REVIEW OF EEF PROJECTS

Summary of Key Findings.

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)

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; and
- 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

Overview

This document summarises key findings from the quantitative strands of a **review of the Education Endowment Foundation (EEF) evaluations** that had reported from the establishment of EEF in 2011 up to January 2019. The quantitative strands summarised include meta-analyses of effect sizes reported for attainment outcomes and descriptive analyses of cost-effectiveness and attrition. Additionally, descriptive univariate analyses of the explanatory variables used in the review are summarised. Complete findings can be found in the main report (Demack et al., 2021). The review conducted qualitative interviews to explore perceptions on effective scale-up and developed and piloted an IPE quality measure, and these are reported separately (Maxwell et al., 2021a & 2021b respectively).

The review initially focused on effect sizes reported for the headline 'intention-to-treat' (ITT) analyses of primary attainment outcomes for 82 EEF trials that had reported by January 2019. As might be expected, even with a common 'attainment' focus, the effect sizes were extremely heterogeneous in nature. For example, some related to GCSE attainment overall or in a specific subject; others to commercial tests at the end of KS3; others to tests at the end of KS2 etc. This breadth in age range and outcome tests prescribed a descriptive approach rather than a more traditional systematic review meta-analysis that focused in on a specific pupil age, intervention type and outcome measure. To provide necessary structure, a theoretical framework was developed with five over-arching themes:

1 The intervention
2 Use of theory & evidence
3 The context of the intervention and evaluation
4 Implementation & fidelity of the intervention
5 The evaluation research design.

The development of this theoretical framework was an iterative process conducted in the early stages of the review project, which resulted in a total of 65 explanatory variables grouped under these five overarching themes.

This framework was adapted for follow-on meta-analyses of effect sizes reported for secondary outcomes that measured pupil attainment, and effect sizes reported for free school meals (FSM) subsample analyses of primary or secondary attainment outcomes. Similarly, the same framework was adapted for the descriptive trial-level analyses into the cost effectiveness of interventions and pupil-level attrition.

Outcome variables

Three groupings of effect sizes reported for attainment outcomes reported by the 82 EEF trials included in the review were included in the meta-analyses:

1 133 effect sizes reported for ITT analyses of primary attainment outcomes (reported by 82 trials)
2 78 effect sizes reported for ITT analyses of secondary attainment outcomes (35 trials)
3 149 effect sizes reported for FSM subsample analyses of primary or secondary attainment outcomes (73 trials).

Effect sizes for psychological outcomes were also collected but unsuitable for inclusion in the meta-analyses. The theoretical framework and selected explanatory variables focused on hypothesised associations with effect sizes reported for attainment outcomes. Psychological outcomes were distinct from attainment and more disparate in nature. Future reviews might examine specific types of psychological outcomes and develop a tailored theoretical framework to help structure a meta-analysis. For this review, the disparate non-attainment nature of psychological outcomes resulted in not including these in the meta-analyses. Please see the main report for detail on the distribution of psychological outcomes and our thoughts on how these might be classified within future reviews.

Cost effectiveness was measured at the evaluation level and defined as the cost per pupil for an effect size of +0.10 SD **given** evidence of a positive impact. Forty of the 82 evaluations in the review (49%) were included in the cost effectiveness outcome and analysis. 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.

Pupil attrition was also measured at the evaluation level and was obtained for 79 of the 82 evaluations, and a mean reported attrition rate of 19% for ITT analyses of primary outcomes was observed.

Explanatory variables

The 65 explanatory variables were collected under five overarching themes key findings from the distribution of these variables are noted under each theme below.

Intervention theme [22 variables]

- A majority of interventions took place in primary schools (51 evaluations, 62%), particularly KS2 (33 evaluations, 40%); 30% (*n* = 25) took place in secondary schools, most commonly KS3 (20 evaluations, 24%).
- 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%).
- The overall mean intensity of interventions was just over 90 minutes per week.
- Teacher-led interventions were the most common (37 evaluations, 45%) followed by externally-led interventions (18 evaluations, 22%) and TA-led interventions (12 evaluations, 15%).
- 52 of the 82 evaluations in the review reported on the quality of supporting resources for an intervention. Among those that did report on this issue, quality was perceived to be variable for over half of the reported projects (*n* = 27 evaluations, 52% that reported on this) and high in just under 40% of evaluations that reported on supporting resources (*n* = 20).
- The mean total cost of delivery of interventions across the 82 trials in the review was just under £500k (median ~£470k), fluctuating over time from a mean and median under £370k in 2014 and 2015, then increasing to a peak mean of over £700k (median just over £600k) in 2017 before falling to just under £500k (median = £450k) in 2018. The mean cost per pupil appears to have reduced over time and for interventions in this review the mean cost was £174 (median = £54).

Theory & evidence theme [three variables]

- Strong prior empirical evidence was presented for interventions in only 21% of reports in the review (*n* = 17 evaluations) and, similarly, highly detailed theoretical discussion was reported in 21% of reports (*n* = 17 evaluations) but few reports presented both strong empirical evidence **and** strong theoretical detail (*n* = 5 evaluations, 6%). Strong empirical evidence and theoretical detail were observed to increase over time, but theoretical detail was more often minimal or omitted than empirical evidence. The increase in empirical and theoretical evidence found may be a result of evaluators providing more detail over time, rather than the interventions being trialled having stronger empirical and theoretical foundations.
- Most interventions (69 evaluations, 84%) focused on pupil learning, 11% focused on wider pupil outcomes (*n* = 9 evaluations) and 4% solely on teacher change outcomes (*n* = 3 evaluations).

Context theme [nine variables]

- The most commonly perceived organisational barrier was staff time and availability (54 evaluations, 66%) followed by specialist facilities and space (35 evaluations, 43%) and workforce capacity (31 evaluations, 38%). In relation to individual characteristics, pupil behaviour was mentioned as a barrier to implementation in 32% of evaluations (*n* = 26); and staff expectations and/or motivations were perceived to be an enabler in 18 evaluations (22%) and a barrier in 12 (15%). Preparations for Ofsted and/or staff perceptions of Ofsted's requirements conflicting with the intervention method were mentioned in 20% of evaluations (*n* = 16).

