic Evaluation*. London: Sage Publications Ltd.

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review – a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy*, *10*(1_suppl), 21–34.

Peck, R., Olsen, C., & Devore, J. L. (2015). *Introduction to Statistics and Data Analysis*. Cengage Learning.

Sharples, J., Webster, R., & Blatchford, P. (2015). *Making the Best Use of Teaching Assistants: Guidance Report*. Available at: **https://educationendowmentfoundation.org.uk/public/files/Publications/Teaching_Assistants/TA_Guidance_Report_MakingBestUseOfTeachingAssistants-Printable.pdf**

Sharples, J., Albers. B., Fraser,S., & Kime, S. (2019) *Putting Evidence to Work: A School's Guide to Implementation*. EEF: London Available at: **https://educationendowmentfoundation.org.uk/public/files/**Publications**/Implementation/EEF_Implementation_Guidance_Report_2019.pdf**

Slavin, R. (2016), *What Works? Lessons from Randomised Trials.* EEF Report: Unpublished.

Tabak, R.G., Khoong, E.C., Chambers, D.A., & Brownson, R.C. (2012). Bridging research and practice: models for dissemination and implementation research. *American Journal of Preventive Medicine*, *43*(3), 337–350.

Vanderkruik R., & McPherson M.E. (2017) A contextual factors framework to inform implementation and evaluation of public health initiatives. *American Journal of Evaluation* 38(3): 348–359

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Vogt, P.W. & Johnson, R.B. (2011) *Dictionary of Statistics and Methodology: A Nontechnical Guide for The Social Sciences*. Sage

Appendix A: Full details of variable descriptions and codes employed in the final analysis

Guide to appendix tables

Presented below is a full list of variables used in the analysis under each thematic area, followed by the list of IPE quality variables. The tables of codes and definitions presented below are the final versions of variables that were operationalised for the analysis and so do not include earlier iterations. Each table presents the full description of the variable and the codes employed within that variable. For each explanatory variable a 'level of confidence' has been denoted using 'high', 'medium' or 'low' with a brief commentary where appropriate. This gives an indication of the reliability and validity of each explanatory variable used in the final analysis.

Variable sources

Where the source of the variable is recorded as 'review', this indicates that all the data in that variable came directly from the review of reports conducted by the review team at SHU. It should be reasserted here that much of the information that was coded from the review of evaluation reports was based on qualitative perceptions and/or survey data from the process evaluation presented in the reports, and the extent of the issue being coded was not always clear from the reports (e.g., codes could be based on a small number of stakeholders' views from a small number of schools), and the information on this varied greatly between reports. It was thus challenging to be consistent in the coding process.

Where the source of the variable is recorded as 'data', this indicates that the variable was obtained from an external source (EPPI, Lortie-Forgues & Inglis, 2019 and Anders et al., 2017) alongside a cross-checking process. There were cross-checks when variables were recorded on multiple data sources. When disagreement was found, the EEF website and evaluation reports were used to check/correct. Finally, the data were updated in order to ensure all 82 evaluations in the review were included. This final update involved drawing on evaluation reports and EEF evaluation websites to gather missing details of variables. This process was led by the quantitative strand lead for this review who has extensive experience in designing, undertaking, analysing and reporting on educational trials (many of which were funded by EEF).

In nearly all cases, variables that drew on data sources were 'factual' rather than 'perception' based – for example, effect sizes, descriptive detail on trial design, the primary ITT outcome, number of schools, EEF padlock rating etc. This means that for these items there was less risk of variability in the data gathering/completion compared with many of the IPE review items in the review. For this reason, we predominantly classify items that drew on a data source as having relatively high reliability. Since we draw on multiple sources and undertook numerous cross-checks, we also predominantly classify these items as having high validity. There are three 'data' items that we classify as medium rather than high in terms of reliability and validity: length of intervention; intensity of trial and cost per pupil. Length of intervention is classed as medium because this length was reported using different metrics (weeks, months, terms, years) and drawn from the EEF website. The different metrics were standardised into weeks using the approach shown below (please see comment below table A5). Therefore, we classify this as having medium reliability and validity. One of our recommendations to EEF is for greater clarity in reporting timings of a trial. Specifically, a standardised table that clearly shows date of randomisation, start/finish dates of CPD, and dates of when intervention activity begins/ends would bring greater precision. Details on intervention intensity (minutes per week) was obtained for 51 of the 82 evaluations in the review and so this variable is classed as having medium reliability and validity. The cost per pupil variable did not suffer the same problem of missing data as intervention intensity did because data were obtained for all 82 evaluations. However on collecting this data we observed notable variation in details on how estimates for cost per pupil were derived. Therefore we are less confident in the comparability across trials for this explanatory variable and therefore classify cost per pupil as medium rather than high.

Table A1: The intervention

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Descriptor	Code
Focus	School phase	Data	High	Factual data	Phase of education for trial	1 Primary (including early years)
						2 Primary to secondary transition
						3 Secondary
Focus	School Key Stage	Data	High	Factual data	Educational Key Stage for trial	1 Early Years
						2 Primary (KS1)
						3 Primary (KS2)
						4 Primary (multiple key stages)
						5 Transition KS2 to KS3
						6 Secondary KS3
						7 Secondary KS4
						8 Secondary (Multiple Key Stages)
Focus	Curriculum focus of intervention	Data	High	Factual data	Curriculum focus of intervention	1 Cross curriculum
						2 English
						3 Maths
						4 Science
Intensity	Minutes per week	Data	Medium	Factual but variable reporting and only available for 51 evaluations)	Minutes per week	30 min or less
						31–60 min
						61–120 min
						Over 120 min per week

Table A1 cont...

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Descriptor	Code
Who implements with direct target?	Direct implementers	Review	High	Confident of consistency and final allocation to this category, clearly defined variable from reports.	Direct implementers	1 External led (e.g., delivery partner, a coach, mentor, tutor outside the school)
						2 Parent-led
						3 Resource-led, implementer is using resources provided directly to pupils (such as financial resources) or to teachers or leaders if they are direct recipients
						4 School leader-led, implementer is a school leader HT/DH etc.
						5 TA-led, the key implementer is the TA
						6 Teacher-led, the key implementer is the teacher
						7 Other school staff-led, key implementer is another school staff member
						8 Other, Not one of the above, or multiple, please identify in a note
						9 Two or more of above categories
Perceived quality of supporting resources	Perceived quality of supporting resources	Review	Medium	Confident of consistency of coding. A sizeable minority of evaluations provided no detail on this, and there was variation in how this was reported on which reduces the confidence in this variable.	Perceived quality of supporting resources	1 High, most stakeholder groups perceive that the quality of the supporting resources are high
						2 Variation, there is variation in perceptions about the quality of supporting resources (between stakeholder groups or across schools or by particular resources)
						3 Low, most stakeholder groups perceive that the quality of the supporting resources is low
Cost	Total cost	Data	High	Clearly defined factual information EEF Source.	Total cost of delivery of intervention	<100k
						100k – <250k
						250k – <500k
						500k – <750k
						750k – <1Mill
						1Mill+
Cost	Cost per pupil (over three years)	Data	Medium	Factual but details on how this cost estimate was derived was variable.	Cost per pupil (over three years)	<£10
						£10 – <£25
						£25 – <£50
						£50 – <£100
						£100 – <£250
						£250 to <£1K
						£1K+

Table A1 cont...

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Descriptor	Code
EEF Intervention themes	EEF intervention themes	Data	High	Clearly defined factual information EEF Source	EEF intervention themes	English/literacy
						Maths/numeracy
						Staff deployment
						School organisation
						Effective learning
						Feedback
						Behaviour
						Character
						Parental engagement
						Science
						Enrichment
						Early years
EEF promising project	Whether identified as promising on EEF website	Data	High	Clearly defined from EEF Source	EEF promising project	Binary

Table A2: Theory & evidence

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Descriptor	Code
Evidence	What does the report tell us about the strength of prior evidence of impact of this or similar interventions?	Review	Medium	This was a matter of judgement, and the distinction between strong evidence and some evidence was marginal in some cases. The variable also depends to some extent on the approach of the evaluator, and is partly a judgement of the report rather than the strength of evidence. For this reason, this variable should be treated with caution.	Strength of prior evidence of impact	1 Strong evidence, the report provides a range of evidence of impact from several sources (e.g., longitudinal/RCT evidence from credible academic sources) or a single VERY robust study such as RCT or quasi-experimental design AND/OR the evaluator indicates that there is strong prior evidence of impact of this or similar interventions
						2 Some evidence, the report provides some evidence of impact (e.g., small-scale study that cannot evidence causal change; developers' own research; research that relates to some but not all aspects of the causal theory described) no comparison group used. Supporting the link between theory and outcomes AND/OR the evaluator indicates that there is a moderate level of prior evidence of impact of this or similar interventions
						3 Minimal or no evidence, there is no or very little credible evidence in the report supporting the link between the theory and outcomes; and/or the evidence provided does not clearly link to the causal theory (e.g., general evidence on impacts of parents supporting schooling, but no specific evidence on how parents using this particular approach support this particular outcome.) AND/OR the evaluator indicates that there is little prior evidence of impact of this or similar interventions.