Implementation & fidelity theme [15 variables]

- Charities were the most common developers of interventions across the evaluations in the review (32 evaluations, 39%) followed by universities (19 evaluations, 23%), private companies and individual schools or academy trusts (*n* = 9 in both categories, 11%) and local authorities (*n* = 8 evaluations, 10%).
- Lead-in time for implementation was perceived to be insufficient in over half of the evaluations (*n* = 24 evaluations, 56%) where lead-in time was mentioned (43 of the 82 evaluations mentioned this), with only 12% of those 82 evaluations indicating that it was sufficient (*n* = 5 evaluations).
- The vast majority of programmes in the review provided one or more forms of Continuing Professional Development (CPD) to support implementation of the intervention (*n* = 77 evaluations, 94%[1]), most commonly taking place before and during the intervention (47 evaluations, 57%). Around 60% of this CPD was subject-

[1] Note that this includes some of the interventions that involved direct delivery by external organisations; in these cases, the CPD was for staff in these external delivery organisations.

specific or curriculum-specific (*n* = 49 evaluations); 65% of CPD was delivered direct by the intervention developer (*n* = 53 evaluations); 90% of interventions included face-to-face CPD (*n* = 74 evaluations); 16% included mentoring and coaching (*n* = 13 evaluations); and 13% involved online training (*n* = 11 evaluations). A majority of developers (60 evaluations, 73%) provided support other than group training sessions, for example through classroom visits and email support.

- Just over half of the evaluations reported some form of monitoring of implementation (42 evaluations, 51%); this was usually done by the delivery partner but varied in nature and intensity across the trials.
- Senior leader support for implementation was mentioned in just under half of all reports and, contrary to the findings of intervention studies in the wider evidence base, few of the evaluations in this review reported that support was limited or minimal support (5 evaluations, 6% of all evaluations, 13% of the 38 evaluations that mentioned SLT support).
- A majority of the evaluations in the review mentioned intended fidelity (68 evaluations, 83%). Of the evaluations that did mention intended fidelity, a small majority indicated that the intervention (by the direct implementer) was intended to be adopted faithfully (37 evaluations, 54%) (54%), and a notable proportion reported that the intervention was adapted to context (31 evaluations, 46%). The majority of evaluations in the review provided details on the actual fidelity of implementation (73 evaluations, 89%). The most common report of actual fidelity was that it varied across schools/settings (46 evaluations, 65% of evaluations that mentioned actual fidelity) but 18% reported high fidelity (*n* = 13) and 19% reported limited fidelity (*n* = 14). Turning to fidelity relating to CPD compliance, a similar pattern of varied fidelity was found in the 44 evaluations (54%) that had sufficient data to assess fidelity.

Evaluation design theme [16 variables]

- The majority of evaluations in the review had a clustered RCT trial design (55 evaluations, 67%), usually with school-level randomisation (48 evaluations, 59%), and clustered CRT designs were observed to become more common over time and comprise 90% of reports published between 2017 and 2019. Using the classification provided by EEF,[2] half of the evaluations were classed as efficacy trials and half as effectiveness trials. A majority of evaluations were undertaken by a university (52 evaluations, 63%).
- The length and size of trial appears to be associated with the type of trial design. On average, evaluations that used a randomised controlled trial (RCT) design were shorter and involved fewer schools and individuals than the evaluations that used a clustered RCT (CRT) design. Intervention length fluctuated over time: the mean length was 16 weeks (median = 13.5 weeks) for the first 16 evaluations published in 2014 rising to a peak of 77 weeks (median = 97 weeks) in 2017 before falling to 48.5 weeks (median = 45 weeks) in 2018. The number of participating schools ranged between three and 205 with a mean of 64 schools, and the number of individual participants (usually pupils) ranged between 36 and 25,000 with a mean of about 3,700 individuals.
- A fall in attrition rates was found from 2014 to 2018 while mean padlock ratings have increased over the same time period.
- Evaluations were commonly found to create a burden on schools. Only 11% (*n* = 9 evaluations) undertook no testing and a majority (49 evaluations, 60%) collected two or more external tests from schools, although the burden of testing can vary in intensity depending on the test administered. A small majority of evaluations (43 evaluations, 52%) collected both survey and interview data from schools and 33% collected survey or interview data but not both.
- The majority of the 82 evaluations in the review used a single primary outcome (50 evaluations, 61%) and the use of a single primary outcome became increasingly common over time. In most evaluations, the review found a direct match between the intervention focus and the primary outcome(s) (47 evaluations, 57%), but in 12% this alignment was limited (10 evaluations).
- Across the 133 effect sizes reported by the 82 evaluations in the review, primary outcomes were most commonly commercial tests (79 effect sizes, 59%) with Granada Learning (GL) being the most common provider (46 effect sizes, 35%). A sizeable minority of outcomes were taken from the National Pupil Database (NPD) (45 effect sizes, 34%) with KS2 SATs being the most common of these (30 effect sizes, 23%).
- 13 specific outcome measures accounted for 68% of the reported effect sizes in the review, with GL New Group Reading Test (GL-NGRT) being the most common specific outcome (13 effect sizes, 17%). Whilst these 13 outcomes are 'specific', this does **not** mean that they are 'the same'. This is because of differences in pupil age year groups that took the 'specific' tests.

[2] There is some variation in 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 the 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 classification contrasts with the one used in the previous IoE review. Restricting the sample to RCT/CRT designs, the IoE data shows 26 efficacy and 21 effectiveness trials. The EEF classification lists four of the IoE efficacy trials as being effectiveness trials and three of the IoE effectiveness trials as efficacy trials.

Meta-analyses of effect sizes reported for attainment outcomes

The meta-analyses of effect sizes reported for ITT analyses of primary attainment outcomes, ITT analyses of secondary attainment outcomes and FSM subsample analyses of primary or secondary attainment outcomes are summarised here. Whilst all three effect size groupings relate to outcomes measuring pupil attainment, comparing meta-analyses findings needs to be done with caution. This is perhaps most critical when comparing the findings for secondary ITT effect sizes. This is because secondary ITT attainment effect sizes were much rarer than primary ITT or FSM effect sizes; they were reported by 35 of the 82 trials included in the review. This means that observed differences between the secondary ITT and other effect size groupings might be due to this notable sample difference. FSM effect sizes were reported by 73 of the 82 trials and therefore do not suffer the same extent of sample issues seen with secondary ITT effect sizes. However, caution is also needed here because the FSM effect sizes relate to both primary and secondary attainment outcomes. Readers are advised to consider findings for each effect size grouping separately before attempting to draw (cautious) comparisons. Finally, please note that all meta-analyses are bivariate and descriptive, and it is therefore not appropriate to draw causal conclusions from the findings. Future systematic reviews might draw on these descriptive findings to help develop an analytical framework with a narrower focus (e.g., pupil age and/or subject area) and adopt multivariate approach for the meta-analyses.