Table A2: cont...

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Descriptor	Code
Theory	What level of detail is provided in the report about how and why the intervention will lead to the intended outcomes/impact?	Review	Medium	This was a matter of judgement. The variable depended strongly on the approach of the evaluator. So to a greater extent than in variables in other themes, this is a judgement on the report rather than the intervention. Changing guidance from EEF means that more recent reports may have more detail. For this reason, this variable should be treated with caution.	Level of detail of assumed implementation pathways(s)	1 Highly detailed, HOW: the report provides a clear and detailed description of the steps by which it is thought that the intervention inputs will lead to the intended pupil-level impacts (e.g., it includes a clear logic model or theory of change) and narrative description of the likely path(s) by which the intervention impacts (usually via one or more intermediate outcomes). WHY: there may be references to theories that predict that the intervention will lead to the intended outcomes/impact AND/OR there may be explanations drawn from empirical studies or professional experience for why the interventions is likely to lead to the intended outputs/impacts.
						2 Some detail, the report includes some description of the steps by which it which it is thought that the intervention inputs will lead to the intended outcomes/impact (e.g., it includes a simple logic model and/or brief description of the theory of change).
						3 Minimal or no detail, the report has no or a weak description of how the intervention leads to the intended final impact (outcome) (e.g., no clear theory of change or logic model or explanation of how the steps by which the intervention should lead to the intended outcome).

Table A2: cont...

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Descriptor	Code
Causal process	Focus of Change	Review	Medium	We are confident where focus of change is identified as learning focus. Other change focuses were more debatable, in that most had a longer term focus on improving learning, and sometimes it was a matter of emphasis whether this was highlighted in the report.		1 Learning focus, the direct outcomes of the intervention are focused on pupil learning
						2 School leader change, the direct outcomes of the intervention are focused on school leaders
						3 Teacher change focus, the direct outcomes of the intervention are focused on teacher change
						4 Wider pupil outcome focus, the direct outcomes of the intervention are focused on pupil outcomes other than learning (e.g., behaviour, attitudes, resilience)
						5 Other focus, not one of the above, please identify in a note

Table A3: Context

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Descriptor	Code	Final recoded version
External context	Geographical location 1	Review	High	Confident of consistency of coding. Factual information.	Were schools involved located in	1 National 2 One geographical location 3 Two or three geographical areas 4 Other	No changes
External context	Ofsted	Review	Medium	Confident of consistency of coding. However the inconsistency in reporting on this variable lowers the level of confidence in this. There were a high number of reports that did not mention this at all. Codes were amalgamated to 'barrier' vs 'not mentioned as a barrier'.	Inspection visits and/or are staff perceptions of Ofsted approval of ways of teaching related to intervention mentioned as a barrier or enabler?	1 Barrier 2 Both barrier and enabler 3 Enabler 4 Not mentioned 5 Unclear	1 Barrier 2 Not mentioned as a barrier
Characteristics of participating organisations	Specialist facilities and space	Review	Medium	Confident of consistency of coding. However inconsistency across reports regarding this variable and a sizeable number of reports that did not mention this.	Were space and or specialist facilities such as availability of computers mentioned as a barrier or enabler?	1 Barrier 2 Both barrier and enabler 3 Enabler 4 Not mentioned 5 Unclear	1 Barrier 2 Not mentioned as a barrier
Characteristics of participating organisations	Staff time and availability	Review	Medium	Confident of consistency of coding. However inconsistency across reports on reporting on this variable.	Was staff time and or availability at required times mentioned as a barrier or enabler?	1 Barrier 2 Both barrier and enabler 3 Enabler 4 Not mentioned 5 Unclear	1 Barrier 2 Not mentioned as a barrier
Characteristics of participating organisations	Workforce capacity	Review	Medium	Confident of consistency of coding but inconsistency across reports on reporting on this variable and a sizeable number of reports that did not mention this.	Was the stability of staff (e.g., level of turnover, maternity and sickness levels mentioned as a barrier or enabler?)	1 Barrier 2 Both barrier and enabler 3 Enabler 4 Not mentioned 5 Unclear	1 Barrier 2 Not mentioned as a barrier

Table A3: cont...

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Descriptor	Code	Final recoded version
Characteristics of participating organisations	Alignment between intervention and existing practice	Review	Low	Confident of consistency of coding. However there was inconsistency across reports regarding this variable, and the majority of reports did not mention this at all.	Was the alignment of the intervention with the school's usual practice mentioned as a barrier or enabler?	1 Barrier	2 Not mentioned as enabler
						2 Both barrier and enabler	1 Enabler
						3 Enabler	
						4 Not mentioned	2 Not mentioned as enabler
						5 Unclear	
Characteristics of participating organisations	Staff teamwork	Review	Medium	Confident of consistency of coding. However there was inconsistency across reports regarding this variable, and a large number of reports did not mention this at all.	Is staff working as a team mentioned as a barrier or enabler?	1 Barrier	2 Not mentioned as enabler
						2 Both barrier and enabler	1 Enabler
						3 Enabler	
						4 Not mentioned	2 Not mentioned as enabler
						5 Unclear	
Characteristics of participating individuals	Pupil behaviour	Review	Medium	Confident of consistency of coding. However there was inconsistency across reports regarding this variable, and a sizeable number of reports did not mention this at all.	Was the behaviour of pupils in the institutions mentioned as a barrier or enabler?	1 Barrier	1 Barrier
						2 Both barrier and enabler	
						3 Enabler	2 Not mentioned as a barrier
						4 Not mentioned	
						5 Unclear	
Characteristics of participating individuals	Staff expectations and motivations	Review	Medium	Confident of consistency of coding. However there was inconsistency across reports regarding this variable.	Expectation and motivations (volunteer or asked to participate), dedication, commitment, advocacy, level of engagement or buy-in-mentioned as barrier or enabler?	1 Barrier	No change to coding as there were sufficient numbers in each category for analysis.
						2 Both barrier and enabler	
						3 Enabler	
						4 Not mentioned	
						5 Unclear	

Table A4: Implementation & fidelity

Subtheme	Variable name	Source	Level of confidence*	Level of confidence notes	Descriptor	Code
Developer characteristics	Developer characteristics	Data	High	Factual information		1 Charity
						2 University
						3 Private company
						4 School or academy or MAT
						5 Council or LA
						6 Mixed
Implementation planning and time	Clarity of implementation plan	Review	Low	Limited data in many reports and where reported this was done in inconsistent ways. Often difficult to consistently distinguish between clearly understood and variation in understanding.	All or nearly all stakeholders understand how the intervention and/or (when appropriate) any associated training is to be implemented	1 Clearly understood
					Some stakeholders report being unclear about how the intervention is to be implemented and/or (when appropriate) how any associated training is to be delivered	2 Variation in understanding
					Many stakeholders report being unclear about how the intervention is to be implemented and/or (when appropriate) how any associated training is to be delivered	3 Unclear
Implementation planning and time	Lead-in time for preparation	Review	Low	This varied on how lead-in time was defined for the project so caution should be taken with this variable. The evaluation reports in general did not clearly define lead-in time as to whether it was from randomisation or from first training session.	Sufficiency of time to prepare for implementing the intervention	1 Sufficient time, all or nearly all stakeholders perceive that there was sufficient time to prepare for implementing the intervention
						2 Variation in perceptions, some stakeholders perceived that there was insufficient time to prepare for implementing the intervention
						3 Insufficient time, most stakeholders perceive that there was insufficient time to prepare for implementing the intervention

Table A4: cont...

Subtheme	Variable name	Source	Level of confidence*	Level of confidence notes	Descriptor	Code
Professional development	Is CPD provided to support implementation?	Review	Medium	Reasonable confidence in consistency of coding. Data available in most reports on this variable.		1 Yes, only to direct implementers, CPD is provided to direct implementers only (e.g., teachers who are going to use the intervention in their classroom)
						2 Yes, only to direct implementers and other stakeholders, CPD provided to direct implementers and one or more other stakeholder group (e.g., teachers who are going to use the intervention in their classroom and leaders in their school)
						3 Yes, only to stakeholders who are not direct implementers, CPD provided only to stakeholder groups who are not direct implementers (e.g., to leaders but not to the teachers who are going to trial the intervention in their classroom)
						4 No CPD, CPD is not provided to any stakeholder group
Professional development	Sequencing of PD	Review	Low	Limited confidence in consistency of coding due to the term 'CPD' not being consistently applied by all coders and a lack of clarity in a minority of reports.	In relation to the intervention	1 Pre-intervention only
						2 During the intervention only
						3 Pre and during the intervention
						4 Not mentioned
Professional development	Subject or curriculum specific or generic	Review	Medium	Confidence in consistency of coding but insufficient data in some reports to make valid judgements.	Focus of CPD	1 Predominantly subject or curriculum specific, CPD content specifically focuses on the subject and/or curriculum that is the focus of the intervention (e.g., maths pedagogy)
						2 Predominantly generic, CPD content is generic and participants are expected to apply it to their own subject or curriculum
						3 Mixed generic and subject specific
						4 Not mentioned

Table A4: cont...