The selection of explanatory variables for inclusion in the meta-analyses summary was by drawing on statistical details (such as statistical significance) and discussion within the review research team. To summarise, a four-stepped approach was taken. First, when a statistically significant association between an explanatory variable and effect size was observed, the variable was flagged for inclusion. Second, each of the flagged explanatory variables was examined and one or more category that had a relatively high or low effect size was selected for inclusion in the summary. Third, the review team agreed the final selection of explanatory variables and categories. Fourth, the review team examined instances where the meta-analyses did not find a statistically significant association, to identify other explanatory variables and categories for inclusion in the summary. This mixture of the objective and subjective reflects the methodological approach of the review of EEF evaluations; purposely broad and descriptive within a thematic theoretical structure.

In this summary, we refer to weighted (meta-analyses) mean effect sizes as being relatively high or low. As outlined above, this was guided in part by whether a statistically significant association was observed. For each of the three effect size groupings, the size of weighted mean effect size for each selected category was considered relative to the overall weighted mean effect size. In other words, effect sizes that are relatively high or low for the primary ITT were relative to the overall primary ITT weighted mean effect size (+0.04 SD). Similarly, for secondary ITT effect sizes (relative to +0.01 SD) and FSM effect sizes (relative to +0.03 SD).

Table 1 summarises the overall weighted mean effect size for primary ITT, secondary ITT and FSM attainment effect sizes, and lists the explanatory variables selected for inclusion in this summary grouped under each of the five overarching themes. The number of explanatory variables and the number of categories selected for inclusion are shown. For example, for primary ITT effect sizes, the summary includes 26 categories from 21 explanatory variables [21 (26)]. Table 1 also indicates when the association between and explanatory variable and effect size was observed to be statistically significant ($p < 0.10$) or not, by using an asterisk (*****).

Table 1: Table summarising selected explanatory variables from meta-analyses of attainment outcomes

	Primary ITT	Secondary ITT	FSM
Overall meta-analysis weighted mean Hedges' *g* effect size	+0.04	+0.01	+0.03
95% CI	+0.03; +0.06	−0.01; +0.03	+0.01; +0.05
Number effect sizes	133	78	149
Number of trials	82	35	73
Selected explanatory variables (categories)#			
All	21 (26)	24 (34)	17 (20)
Intervention theme: selected explanatory variables	6 (8)	7 (10)	5 (7)
	Key Stage curriculum focus* direct implementer** total cost** cost per pupil** EEF promising**	Key Stage** curriculum focus** direct implementer* quality of resources* total cost** cost per pupil* EEF promising**	Key Stage* direct implementer quality of resources* total cost* EEF promising*
Theory & evidence theme: Selected explanatory variables	2 (2)	2 (4)	2 (2)
	empirical evidence** focus of change**	empirical evidence** theoretical detail*	empirical evidence* focus of change**
Context theme: selected explanatory variables	3 (4)	4 (4)	2 (2)
	geography** specialist facilities** staff teamwork**	specialist facilities* staff teamwork* staff expectations* pupil behaviour*	geography specialist facilities*
Implementation & fidelity theme: selected explanatory variables	5 (5)	6 (8)	4 (4)
	type of developer* CPD provision CPD sequencing subject-specific CPD CPD fidelity*	type of developer* clarity of plan* implement support* CPD provision* CPD sequencing* CPD fidelity*	type of developer* monitoring* subject-specific CPD CPD fidelity
Evaluation design theme: selected explanatory variables	5 (7)	5 (8)	4 (5)
	trial design* level of randomisation* no. of pupils* alignment between intervention & primary outcome* type of outcome	Evaluator type* no. of schools* evaluation length* EEF padlocks* type of outcome*	level of randomisation no. of schools* evaluation length* EEF padlocks*

Key: # The specific explanatory variables included in this summary are shown first (e.g., 21 explanatory variables are included from the primary ITT analyses). All of the explanatory variables included in the meta-analyses were categorical in nature (i.e., nominal or ordinal) and this summary selects one or more category from each of the included explanatory variables. The number of categories selected is shown in brackets. For example, in the primary ITT summary, a total of 26 categories from the 21 selected explanatory variables are included.
***** indicates when the association between effect size and explanatory variable was statistically significant at $p = 0.10$ or lower.

Primary ITT effect sizes

Figure 1 shows error bars and weighted (meta-analyses) mean effect sizes for a selection of categories within explanatory variables organised into the five overarching themes.

The overall weighted mean for all 133 primary ITT effect sizes (+0.04 SD) that was reported by the 82 evaluations in the review is shown at the top of Figure 1 as a solid black circle and as a dotted red vertical red line running parallel to the categorical axis. For each weighted mean, 95% confidence intervals (CI) are shown (+0.03 to +0.06 SD for all 133 effect sizes).

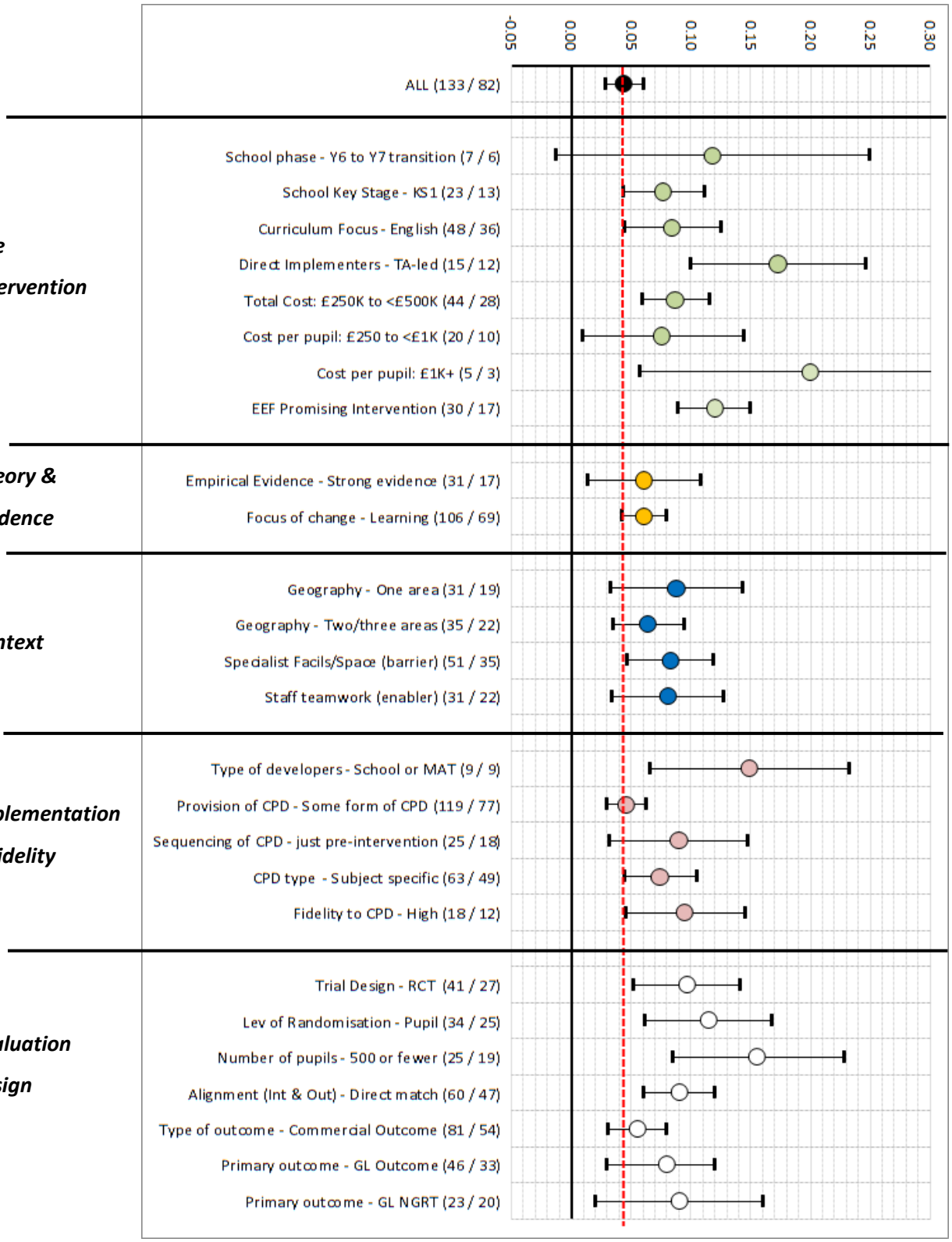
Mean weighted effect size estimates for 26 selected categories from 21 explanatory variables are shown using circles colour coded according to the overarching theme where the explanatory variable was placed.

In most cases, the selected category is drawn from an explanatory variable found to be statistically significantly associated with primary ITT effect sizes and where the category was notably higher than the overall weighted mean (of +0.04 SD).

Figure 1: Effect sizes for ITT analyses of primary attainment outcomes

Meta-analyses summary using error bars

Weighted mean effect sizes and 95% CI from meta-analyses for selected categories of explanatory variables across the five overarching themes.



This was not the case for the two error bars relating to school phase/Key Stage and the three error bars relating to the provision of CPD, sequencing and types of CPD. Additionally, Figure 1 provides closer detail on commercial primary outcome measures. These are included in Figure 1 for interest.

Intervention theme

Eight categories from six variables: school Key Stage, curriculum focus, direct implementers, total cost, cost per pupil and EEF promising interventions.

Relatively **higher** weighted mean effect sizes were associated with:

- interventions involving Year 6 to Year 7 transition (weighted mean effect size = +0.12 SD; 95% CI: −0.01; +0.25 seven primary ITT effect sizes reported by six evaluations) or KS1 pupils (+0.08 SD; 95% CI: +0.04; +0.11; 23 primary ITT effect sizes reported by 13 evaluations). Note that this association was not statistically significant (p > 0.10).
- interventions with an English curriculum focus (+0.08 SD; 95% CI: +0.04; +0.13; 48 primary ITT effect sizes reported by 36 evaluations).
- TA-led interventions (+0.17 SD; 95% CI: +0.10; +0.25; 15 effect sizes reported by 12 evaluations).
- interventions with a **total cost** of from £250k to less than £500k (+0.09 SD; 95% CI: +0.10; +0.25; 44 effect sizes reported by 28 evaluation).
- interventions with a **per pupil cost** of more than £1,000 (weighted mean effect size = +0.20 SD; 95% CI: +0.06; +0.34; five primary ITT effect sizes reported by three evaluations) or from £250 to £1,000 (weighted mean effect size = +0.08 SD; 95% CI: +0.01; +0.14; 20 primary ITT effect sizes reported by 10 evaluations).
- interventions that were identified as 'promising' by EEF (weighted mean effect size = +0.12 SD; 95% CI: +0.09; +0.15; 30 effect sizes reported by 17 evaluations).

Theory & evidence theme

Two categories from two variables: use of empirical evidence and focus of change.

Relatively **higher** weighted mean effect sizes were associated with:

- strong empirical evidence for the intervention being present in the evaluation report (weighted mean effect size = +0.06 SD; 95% CI: +0.01; +0.11; 31 primary ITT effect sizes reported by 17 evaluations).
- interventions with a learning focus of change (+0.06 SD; 95% CI: +0.04; +0.08; 106 effect sizes reported by 69 evaluations).

Context theme

Four categories from three variables: geography, specialist facilities/space and staff teamwork.

Relatively **higher** weighted mean effect sizes were associated with:

- interventions/evaluations taking place in one geographical area (weighted mean effect size = +0.09 SD; 95% CI: +0.03; +0.14; 31 primary ITT effect sizes reported by 19 evaluations), or in two/three geographical areas (weighted mean effect size = +0.06 SD; 95% CI: +0.04; +0.09; 35 primary ITT effect sizes reported by 22 evaluations).
- reported perceptions that specialist facilities/space were a barrier for the intervention (+0.08 SD; 95% CI: +0.05; +0.12; 51 effect sizes reported by 35 evaluations).
- reported perceptions that staff teamwork was an enabler for the intervention (+0.08 SD; 95% CI: +0.03; +0.13; 31 effect sizes reported by 22 evaluations).