Subtheme	Variable name	Source	Level of confidence *	Level of confidence notes	Descriptor	Code
Professional development	Who delivers CPD direct to implementers?	Review	Medium	Confidence in consistency of coding. Some reports lacked sufficient detail and/or clarity in relation to this variable.		1 Delivery partner, most CPD is delivered directly by the delivery partner
						2 Other external organisation, most CPD is delivered directly by other external organisations or consultants
						3 Leaders and/or teachers selected from schools participating in the trial. Most CPD is delivered by leaders and/or teachers selected from the participating schools/ organisations
						4 Mixed, more than one type of deliverer
						5 Not mentioned
Professional development	Types of CPD – Face to face training	Review	Medium	Confidence in consistency of coding and most reports provided sufficient data to make a judgement.	e.g., workshops, training session where the CPD leaders works directly with one of more stakeholder group	1 Yes
						2 No, or not mentioned
						3 Mentioned but unclear
Professional development	Types of CPD – Online training	Review	Medium	Confidence in consistency of coding. Many reports omit any data on this so it is not possible to be certain whether this is not because there was no online training or it had not been made explicit that training was online.	Training activities and/or resources are provided online and one or more stakeholder group is expected to engage with them – this may also include discussion forums for participants	1 Yes
						2 No, or not mentioned
						3 Mentioned but unclear
Professional development	Types of CPD – Coaching or mentoring	Review	Medium	This code was restricted to instances of specific mentions of the words 'coaching' or 'mentoring' and we have confidence in consistency of coding on this basis. However, there are issues of validity as the coding is reliant on the developer and/or the evaluator labelling a practice as coaching or mentoring rather than for example 'support', 'feedback' or 'classroom visit'.	Coaching and or mentoring provided to any stakeholder group to support implementation – this may be face to face or online	1 Yes
						2 No, or not mentioned
						3 Mentioned but unclear

Table A4: cont...

Subtheme	Variable name	Source	Level of confidence*	Level of confidence notes	Descriptor	Code
Professional development	Types of CPD – Cascade train-the-trainer training	Review	Medium	Confidence in consistency of coding. Many reports omit any data on this so it is not possible to be certain whether this is not because there was no train-the-trainer model or it has not been made explicit that a train-the-trainer model was being deployed.	Trainers are trained by the delivery partner. These trainers then train the direct implementers (e.g., The delivery partners train middle leaders in schools). The middle leaders then train teachers in their department to implement the intervention.	1 Yes 2 No, or not mentioned 3 Mentioned but unclear
Support and monitoring	Delivery partner support (excludes CPD)	Review	Medium	Confident in consistency of coding where this is made explicit in reports but significant missing data and reliance on the 'labels' developer and evaluators attach to support and related activities such as coaching and mentoring (see above) which limits validity.	Any direct support for implementation from the delivery partner EXCLUDES face to face training, online training programme and coaching and mentoring.	1 Before the intervention only 2 During the intervention only 3 Before and during the intervention 4 Neither before nor during the intervention 5 Not mentioned
Support and monitoring	Monitoring of implementation (internal or external)	Review	Low	Confident in consistency of coding but data missing across a sizeable number of reports	Extent to which monitoring was conducted by the delivery team, or organisation commissioned by the delivery team, or the school involved.	1 Robust monitoring, robust monitoring of the implementation of the intervention either by the delivery team, or organisation commissioned by the delivery team, or the school involved. 2 Some monitoring, there is some monitoring but it is not very strong 3 No monitoring 4 Not mentioned
Support and monitoring	Senior leader support	Review	Medium	Some issues in distinguishing between 'strong' and 'some' codes. Inconsistency across projects in whether or how this was reported.	Senior leader support for implementing the intervention	1 Strong, stakeholders perceive that SLs provide a high level of support to enable the intervention to be implemented 2 Some, stakeholders perceive that SLs provide some support to enable the intervention to be implemented, but this could be higher 3 Limited or minimal, stakeholders perceive that SLs provided limited or no support to enable the intervention to be implemented 4 Not mentioned

Table A4: cont...

Subtheme	Variable name	Source	Level of confidence*	Level of confidence notes	Descriptor	Code
Fidelity	Fidelity related to CPD (provided for direct implementers)	Review	Medium	Generally confident in consistency of coding but some limitation as codes combine two dimensions – fidelity to intended delivery of CPD and attendance. Inconsistency in approach to reporting fidelity across reports and in some insufficient data.	Extent to which the content of CPD was delivered as intended and the level of attendance of intended participants	1 High, the content was delivered as intended and there was a high level of attendance by target participants across most schools.
						2 Varied or moderate, some variation in the extent to which the same content was delivered and/or variation in levels of attendance by participants in different schools or moderate levels of participation across most schools.
						3 Limited, some variation in the extent to which the same content was delivered and/or variation in levels of attendance by participants in different schools or moderate levels of participation across most schools.
						4 Not mentioned
						5 No CPD
Fidelity	Intended fidelity (of implementation by direct implementers)	Review	Medium	Confidence in consistency of coding but many of reports did not explicitly identify whether faithful adoption or adaptation to context was intended. Therefore default code was faithful adaptation unless there was an express statement about adaptation, where it was less clear or contradictory, not mentioned was used.	Whether the direct implementers were expected to faithfully adopt the intervention or they were expected to adapt the intervention to their context.	1 Faithful adoption, direct implementers were expected to deliver the intervention as exactly as specified by the deliver partners
						2 Adaptation to context, direct implementers were expected to adapt the intervention to fit their own context.
						3 Not mentioned
Fidelity	Actual fidelity of direct implementation	Review	Medium	Confidence in consistency of coding but inconsistent reporting limits validity.	Are direct implementers implementing the intervention as intended (i.e., if faithful adoption are they following the protocol); if adaptation to context are they following the core features that they are expected to be maintained as they adapt to context (e.g., principals specified, any sequencing of specific content).	1 High
						2 Varied or moderate
						3 Limited
						4 Not mentioned
						5 unclear

Table A5: Evaluation design

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Code
Trial design description	Type of trial (RCT/CRT)	Data	High	Factual data	1 RCT
					2 Clustered RCT
Trial design description	Level of randomisation	Data	High	Factual data	1 School
					2 Pupil
					3 Class or Teacher
					4 Year or Key Stage
					5 Parent
					6 Other / Complex
Trial design description	Efficacy/effectiveness	EEF	Medium	Obtained from EEF, however some discrepancies on trial definitions found.	1 Efficacy
					2 Effectiveness
Trial design description	Type of evaluator	Data	High	Factual data	1 University
					2 Non-university
Size and intervention length	Intervention length (weeks)*	Data	Medium	See note below*	1 Within one term (up to 15 weeks)
					2 Within two terms (15–30 weeks)
					3 Within 3 Terms(1 year, 30–14 weeks)
					4 More than one academic year
Size and intervention length	Number of schools	Data	High	Factual data	20 or less
					21 to 40
					41 to 60
					61 to 80
					81 to 100
					101 or more
Size and intervention length	Number of pupils	Data	High	Factual data	500 or less
					501 to 1,000
					1,001 to 2,500
					2,501 to 5,000
					5,001 or more
Statistical sensitivity, attrition and trial quality	Statistical sensitivity (MDES estimate)	Data	High	Factual data	Continuous data
Statistical sensitivity, attrition and trial quality	Pupil level % attrition	Data	High	Factual data	Continuous data
Statistical sensitivity, attrition and trial quality	Trial Quality (EEF padlocks)	Data	High		0
					1
					2
					3
					4

					5

Table A5: cont...