Implementation & fidelity theme

Five categories from five variables: type of developer, provision of CPD, sequencing of CPD, subject-specific CPD and CPD fidelity.

Relatively **higher** weighted mean effect sizes were associated with:

- interventions/evaluations developed by schools, academies or multi-academy trusts (weighted mean effect size = +0.15 SD; 95% CI: +0.07; +0.23; 9 primary ITT effect sizes reported by 9 evaluations).
- when some form of CPD was part of the intervention (+0.05 SD 95% CI: +0.03; +0.06; 119 primary ITT effect sizes reported by 77 evaluations). Note that this association was not statistically significant (p > 0.10).
- when CPD was provided before the intervention (+0.09 SD 95% CI: +0.03; +0.15; 25 primary ITT effect sizes reported by 18 evaluations). Note that this association was not statistically significant (p > 0.10).
- when CPD was subject-specific (+0.07 SD; 95% CI: +0.04; +0.11; 63 primary ITT effect sizes reported by 49 evaluations). Note that this association was not statistically significant (p > 0.10).

- when fidelity of training delivered to intended CPD was high (+0.09 SD; 95% CI: +0.05; +0.14; 18 effect sizes reported by 12 evaluations).

Evaluation design theme

Seven categories from five variables: trial design, level of randomisation, number of pupils, trial sensitivity, alignment between intervention and primary outcome.

Relatively **higher** weighted mean effect sizes were associated with:

- RCT designs with individual level randomisation (+0.10 SD; 95% CI: +0.05; +0.14; 41 effect sizes reported by 27 evaluations); particularly with randomisation at the pupil level (+0.11 SD; 95% CI: +0.06; +0.17; 34 effect sizes reported by 25 evaluations).
- trials with 500 pupils or fewer (weighted mean effect size = +0.16 SD; 95% CI: +0.09; +0.23; 25 effect sizes reported by 19 evaluations).
- when there was a direct match between the intervention and primary outcome (+0.09 SD; 95% CI: +0.06; +0.12; 60 effect sizes reported by 47 evaluations).
- when a commercial test was used as the primary outcome (+0.05 SD; 95% CI: +0.03; +0.08; 79 effect sizes reported by 51 evaluations), particularly if the test was from GL (+0.08 SD; 95% CI: +0.03; +0.12; 46 effect sizes reported by 33 evaluations) with GL-NGRT being the specific outcome observed with the highest weighted mean effect size (+0.09 SD; 95% CI: +0.02; +0.16; 23 effect sizes reported by 20 evaluations). Whilst we include these variables and categories in the summary, please note that the observed association was not statistically significant ($p > 0.10$).

Secondary ITT effect sizes

Figure 2 shows error bars and weighted mean effect sizes for a selection of categories within explanatory variables.

The overall weighted mean (= +0.01 SD) for all 78 secondary ITT effect sizes reported by 35 of the 82 evaluations in the review is shown at the top of Figure 2 as a solid black circle and as a dotted red vertical red line running parallel to the categorical axis. For each weighted mean, 95% confidence intervals (CI) are shown (−0.01 to +0.03 SD for all 78 effect sizes).

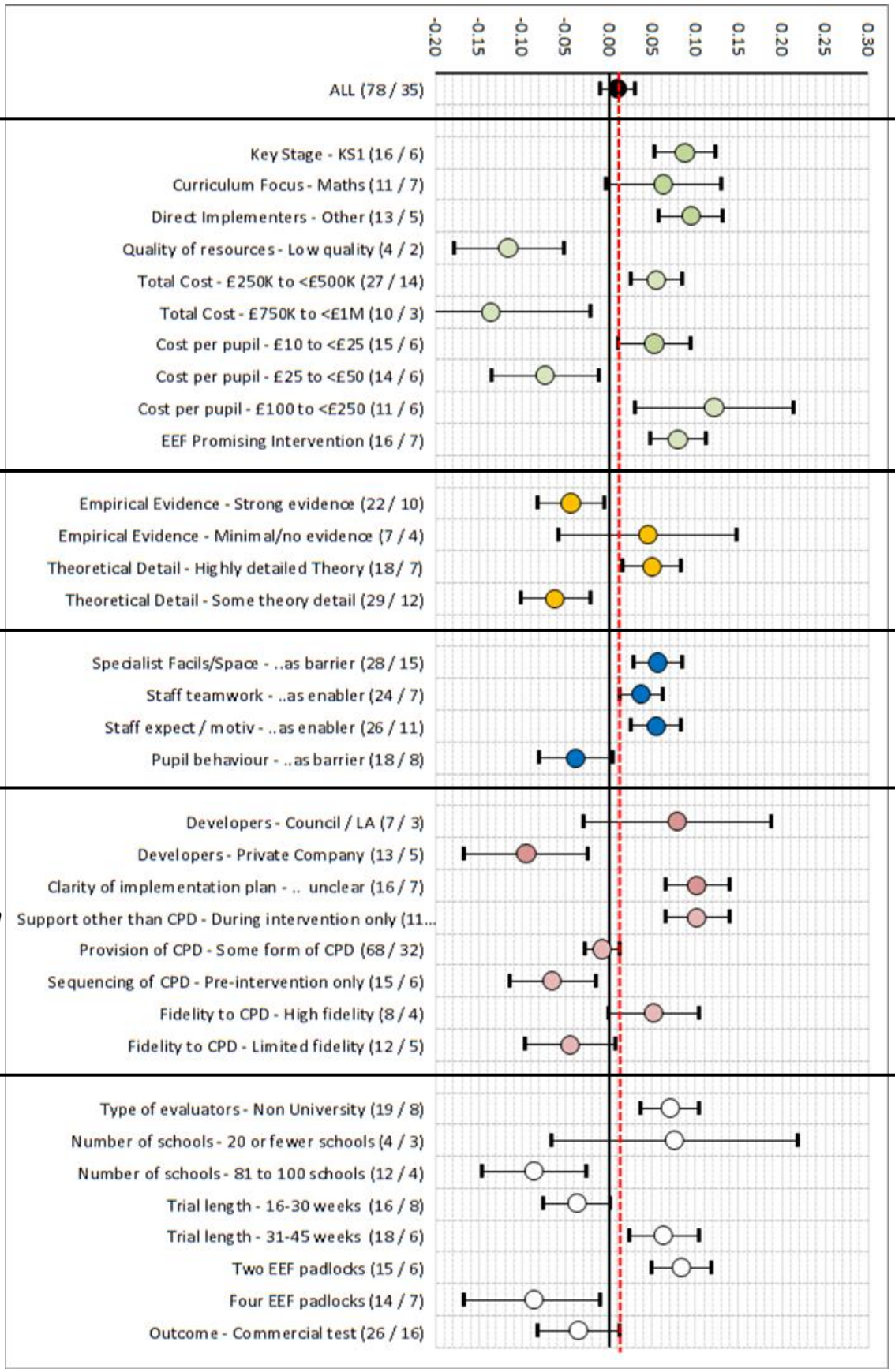
Mean weighted effect size estimates for 34 categories of 24 explanatory variables are shown using circles colour coded according to the overarching theme where an explanatory variable was placed.