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Descriptor	Code
Evaluation burden	Testing burden	Review	High	Factual data		Low (just NPD) Medium (one external test) High (two or more external tests)
	IPE data collection burden	Review	Medium	Factual data but not always clear information on this in evaluation reports		Low (no surveys or interviews) Medium (just surveys or just interviews) High (interviews and surveys)
Primary outcome	Type	Data	High	Factual data		Commercial Statutory Other or mixed
Primary outcome	Outcome curriculum area	Data	High	Factual data		Cross-Curriculum English / Literacy Maths / Numeracy Science
Primary outcome	Number of primary outcomes	Data	High	Factual data		One Two Three or more
Primary outcome	Alignment of Primary outcome(s) and measure(s)	Review	Medium	Confident of consistency of coding, although this was relatively clear-cut there was a judgement made here which reduces the confidence level slightly.	Direct match (e.g., the intervention targets improvement in maths in Y10 and Y11 and the primary outcome measure is maths GCSE). Associated match (e.g., the focus of the intervention is improving problem solving in geography in Y6 and the primary measure is SATs scores). Limited match (e.g., the focus of the intervention is improving engagement in school through	1 Direct match between primary outcome and intervention focus 2 Associated match between primary outcome and intervention focus 3 Limited match between primary outcome and intervention focus

					adventure learning and the primary outcome measure is attainment in English)	

***** Intervention length: The UCL/IoE data have length of intervention in weeks; for more recent trials EEF lists this on the trial website. However, the length of time is reported in a range of units (weeks, months, terms and years). To standardise the lengths of the 82 trials in the review into weeks, the following approaches were taken:

Months: To convert from months to weeks the number of months is multiplied by (52/12 =) 4.3' and then rounded to the nearest week.

- 1 Month = 4 weeks (rounded)
- 3 Months = 13 weeks (rounded)
- 6 Months = 26 weeks (rounded)

Years: These tend to relate to academic years and a full 12 months is not used in the first year:

- 1 Year: Sept–July = 10.5 months = 10.5*4.3 = 45 weeks (rounded)
- 2 Years: Sept – Sept–July = 22.5 months = 97 weeks (rounded)
- 3 Years: Sept–Sept–Sept–July = 34.5 months = 149 weeks (rounded)

Terms: There are three terms in a 10.5 month academic year

- 1 Term: 10.5/3 months = 3.5 months = 15 weeks (rounded)
- 2 Terms: 21/3 = 7 months = 30 weeks (rounded)
- 3 Terms: 10.5 months = 45 weeks (rounded)
- 4 Terms: This will span the summer hols and so an extra 1.5 months are added: 10.5 + 1.5 + 3.5 = 15.5 months = 67 weeks
- 5 Terms: 10.5+1.5+7 = 19 months = 82 weeks

This results in having a near complete list of 'length of intervention' data at the trial level – with the following data added from reports in order to complete this field.

Table A6: IPE quality

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Code
IPE quality	Sufficiency of data sources	Review	Medium	Confidence in consistency of coding with some limitations as each code includes more than one dimension. In a limited number of reports (particularly earlier reports) insufficient information to make a valid judgement.	1 High – Data is collected from all the groups that are necessary to answer the RQs (e.g., as appropriate leaders, teachers, pupils, delivery partners, others connected to the intervention). Where the focus of the intervention is pupil change this includes data collected directly from pupils or through observation of pupils engaging in the intervention. Data is also collected from the control group to the extent necessary to establish the 'business as usual' condition.
					2 Medium – Some gaps in data collection from groups that are necessary to answer the RQs (e.g., as appropriate leaders, teachers, pupils, delivery partners, others connected to the intervention). And/or insufficient data is collected from the control group to establish the 'business as usual' condition.
					3 Low – Significant gaps in data collection from groups that are necessary to answer the RQs (e.g., as appropriate leaders, teachers, pupils, delivery partners, others connected to the intervention). And/or no data is collected from the comparison group.
IPE quality	Quality of data collection methods	Review	Medium	Confidence in consistency of coding with some limitations as each category covers more than one dimension. Also within some reports some methods of data collection were more clearly specified and/or more appropriate than others. In a few reports insufficient data to make a judgement.	1 High, all methods of collecting data are clearly specified and valid (i.e., they measure what they are supposed to measure).
					2 Medium, methods of collecting data are variably specified and/or variably valid (i.e., they do not all measure what they are supposed to measure).
					3 Low, methods of collecting data are poorly specified and/or lack validity (i.e., most do not measure what they are supposed to measure).
IPE quality	Quality of sampling	Review	Medium	Confidence in consistency of coding with some limitations as within some reports some sampling methods (usually for a survey) were more clearly specified than for other data collection activities. In a few reports insufficient data to make a judgement.	1 High, sampling approach is clear, justified and appropriate in relation to all methods used. For qualitative work the sample does not need to be statistically representative but to be categorised as high it would require a sample that is random or purposive rather than a convenience sample.
					2 Medium, sampling approach is largely clear, reasonably well justified and appropriate to the methods used but does not fully meet the criteria for high.
					3 Low, sampling approach is unclear and/or poorly justified and/or not appropriate to the methods used.
IPE quality	Quality of analysis methods	Review	Low	Confidence in consistency of coding with limitation as each category covers more than one dimension. Also limited (or no) description of analysis methods in a significant number of reports and limited reporting of analysis in some (mostly earlier) reports undermines validity.	1 High, Methods of analysis are clearly set out and appropriate in relation to the type/s of data and to answer the research questions.
					2 Medium, Methods of analysis are variably set out and/or vary in appropriateness in relation to the type/s of data and to answer the research questions.
					3 Low, Methods of analysis are largely missing and/or are inappropriate in relation to the type/s of data and to answer the research questions.

Table A6: cont...

Subtheme	Variable name	Source	Level of confidence	Level of confidence notes	Code
IPE quality	IPE Conduct	Review	Medium	Dependent on the detail in the description of methods and reporting of findings.	1 High, Intended data collection and analysis methods are followed or any changes to methods are justified and appropriate.
					2 Medium, Intended data collection and analysis methods are not always followed and/or changes to methods are not always clearly justified and/or are not always appropriate.
					3 Low, there is low adherence to intended data collection methods or it is unclear whether intended data collection and analysis methods are followed and/or any changes to data collection or analysis methods are generally not justified or not appropriate.

Appendix B: List of omitted variables

Once the coding process was completed, a univariate analysis was conducted and a meeting held to discuss the reliability of all the original variables set up for the coding process. At this point a number of variables were omitted from the final analysis for reasons such as lack of information in the reports and thus low numbers for the variable, or lack of consistency across the reports. Some of the original variables were used as confirmatory variables for other external data that was brought in to the dataset. The variables omitted from the final analysis are presented below under each theme.

Table B1: The intervention

Variable	Reason omitted
Clarity of specification of the intervention	Dropped due to lack of confidence in consistency of this. There was some overlap with intended fidelity which was considered to be a better quality variable.
Is there more than one year group involved in the trial at the same time?	Used as confirmatory variable only, overlap with other data.
Do pupils in one year group continue in the intervention beyond one academic year?	Used as confirmatory variable only, overlap with other data.

Table B2: Theory & evidence

Variable	Descriptor	Reason omitted
Direct or training based	Training-based intervention or developer provides the implementation direct (e.g., to students).	The dropped variable significantly overlaps the direct implementer variable and CPD variables, in practice, so has been removed.

Table B3: Context

Variable	Descriptor	Reason omitted
Cost (financial) resources at the institution	Was this mentioned in the evaluation report?	Dropped due to a high number of 'unclear or not mentioned'.
Geographical location 2 (urban/rural/town/fringe)		Dropped due to a high proportion of not mentioned, little variation in responses (most are 'mostly urban').
% of academies in sample (or not mentioned)	Were there any academy schools in the sample? put %	Dropped due to high proportion of missing data, variation on how this is presented in reports (some give intervention only, some give intervention and control).
Ofsted ratings of schools (at analysis not randomisation) mentioned or not? if yes complete below	Were the Ofsted ratings of the schools in the sample mentioned?	Dropped, reporting of the Ofsted ratings varied in each report, some did not report, some reported intervention only and some reported intervention and control together. Additionally it was unclear at what time point the ratings had been gathered.
Outstanding (number of schools)		Dropped
Good (number of schools)		Dropped
Requires improvements (number of schools)		Dropped
Inadequate (number of schools)		Dropped
missing Ofsted details (number of schools)		Dropped
Other interventions	Were other interventions happening in the school at the same time?	Dropped; only 10% of reports had information on this
Religious character of schools	Are there institutions with religious characteristics in the sample?	Dropped; only 14% of reports had information on this.
School ethos	Alignment of intervention with institution values/ethos – was	Dropped; too much variation in the understanding of what this means and a high number of 'not mentioned'.