The selected category is drawn from an explanatory variable found to be statistically significantly associated with primary ITT effect sizes and where the category was notably higher or lower than the overall weighted mean (of +0.01 SD). Unlike primary ITT effect sizes, there are a number of categories associated with relatively low and negative secondary ITT effect sizes. Of the 34 categories shown in Figure 2, 21 are associated with relatively high effect sizes and 13 with relatively low effect sizes.

Figure 2: Effect sizes for ITT analyses of secondary attainment outcomes

Meta-analyses summary using error bars

Weighted mean effect sizes and 95% CI from meta-analyses for selected categories of explanatory variables across the five overarching themes.



Intervention theme

Ten categories from seven variables: pupil key stage, curriculum focus, direct implementers, quality of support resources, total cost, cost per pupil and EEF promising interventions.

Relatively **higher** weighted mean secondary ITT effect sizes were associated with:

- interventions undertaken with KS1 pupils (+0.09 SD; CI: +0.05; +0.12; 16 effect sizes reported by six evaluations).
- interventions with a maths curriculum focus (+0.06 SD; CI: 0.00; +0.13; 11 effect sizes reported by seven evaluations).
- interventions when the direct implementers were someone other than teachers, TAs, other school staff, parents or externally led (+0.10 SD; CI: +0.06; +0.13; 13 effect sizes reported by five evaluations).
- interventions with a total cost of from £250k to less than £500k (+0.06 SD; CI: +0.03; +0.09; 27 effect sizes reported by 14 evaluations).
- interventions costing from £100 to less than £250 per pupil to implement (+0.12 SD; CI: +0.03; +0.21; 11 effect sizes reported by six evaluations) or costing from £10 to less than £25 per pupil to implement (+0.05 SD; CI: +0.01; +0.09; 15 effect sizes reported by six evaluations).
- interventions that were identified as 'promising' by EEF (size = +0.08 SD; CI: +0.05; +0.11; 16 effect sizes reported by seven evaluations).

Relatively **lower** weighted mean secondary ITT effect sizes were associated with:

- when it was reported that intervention support resources were perceived to be of low quality (−0.12 SD; CI: −0.18; −0.05; four effect sizes reported by two evaluations).
- interventions with a **total cost** of from £750k to less than £1 million (−0.14 SD; CI: −0.25; −0.02; 10 effect sizes reported by three evaluations).
- interventions with a **per-pupil cost** from £25 to less than £50 per pupil to implement (−0.07 SD; CI: −0.14; −0.01; 14 effect sizes reported by six evaluations).

Theory & evidence theme

Four categories from two variables: use of empirical evidence and theoretical detail.

Relatively **higher** weighted mean secondary ITT effect sizes were associated with:

- high theoretical detail for the intervention being present in the evaluation report (+0.05 SD; CI: +0.02; +0.08; 18 effect sizes reported by seven evaluations).
- minimal or no empirical evidence for the intervention being present in the evaluation report (+0.05 SD; CI: −0.06; +0.15; seven effect sizes reported by four evaluations).

Relatively **lower** weighted mean secondary ITT effect sizes were associated with:

- some limited theoretical detail for the intervention being present in the evaluation report (−0.06 SD; CI: −0.10; −0.02; 29 effect sizes reported by 12 evaluations).
- strong empirical evidence for the intervention being present in the evaluation report (−0.04 SD; CI: −0.08; −0.01; 22 effect sizes reported by 10 evaluations).

Context theme

Four categories from four variables: perceptions on specialist facilities and space, on staff teamwork, staff expectations and motivations and pupil behaviour).

Relatively **higher** weighted mean secondary ITT effect sizes were associated with:

- reported perceptions that specialist facilities/space were a barrier for the intervention (+0.06 SD; CI: +0.03; +0.09; 28 effect sizes reported by 15 evaluations).
- reported perceptions that staff teamwork was an enabler for the intervention (+0.04 SD; CI: +0.01; +0.06; 24 effect sizes reported by seven evaluations).
- reported perceptions that staff expectations and motivations were an enabler for the intervention (+0.06 SD; CI: +0.03; +0.08; 26 effect sizes reported by 11 evaluations).

Relatively **lower** weighted mean secondary ITT effect sizes were associated with:

- reported perceptions that pupil behaviour was a barrier for the intervention (−0.04 SD; CI: −0.08; +0.01; 18 effect sizes reported by eight evaluations).

Implementation & fidelity theme

Eight categories from six variables: developer type, clarity of implementation plan, implementation support, provision of CPD, sequencing of CPD and fidelity to CPD.

Relatively **higher** weighted mean secondary ITT effect sizes were associated with:

- interventions developed by councils or local authorities (+0.08 SD; CI: −0.03; +0.19; seven effect sizes reported by three evaluations).
- interventions where the implementation plan was unclear or not mentioned in the report (+0.10 SD; CI: +0.07; +0.14; 16 effect sizes reported by seven evaluations).
- when delivery partners provided support (other than CPD) during the intervention (+0.10 SD; CI: +0.07; +0.14; 11 effect sizes reported by four evaluations).
- when fidelity relating to CPD was high (+0.05 SD; CI: 0.00; +0.11; eight effect sizes reported by four evaluations).

Relatively **lower** weighted mean secondary ITT effect sizes were associated with:

- interventions developed by private companies (−0.10 SD; CI: −0.17; −0.02; 13 effect sizes reported by five evaluations).
- when the intervention involved some form of CPD (−0.01 SD; CI: −0.03; +0.01; 68 effect sizes reported by 32 evaluations).
- when CPD was delivered pre-intervention only (−0.07 SD; CI: −0.12; −0.02; 15 effect sizes reported by six evaluations).
- when fidelity relating to CPD was limited (−0.05 SD; CI: −0.10; +0.01; 12 effect sizes reported by five evaluations).