Variable	Descriptor	Reason omitted
	this mentioned as a barrier or enabler?	
Special school or PRUs in sample?		Dropped; high proportion of not mentioned and of those that mentioned this was mostly 'no'.
Setting or streaming	Where setting/streaming (grouping pupils by prior attainment- in class, by class or differentiation) arrangements mentioned as a barrier or enabler?	Dropped; only 8% of reports mentioned this
External organisations (NOT Ofsted, not other schools)	Other institutions, government, external bodies/organisations, LA, mentioned as a barrier or enabler? EXCLUDES OFSTED and other schools	Dropped; only 11% of reports clearly mentioned this
Parents or carers	Parents/carers mentioned as a barrier or enabler?	Dropped; large number of 'not mentioned'.
Policy	Wider economic local and national policy – including curriculum and assessment policy	Dropped; over 80% not mentioned/unclear.
Support from other schools		Dropped; over 80% not mentioned/unclear.
Staff experience	Is staff previous involvement in trial or evidence informed teaching, relationship to research evidence mentioned as a barrier or enabler?	Dropped; a high number of 'not mentioned' and lack of consistency about what this defined
Wider staff involvement	Other staff in the school who are not the direct focus on the intervention (e.g., when teachers are the main focus of the intervention this could be TAs) mentioned as a barrier or enabler?	Dropped; a high number of 'not mentioned'.
Other CONTEXT enablers or barriers not coded elsewhere		Dropped

Table B4: Implementation & fidelity

Variable	Descriptor	Reason omitted
Structured peer support	Peer to peer collaboration that is structured (e.g., lesson study or where there are clear protocol and/or proforma for working together). EXCLUDES COACHING AND MENTORING	Dropped; hard to define and not enough consistency across reporting, large number of not mentioned/unclear.
Weakly structured peer to peer support	e.g., teachers are encouraged to collaborate and share ideas but no formal structure to do this EXCLUDES COACHING AND MENTORING	Dropped; large number of not mentioned/unclear.
Other	Any other form of CPD not coded above	Dropped.
Amount of CPD for an individual Senior leader (in days)	The total number of days (and half days) of CPD that one senior leader (if applicable) would receive as part of the programme.	Dropped; information on this was not always clear in evaluation reports and there was inconsistency in reporting this.
Amount of CPD for an individual teacher (in days)	The total number of days (and half days) of CPD that one teacher (if applicable) would receive as part of the programme.	
Amount of CPD for an individual TA (in days)	The total number of days (and half days) of CPD that one TA (if applicable) would receive as part of the programme.	
Amount of CPD for an individual Other (e.g., youth worker, parent, external tutor) (in days)	The total number of days (and half days) of CPD that another stakeholder (if applicable) would receive as part of the programme.	

Table B5: Evaluation design

Variable	Reason omitted
Start date of intervention (day, month and year of randomisation if available)	Checked and combined with end date of intervention to give 'length of intervention' variable. Verified against the 'length of intervention' variable from IoE source. Used 'length of intervention' from IoE source in analysis.
End date of intervention	Checked and combined with start date of intervention to give 'length of intervention' variable. Verified against the 'length of intervention' variable from IoE source. Used 'length of intervention' from IoE source in analysis.
Dates of outcome measure1 (e.g., commercial or developer designed test)	Used to create time between intervention end and testing date. Not always clear or possible to define 'intervention end date'.
Dates of outcome measure 2 if applicable (e.g., date of test or exam, such as SATs, if available.)	Not used
Are schools required to provide details on teachers / pupils (e.g., teacher names, class lists etc.) to evaluators?	Dropped; assumed this is the case for nearly all trials.
If yes, how many times are schools required to do this over the period of the trial?	Dropped; not enough clear information on this in the reports.
What is the average pupils per school included in the evaluation (and hence testing)?	Dropped; this was verified with other external sources.
IF pupils take a test prior to randomisation, what is the average number of days used to administer/collect this test per school?	Dropped; lack of information on this in the reports.
Is a pre-test (baseline) measure included in the impact analyses? [other than a Key Stage test]	These school burden variables were simplified and used to create a single variable 'testing burden' which was based on the number of external tests included overall.
If yes, How is the pre-test (baseline) measure obtained / collected?	
Following randomisation, are any test measures [other than Key Stage tests/external exams such as GCSE/A Level] included in the impact analyses?	
if yes, how many?	
if yes, How are these test measure(s) obtained / collected?	
What is the average number of days used to administer/collect these tests? per school?	Dropped – lack of information on this in the reports.
In how many intervention schools did the evaluators collect interview or focus group data? (Face to face, by phone or electronically)	These variables were amalgamated into a single variable 'IPE burden'.

Variable	Reason omitted
In total how many school staff took part in interviews and focus groups conducted by the evaluators in intervention schools (leaders, teachers, TAs ,other)	
In total how many pupils took part in interviews and focus groups conducted by the evaluators in intervention schools (leaders, teachers, TAs ,other)	
In how many control schools did the evaluators collect interview or focus group data? (Face to face, by phone or electronically)	
In total how many school staff took part in interviews and focus groups conducted by the evaluators in control schools (leaders, teachers, TAs ,pupils other)	
In total how many pupils took part in interviews and focus groups conducted by the evaluators in control schools ((leaders, teachers, TAs ,pupils other)	
Was any other qualitative data collected from school participants?	
If yes – specify method of data collection and total number of school staff and pupils who took part for each method	
Did any Teacher surveys take place in intervention schools?	
Did any Teacher surveys take place in control schools?	
Did any pupil surveys take place in intervention schools?	
Did any pupil surveys take place in control schools?	
Did any other surveys take place in intervention schools?	
Did any other surveys take place in control schools?	

Appendix C: Psychological outcomes

The psychological outcomes might be categorised in a number of ways, which will determine how 'fine grained' analyses might be. On inspecting the 88 effect sizes for psychological outcomes reported by 21 of the 82 evaluations in the review, four over-arching categories have been identified: Attitudes and beliefs; cognition; social behaviour and mental health. These four over-arching categories could in some cases be collapsed further to provide more specific sub-categories. However, the feasibility of this will be determined by the subsample size. The rest of this appendix considers each over-arching category and the range of outcome measures that these would include.

Attitudes and beliefs

Attitudes

Attitudes are evaluations where we are stating a preference towards an attitude object. For example (in the context of the EEF reports) 'I love English', 'I dislike maths'. When an attitude is expressed the individual is describing the relationship between themselves and the attitude object (Mathematics, English, swimming etc.) Attitudes are also important to an individual's self-concept.

Attitudes are comprised of three main elements, cognitive (thoughts and beliefs about something), behavioural (how you act towards the attitude object) and affective elements (emotional reactions towards the attitude object) to attitudes.

This category could be split further into 'self-attitudes', for example: self-esteem, self-confidence AND attitudes about others/objects (i.e., family, parent, subjects).

Attitudes about self

Self-esteem

Self-esteem is an attitude about you and is comprised of cognitive and affective elements. Self-esteem can be positive or negative attitudes about yourself in its entirety (global) or specific. We should not conflate global self-esteem with specific self-esteem. For example, poor academic self-esteem does not necessarily equate to low general/global self-esteem. Research has suggested that specific self-esteem is most relevant to behaviour and global self-esteem most relevant to psychological wellbeing (Rosenberg et al., 1995).

Self-confidence

Trust in yourself and ability to deal with challenges.

Self-concept

How someone thinks about or evaluates themselves.

Self-efficacy (situation-specific construct)

Belief in capacity to execute behaviours necessary to attain a specific outcome.

Attitudes towards others

These are attitudes about things other than self. For example, attitudes towards subjects at schools.

Example: Attitudes towards Mathematics– Catch up Numeracy Effectiveness Trial

Beliefs

Beliefs are different to attitudes – idea accepted to be true but without any facts.

Example: Self-confidence – Youth Social Action Trials

Cognition and metacognition

Cognition

Cognition is about mental processes and abilities. For example, memory, learning, problem solving–attention and decision making are all cognitive processes.

Example: GL-CAT4-Philosophy for children project

Metacognition

Metacognition is multifaceted and most commonly thought of as 'thinking about thinking', but more specifically it is about the process of planning, monitoring and assessing ones understanding and performance–an evaluation of cognitive processes.

Metacognition is comprised of; metacognitive knowledge (knowledge – i.e., how learning works), metacognitive monitoring (assessing (i.e., how well you are understanding what you learn), metacognitive control (regulating–i.e., deciding to try a new method to a difficult problem).

Example: Pupil views template – Mind the Gap project

Social behaviours

This was a difficult category and I have titled it social behaviours, but I'm not convinced that is the most appropriate title to give this category, so I will keep thinking this one through. This seems to logically split into category can collapse into two sub-categories: social skills and motivation.

Social skills

Social skills are learnt and socially acceptable behaviours that allow people to positively interact and communicate with others.

Motivation

Initiates, guides and maintains goal orientated behaviour and can be thought of as either intrinsic or extrinsic. Intrinsic motivation comes from within, it is personally rewarding to you and extrinsic is motivation from an external source, the reward comes from an external source avoid punishment.

Example: Teamwork– Children's University Project

Mental health

Some of the measures used are more clinical in nature and therefore I think it would be beneficial to have a separate category that encompasses these. I don't think this can really be split any further, but the depression and anxiety measures fit here, equally the SDQ is seen as a clinical measure.

Example: Penn States Worry Questionnaire – FRIENDS for Life

Appendix D: Statistical detail on meta-analyses

The meta-analyses were undertaken using the R statistical program, specifically the Metafor R package. This appendix draws on Borenstein et al. (2009)[37] to illustrate how the weighted mean estimates were obtained. The appendix also briefly notes the statistical power of the meta-analyses.

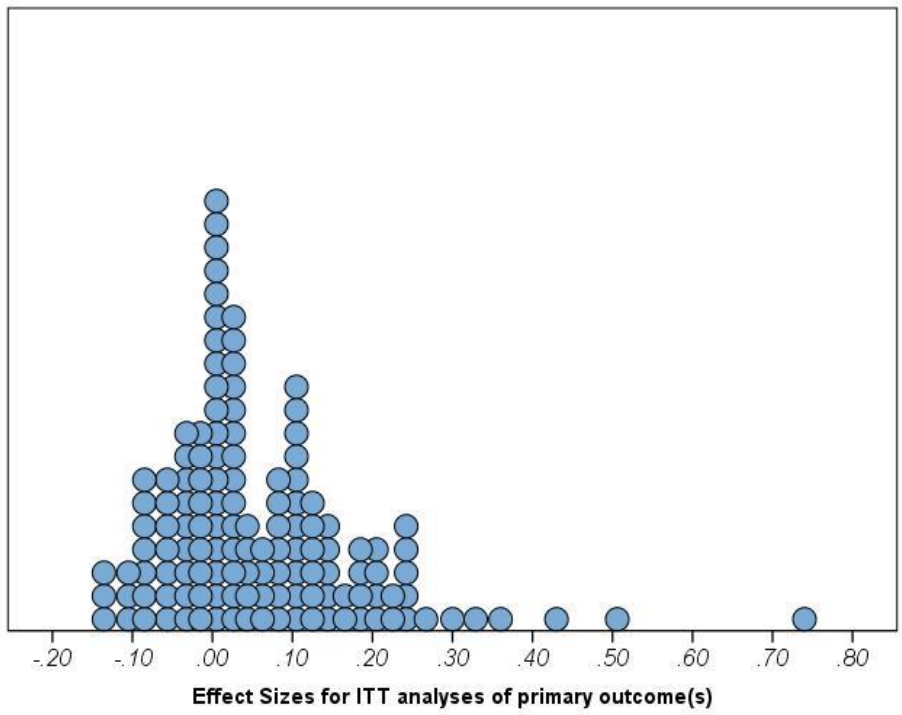
Table D1 below is taken from the report, and shows a descriptive statistical summary of the 133 primary ITT effect sizes reported by the 82 EEF trials that had reported up to January 2019. Below this, Figure D1 shows the distribution of these 133 primary ITT effect sizes.

Table D1: 133 reported effect sizes for ITT analyses of 82 EEF trials in the review: descriptive/unweighted analyses of effect sizes

	Number of trials*	No outcomes /effect sizes	Unweighted Median (IQR)	Unweighted Mean (SD)	Min	Max
SHU review (trial level)	82	–	+0.03	+0.07 (0.135)	–0.13	+0.74
SHU review (ES level)	–	133	+0.03	+0.06 (0.128)	–0.14	+0.74

*excludes quasi-experimental designs

Figure D1: Dot plot: distribution of 133 effect sizes (effect size level)



This approach assumes that all effect sizes have the same statistical reliability. For example, an effect size reported by an evaluation involving relatively few pupils across few schools would be given the same weight as an effect size reported by a large-scale national trial involving over 100 schools and 2,000 pupils. A meta-analysis approach re-weights the effect sizes to account for this variation in statistical reliability – the focus is on the standard error attached to each primary ITT effect size.

[37] Borenstein et al. (2009).

The Standard Error (SE) is the standard deviation of the sampling distribution of an effect size. It is a measure of sampling error; it refers to error in the effect size estimates due to random fluctuations across the samples. SE goes down as the number of pupils and schools go up. Adapted from Vogt & Johnson, 2011[38]*.*

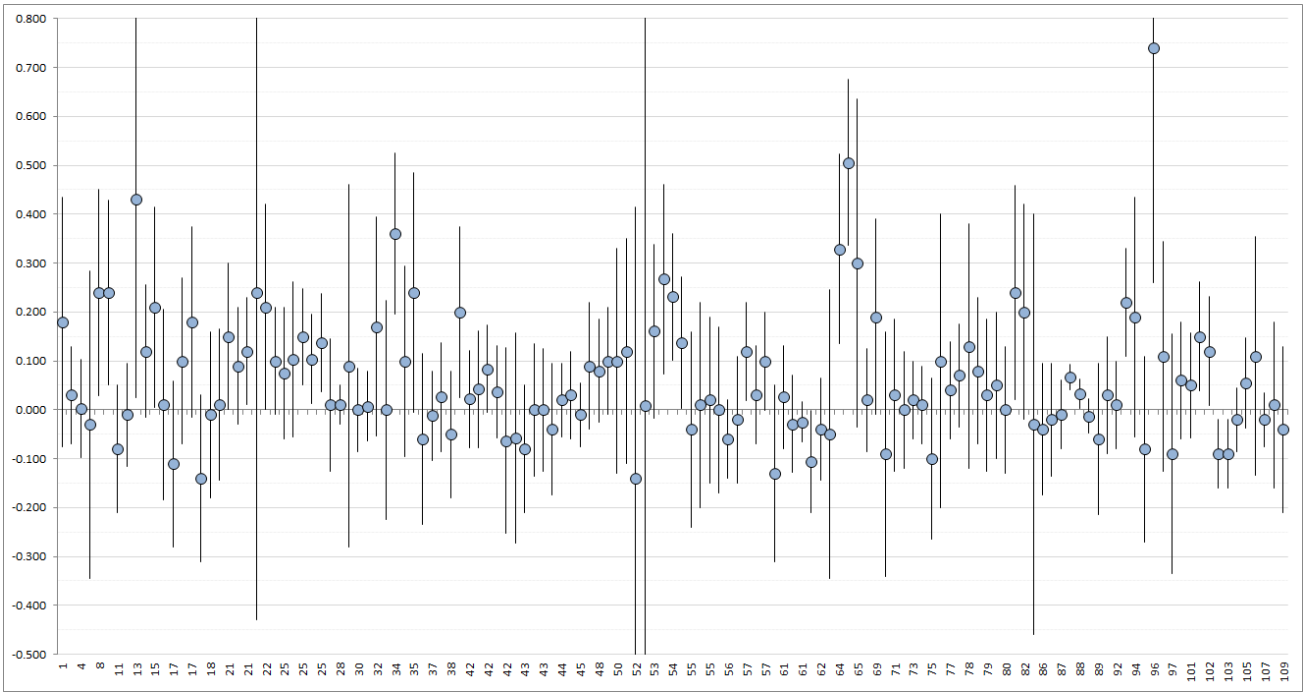
One way to visualise the variation in statistical error across effect sizes is a forest plot that shows the reported mean effect size along with 95% CI. Figure D2 and D3 below do this for the 133 primary ITT effect sizes. Figure D2 shows this ordered by the SHU_ID code (i.e., effect sizes within evaluations/trials) whilst Figure D3 shows this rank ordered high to low by the reported mean effect size.

Fixed and random effects meta-analysis

Borenstein et al. (2009) highlight two broad approaches for undertaking meta-analyses: a fixed-effects and a random effects approach. A fixed-effects approach assumes that there is a single 'real' effect size to be estimated. This approach is suitable for meta-analyses of studies focusing on similar interventions that use similar outcomes (e.g., Key Stage 2 Maths CPD programmes that use KS2 maths as an outcome). A random-effects approach assumes that there are multiple 'real' effect sizes behind the 133 evaluation estimates. Given that the meta-analyses undertaken for the review were descriptive and purposely broad, it is clear that a random-effects approach is the more suitable; and was the approach adopted within the R Metafor package.