Evaluation design theme

Nine categories from six variables: level of randomisation, type of evaluator, number of schools, length of evaluation, EEF padlocks and type of outcome test.

Relatively **higher** weighted mean secondary ITT effect sizes were associated with:

- when the evaluator was **not** a university (+0.07 SD; CI: +0.04; +0.10; 19 effect sizes reported by eight evaluations).
- evaluations with 20 or fewer schools (+0.08 SD; CI: −0.07; +0.22; four effect sizes reported by three evaluations).
- evaluations that lasted between 31–45 weeks and one year (+0.06 SD; CI: +0.02; +0.10; 18 effect sizes reported by six evaluations).
- evaluations that were given two EEF padlocks (+0.08 SD; CI: +0.05; +0.12; 15 effect sizes reported by six evaluations).

Relatively **lower** weighted mean secondary ITT effect sizes were associated with:

- evaluations involving 81 to 100 schools (−0.09 SD; CI: −0.15; −0.03; 12 effect sizes reported by four evaluations).
- evaluations that lasted 16–30 weeks (−0.04 SD; CI: −0.08; 0.00; 16 effect sizes reported by eight evaluations).
- evaluations that were given four EEF padlocks (−0.09 SD; CI: −0.17; −0.01; 14 effect sizes reported by seven evaluations).
- when the secondary ITT attainment outcome was a commercial test (−0.04 SD; CI: −0.08; +0.01; 26 effect sizes reported by 16 evaluations).

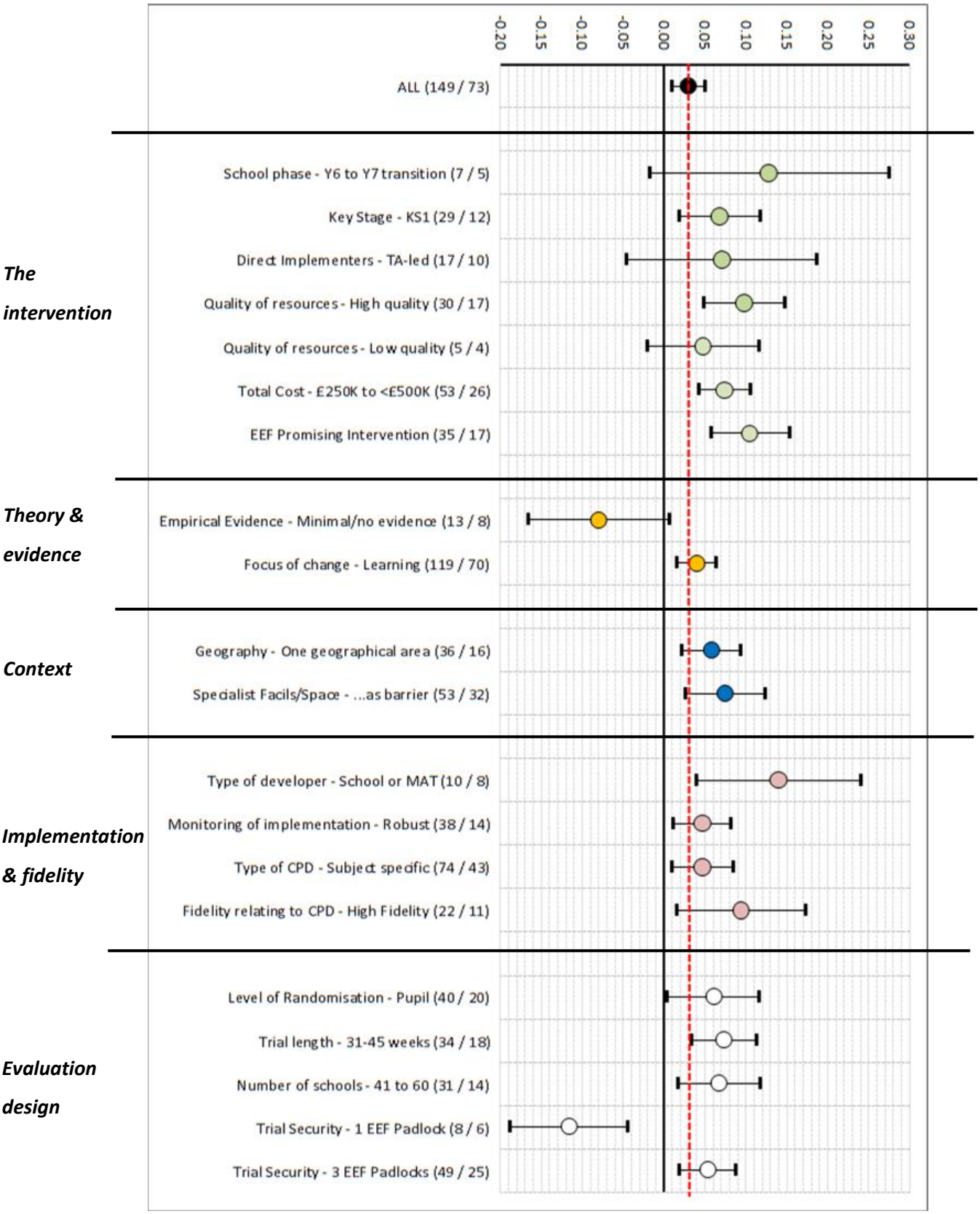
FSM effect sizes

Figure 3 shows error bars and weighted mean effect sizes for a selection of categories within explanatory variables. The overall weighted mean for all 149 FSM effect sizes (+0.03 SD) that were reported by 73 of the 82 evaluations in the review is shown at the top of Figure 3 as a solid black circle and as a dotted red vertical red line running parallel to the categorical axis. For each weighted mean, 95% confidence intervals (CI) are shown (+0.01 to +0.05 SD for all 149 effect sizes).

Figure 3: Effect sizes for FSM subsample analyses of primary or secondary attainment outcomes

Meta-analyses summary using error bars

Weighted mean effect sizes and 95% CI from meta-analyses for selected categories of explanatory variables across the five overarching themes.



Mean weighted effect size estimates for 20 selected categories of 17 explanatory variables are shown using circles colour coded according to the overarching theme where an explanatory variable was placed.