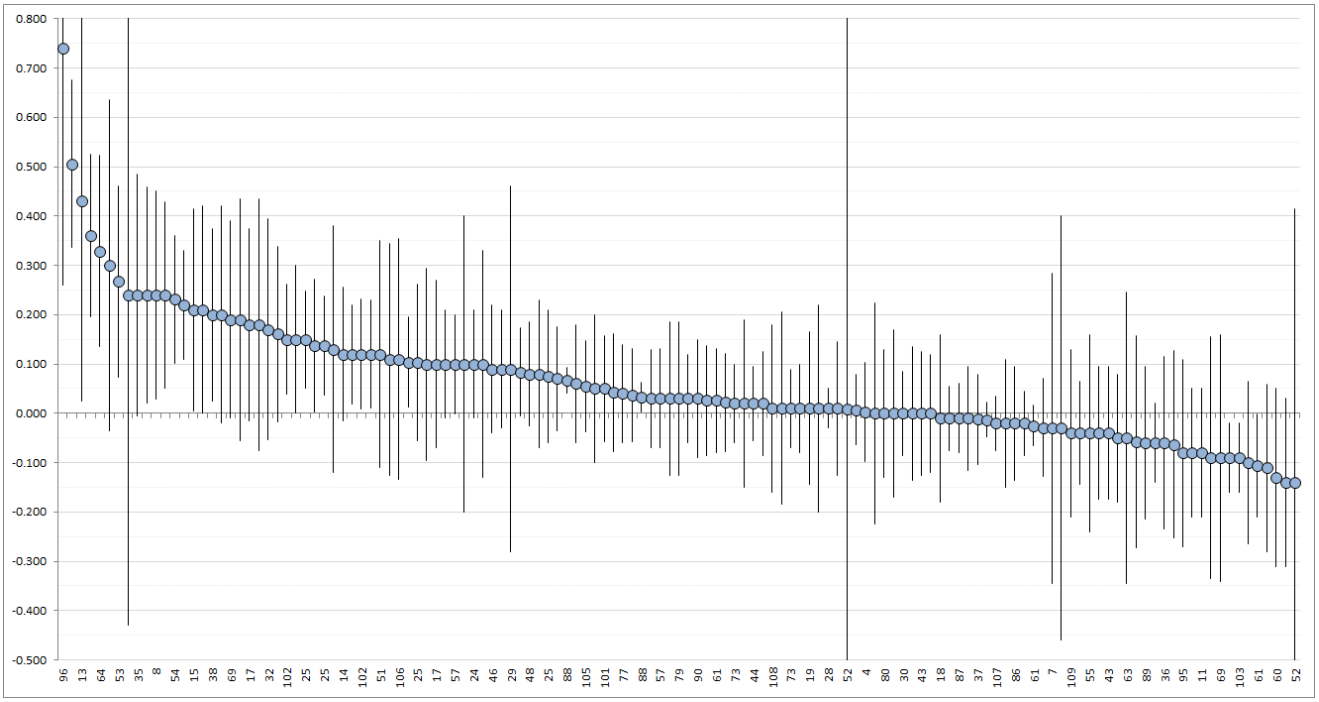
To unpack the statistical theory behind the meta-analyses, we first consider a fixed-effects approach. This is then extended to a form random-effects model.

Figure D2: Forrest chart showing effect sizes and 95% CIs across 133 primary ITT outcomes (ordered by trial ID)



[38] Vogt & Johnson (2011).

Figure D3: Forrest chart showing effect sizes and 95% CIs across 133 primary ITT outcomes (rank ordered by mean effect size)



Blue dot = reported mean effect size; **Bars** = upper/lower 95% CI

Fixed effects meta-analysis

Calculating an overall (summary) effect size – weighted mean (M)

See chapters 10–14 of Borenstein et al. (2009)

Meta-analyses weights data in order for the analysis to account for the statistical uncertainty behind each effect size in a review. When a fixed effects model is assumed, the weighting is calculated using equation 11.2 from Borenstein et al. (2009)

11.2 (p. 65): $W_i = \frac{1}{V_{Y_i}}$

Where is the W_i (fixed effects model) weighting and $V_{Y_i}$is the variance of effect size Y for evaluation i. $V_{Y_i}$is estimated by squaring the standard error for the effect size (i.e., $V_{Y_i} = (SE_{Y_i})^2$).

Equation 11.3 shows how to calculate the summary effect size (i.e., the weighted mean), M:

11.3 (p. 66): $M_{FE} = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i}$

For the 133 primary ITT effect sizes, $M_{FE} = \frac{\sum_{i=1}^{133} W_i Y_i}{\sum_{i=1}^{133} W_i} = +0.031$

The variance of M_{FE} is shown in equation 11.4

11.4 (p. 66): $V_{M_{FE}} = \frac{1}{\sum_{i=1}^{k} W_i}$ which can be converted into a standard error $SE_{M_{FE}} = \sqrt{V_{M_{FE}}}$

and this standard error can be used to calculate 95% CI for the summary effect $M_{FE} \pm 1.96(SE_{M_{FE}})$

For the 133 primary ITT effect sizes:

$V_{M_{FE}} = \frac{1}{\sum_{i=1}^{k} W_i} = +0.0000199$; $SE_{M_{FE}} = 0.00447$ and the lower and upper 95% CI for M_{FE}: +0.022; +0.040 [99%: +0.019; +0.042]

To summarise, assuming fixed effects results in an estimated weighted mean effect size of +0.03 (se = 0.004) with a 95% CI between +0.02 and +0.04.

The 95% CI are both positive and so this illustrates that, whilst small, the weighed mean effect size is statistically significantly greater than zero. This can also be tested directly using a z-test (not shown).

Random effects meta-analysis

Calculating an overall (summary) effect size-weighted mean (M)*

See chapters 10–15 of Borenstein et al. (2009)

If it is unreasonable to assume a single 'true' effect size across all cases in a meta-analysis, the assumption is that there will be some heterogeneity in true effect sizes. To address this, a random effects model is used.

For random effects models, Borenstein et al. introduce a key parameter – tau (τ) and tau-squared (τ^2). Tau (τ) is the standard deviation (or variance when squared) of the distribution of 'true effects' across studies. This means that a single 'true effect' is no longer assumed. This is more suitable for the EEF review given the wide variety of programmes and outcomes included in the analyses.

τ^2 is estimated using T^2 and defined as the 'between studies variance' and 'the variance of the effect size parameters across the population of studies' (p. 72).

Equation 12.2: $T^2 = \frac{Q-df}{C}$

Q, df and C are defined in equations 12.3 to 12.5

The Q statistic is the weighted sum of squares (WSS, p109) and this can be used to evaluate whether a random effects model is suitable. Q can be calculated using Equation 12.3:

$Q = \sum_{i=1}^{k} W_i {Y_i}^2 - \frac{\left(\sum_{i=1}^{k} W_i Y_i\right)^2}{\sum_{i=1}^{k} W_i}$ = 284.22 for 133 Primary ITT effect sizes

The expected value of Q assuming that all cases share a common 'true' effect size (i.e., fixed effects) is shown to be the degrees of freedom;

Equation 12.4: $df = k - 1$ (k = number of effect sizes) = 132

Because Q is the observed WSS and df is the expected WSS assuming all studies share a common effect, Q – df measures the excess variation (i.e., differences in the true.)

Excess variation = $(Q - df)$ = 152.22

Note: $(Q - df) \geq 0$. When $(Q - df) = 0$; $Q = df$ there is no excess variation and so the assumption of a single 'true' effect size is statistically justifiable (i.e., Fixed Effects is OK). It is also technically possible for $(Q - df) < 0$ but in these cases, the statistic (and T^2 to follow) are set to zero (pp. 109–113). When $(Q - df) > 0$; $Q > df$

Equation 12.5: $C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} {W_i}^2}{\sum_{i=1}^{k} W_i}$ = 48383.81

$T^2 = \frac{Q-df}{C}$ = 0.00315

As seen with fixed-effects, a random-effects meta-analyses weights data in order for the analysis to account for the statistical uncertainty behind each effect size in the review.

When a random effects model is assumed, the weighting is calculated using equation 12.6 from Borenstein et al. (2009) 12.6 (p. 73): $W_i^* = \frac{1}{V_{Y_i}^*}$

Where is the W_i^* (random effects model) weighting and $V_{Y_i}^*$is the within-study variance of effect size Y for evaluation **PLUS** the between studies variance, τ^2 (i.e., $V_{Y_i}^* = V_{Y_i} + T^2$).

$V_{Y_i}$is estimated by squaring the standard error for the effect size (i.e., $V_{Y_i} = (SE_{Y_i})^2$).

Equation 12.7 shows how to calculate the summary effect size (i.e., the weighted mean), M*: 12.7 (p. 73): $M^* = \frac{\sum_{i=1}^{k} W_i^* Y_i}{\sum_{i=1}^{k} W_i^*}$

For the 133 primary ITT effect sizes, $M_{RE}{}^* = \frac{\sum_{i=1}^{133} W_i^* Y_i}{\sum_{i=1}^{133} W_i^*} = +0.043$

The variance of the summary effect is shown in equation 12.8

11.4 (p. 66): $V_{M^*} = \frac{1}{\sum_{i=1}^{k} W_i^*}$ which can be converted into a standard error $SE_{M^*} = \sqrt{V_{M^*}}$

..and this standard error can be used to calculate 95% CI for the summary effect $M^* \pm 1.96(SE_{M^*})$

For the 133 primary ITT effect sizes:

$V_{M_{RE}^*} = \frac{1}{\sum_{i=1}^{k} W_i} = +0.0000594$; $SE_{M_{RE}^*} = 0.00771$ and the lower and upper 95% CI for M_{FE}: +0.028; +0.058 [99%: +0.023; +0.063]

To summarise, assuming random effects results in an estimated weighted mean effect size of +0.04 (se = 0.008) with a 95% CI between +0.03 and +0.06.

Adopting a random effects model results in an increased summary effect with wider CI (which reflects the increased variance introduced by T^2).

Table D2: Weighted meta-analyses of effect sizes

	No outcomes / effect sizes	Weighted Mean (SE)	Lower	Upper
Descriptive analyses	133	+0.06 (n/a)	–	–
Meta-analysis assuming fixed effects	133	+0.03 (0.004)	+0.02	+0.04
Meta-analysis assuming random effects	133	+0.04 (0.008)	+0.03	+0.06

It was the random-effects approach used in the review but calculations were undertaken using the 'metafor' R package[39] with the output agreeing with the random effects model calculated by hand above.

[39] Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1–48. https://www.jstatsoft.org/v36/i03/

Meta-analyses with subgroups

All of the explanatory variables included in the review were categorical in nature therefore meta-analyses with subgroups was used to estimate weighted means for each category in a variable. The statistical theory behind the weighted mean estimates is illustrated in two ways.

- First, using a variable from the Evaluation Design theme to estimate the weighted mean primary ITT effect size for RCT and clustered RCT trials.

Second, using a variable from the implementation *& fidelity* theme to estimate the weighted mean FSM effect size for X different types of developers.The first example is used to illustrate the calculations undertaken by the R Metafor package and the second is included because the meta-analyses that included this variable within Metafor resulted in the estimated model parameters failing to converge[40]. The first example is to replicate whilst the second is to manually supplement output from the R Metafor package.

Evaluation design theme: RCT and clustered RCT trial designs

Of the 133 Primary ITT effect sizes in the review, 41 were reported by evaluations with an RCT design whilst 92 were reported by evaluations with a clustered RCT (CRT) design. The descriptive analyses show that the mean effect sizes from RCTs is higher (+0.11 SD) compared with the mean effect size from CRTs (+0.02 SD).

First, assuming fixed effects, the summary effect (M_{FE}), its variance ($V_{M_{FE}}$) and standard error ($SE_{M_{FE}}$)are calculated for each group and for the whole sample as shown above.

Table D3: Effect size by trial design (primary ITT attainment outcomes) – fixed effects model

	k	M_{FE}	$V_{M_{FE}}$	$SE_{M_{FE}}$	Lower	Upper
RCT	41	+0.08	0.00019	0.014	+0.05	+0.11
CRT	92	+0.02	0.00002	0.005	+0.02	+0.03
ALL	133	+0.03	0.00002	0.004	+0.02	+0.04

Table D4: Trial design: calculating Q* and T*2

	Q	$Q - df$	T^2
RCT	93.7	53.7	0.01071
CRT	175.6	84.6	0.00197
ALL	284.2	152.2	0.00315

This process provides the T^2 estimate for τ^2. Borenstein et al. (2009) note two approaches for the random effects model; to use separate T^2 estimates for each subgroup or to use a pooled T^2 estimate (pp. 162–163). If it is assumed that the case-to-case dispersion is the same for RCT and CRT designs – then observed differences in T^2 must be due to sampling variation and so a pooled estimate of T^2 is appropriate. However, if the between-case dispersion for one group (e.g., RCTs) will be different to another (CRTs), we would use separate estimates of T^2 for each group.

For the review, we assumed that the between-case dispersion across categories of explanatory variables would be different and so used separate estimates of T^2 for RCT and CRT designs.

This means that for each of the 41 RCT effect sizes, a value of 0.01071 (T^2_{RCT}) is added to the variance and for each of the 92 CRT effect sizes, a value of 0.00197 (T^2_{CRT}) is added to the variance. The weighting process is then redone to provide the random effects weighted mean estimate with 95% CI.

[40] Error in rma …Fischer scoring algorithm did not converge (R 'metafor' error message). On investigation, this relates to the Tau^2 estimate (T^2) for the council/local authority grouping which was close to zero but negative (-0.003). Whilst Tau^2 cannot be negative, methods to estimate this with T^2 can result in negative values (Borenstein et al., 2009). The manual approach to resolve this is to set the T^2 estimate to zero (Borenstein et al., 2009) but seems to have presented the R Metafor package with a problem.

Table D5: Effect size by trial design (primary ITT attainment outcomes) – random effects model

	k	M_{RE}	$V_{M_{RE}}$	$SE_{M_{RE}}$	Lower	Upper
RCT	41	+0.10	0.00049	0.022	+0.05	+0.14
CRT	92	+0.03	0.00006	0.008	+0.01	+0.04

These estimates agree with those produced by R 'metafor' for the EEF review.

FSM attainment outcomes, implementation & fidelity theme: types of developers

One particular meta-analysis hit problems in the R metafor package and so was undertaken by hand drawing on Borenstein et al. (2009). Specifically, the meta-analysis that estimated weighted mean effect sizes for FSM attainment outcomes across categories of the 'types of developers' explanatory variable found within the 'implementation & fidelity' overarching theme. This section sets out how the weighted mean effect sizes for these meta-analyses were estimated by hand.

First, assuming fixed effects, the summary effect (M_{FE}), its variance ($V_{M_{FE}}$) and standard error ($SE_{M_{FE}}$) were calculated for each group and for the whole sample.

Table D6: Effect size by types of developers (FSM attainment outcomes) – fixed effects model

	k	M_{FE}	$V_{M_{FE}}$	$SE_{M_{FE}}$	Lower	Upper
Charity / non-profit	47	+0.02	0.00013	0.012	0.00	+0.05
University	41	+0.03	0.00012	0.011	0.00	+0.05
Private company	24	–0.01	0.00033	0.018	–0.04	+0.03
School / MAT	10	+0.12	0.00148	0.038	+0.04	+0.19
Council / LA	15	–0.01	0.00070	0.026	–0.06	+0.05
Mixed	12	+0.07	0.00102	0.032	+0.01	+0.13
ALL	149	+0.02	0.00005	0.007	+0.01	+0.04

Table D7: Types of developers: calculating Q* and T*2

	Q	$Q - df$	T^2
Charity / non-profit	102.7	56.7	0.00821
University	49.8	9.8	0.00122
Private company	36.0	13.0	0.00460
School / MAT	11.5	2.5	0.00534
Council / LA	10.2	–3.8	–0.0033
Mixed	13.5	2.5	0.05476
ALL	235.7	87.7	0.00405

If it is assumed that the case-to-case dispersion is the same for all types of developers – then observed differences in T^2 must be due to sampling variation and the use of a pooled estimate of T^2 is appropriate. However, if the between-case dispersion for one group is different to another, separate estimates of T^2 for each group are used.

For the review, we assumed that the between-case dispersion across categories of explanatory variables would be different and therefore used separate estimates of T^2.

This means that:

- for each of the 47 effect sizes within the 'charity/non-profit' developer category, a value of 0.00821 ($T^2_{Charity}$) is added to the variance
- for each of the 41 University developer effect sizes, a value of 0.00122 (T^2_{CRT}) is added to the variance
- for each of the 24 effect sizes within the 'private company' developer category, a value of 0.00461 (T^2_{PrivCo}) is added to the variance
- for each of the 10 effect sizes within the 'School/MAT' developer category, a value of 0.00534 (T^2_{School}) is added to the variance
- for each of the 14 effect sizes within the 'Council/LA' developer category, a value of zero is added to the variance. This is because for this category Q is (10.2) is lower than the degrees of freedom (14). Borenstein et al. (2009) specify that this then sets $T^2_{Council}$ to be zero
- for each of the 12 effect sizes within the 'Mixed' developer category, a value of 0.002999 (T^2_{Mixed}) is added to the variance

In undertaking these calculations, a potential reason for the R Metafor package failing to converge was identified. This relates to the estimated T^2 value for the 'Council/LA' category of the 'types of developers' variable being less than zero ($T^2_{Council} < 0$). When undertaken manually, this value can be set to zero (as advised by Borenstein et al. (2009)) but this may have presented R Metafor with a problem. Follow-on analyses that excluded the 'Council/LA' category did not have problems converging but analyses that excluded another category but included the 'Council/LA' category did.

The T^2 values can then be used to calculated estimated weighted mean effect sizes across categories of the 'types of developers' variable using the random effects model, as shown below in Table D8.

Table D8: Effect size by types of developers (FSM attainment outcomes) – random effects model

	k	M_{FE}	$V_{M_{FE}}$	$SE_{M_{FE}}$	Lower	Upper
Charity / non-profit	47	+0.03	0.0004	0.020	–0.01	+0.07
University	41	+0.03	0.0002	0.013	0.00	+0.05
Private company	24	–0.02	0.0006	0.025	–0.07	+0.03
School / MAT	10	+0.14	0.0026	0.051	+0.04	+0.24
Council / LA	15	–0.01	0.0007	0.026	–0.06	+0.04
Mixed	12	+0.06	0.0014	0.038	–0.01	+0.13

This document is available for download at https://educationendowmentfoundation.org.uk

Education Endowment Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

https://educationendowmentfoundation.org.uk
@EducEndowFoundn
Facebook.com/EducEndowFoundn