The selected category is drawn from an explanatory variable found to be statistically significantly associated with primary ITT effect sizes and where the category was notably higher or lower than the overall weighted mean (of +0.03 SD). Unlike primary ITT effect sizes, but similar to secondary ITT effect sizes, there are some categories associated with relatively low and negative FSM effect sizes. Of the 21 categories shown in Figure 3, 16 are associated with relatively high effect sizes and two with relatively low effect sizes.

Intervention theme

Seven categories from five variables: pupil key stage, direct implementers, total cost and EEF promising interventions.

Relatively **higher** weighted mean FSM effect sizes were associated with:

- interventions undertaken with KS1 pupils (+0.07 SD; CI: +0.02; +0.12; 29 effect sizes reported by 12 evaluations) or Y6 to Y7 transition (+0.13 SD; CI: −0.02; +0.28; seven effect sizes reported by five evaluations).
- TA-led interventions (+0.07 SD; CI: −0.05; +0.19; 17 effect sizes reported by 10 evaluations). Note that this association was not statistically significant ($p > 0.10$).
- when perceptions on intervention support resources was of high quality (+0.10 SD; CI: +0.05; +0.15; 30 effect sizes reported by 17 evaluations) or was of low quality (+0.05 SD; CI: −0.02; +0.12; five effect sizes reported by four evaluations).
- interventions with a **total cost** of from £250k to less than £500k (+0.07 SD; CI: +0.04; +0.11; 53 effect sizes reported by 26 evaluations).
- interventions that were identified as 'promising' by EEF (size = +0.11 SD; CI: +0.06; +0.15; 35 effect sizes reported by 17 evaluations).

Theory & evidence theme

Two categories from two variables: use of empirical evidence and focus of change process.

Relatively **higher** weighted mean FSM effect sizes were associated with:

- interventions with a learning focus of change (+0.04 SD; CI: +0.02; +0.06; 119 effect sizes reported by 70 evaluations).

Relatively **lower** weighted mean FSM effect sizes were associated with:

- minimal or no empirical evidence for the intervention being present in the evaluation report (−0.08 SD; CI: −0.17; +0.01; 13 effect sizes reported by eight evaluations).

Context theme

Two categories from two variables: geography and perceptions on specialist facilities and space.

Relatively **higher** weighted mean FSM effect sizes were associated with:

- interventions/evaluations taking place in one geographical area (+0.06 SD; 95% CI: +0.02; +0.09; 36 primary ITT effect sizes reported by 16 evaluations). Note that this association was not statistically significant ($p > 0.10$).
- reported perceptions that specialist facilities/space were a barrier for the intervention (+0.08 SD; CI: +0.03; +0.12; 53 effect sizes reported by 32 evaluations).

Implementation & fidelity theme

Four categories from four variables: developer type, monitoring of implementation, type of CPD and fidelity to CPD.

Relatively **higher** weighted mean FSM effect sizes were associated with:

- interventions developed by schools or multi academy trusts (+0.14 SD; CI: +0.04; +0.24; 10 effect sizes reported by eight evaluations).
- when evaluation reports noted that monitoring of implementation was robust (+0.05 SD; CI: +0.01; +0.08; 38 effect sizes reported by 14 evaluations).
- when CPD was subject-specific (+0.09 SD; CI: +0.05; +0.14; 18 effect sizes reported by 11 evaluations). Note that this association was not statistically significant ($p > 0.10$).
- when fidelity relating to CPD was high (+0.09 SD; CI: +0.05; +0.14; 18 effect sizes reported by 11 evaluations). Note that this association was not statistically significant ($p > 0.10$).

Evaluation design theme

Five categories from four variables: level of randomisation, number of schools, length of evaluation and EEF padlocks.

Relatively **higher** weighted mean secondary ITT effect sizes were associated with:

- pupil-level randomisation (+0.06 SD; CI: 0.00; +0.12; 40 effect sizes reported by 20 evaluations). Note that this association was not statistically significant ($p > 0.10$).
- evaluations with 41–60 schools (+0.07 SD; CI: +0.02; +0.12; 31 effect sizes reported by 14 evaluations).
- evaluations that lasted between 31–45 weeks and one year (+0.07 SD; CI: +0.03; +0.11; 34 effect sizes reported by 18 evaluations).
- evaluations that were given three EEF padlocks (+0.05 SD; CI: +0.02; +0.09; 49 effect sizes reported by 25 evaluations).

Relatively **lower** weighted mean secondary ITT effect sizes were associated with:

- evaluations that were given one EEF padlock, indicating low security (−0.12 SD; CI: −0.19; −0.05; eight effect sizes reported by six evaluations).

Cost effectiveness

Cost effectiveness was defined as the cost per pupil for an effect size of +0.10 standard deviations **given** evidence of a positive impact. 40 of the 82 (49%) evaluations in the review were identified has providing some evidence of a positive impact and included into the cost effectiveness analyses. The theoretical framework developed for the meta-analyses of primary ITT effect sizes was adapted for these descriptive evaluation-level analyses of cost effectiveness. Key findings are summarised below under the five overarching themes.

The intervention

- Successful interventions in secondary schools tend to be less cost effective (median = £69; mean = £109) than in primary schools (median = £43; mean = £130).
- There was little evidence of difference in terms of cost effectiveness for successful English, maths or cross-curriculum interventions. Successful teacher-led interventions were the most cost effective and externally-led interventions the least.
- Successful interventions with relatively high intensity (over two hours delivery per week) were associated with lower cost effectiveness (median = £183; mean = £285) compared to less intense interventions (median ≤ £69; mean ≤ £162).
- Successful interventions that cost between £250k and £500k were associated with greater cost effectiveness (median = £34; mean = £96) compared with interventions with higher or lower cost (median > £60; mean > £160).
- The only statistically significant association found with EEF intervention themes was the lower cost effectiveness of the 'organising your school' theme, which is likely to relate to the higher cost of implementing whole-school interventions.

Theory & evidence

- On average, successful interventions that drew on strong empirical evidence were more cost effective (median = £43; mean = £74) than evaluations presenting limited evidence (median = £48; mean = £126).

Context

- The cost effectiveness of successful interventions improved over the five years from 2014 (median = £100; mean = £238) to 2018 (median = £33; mean = £37) although this is not statistically significant. There was no clear or statistically significant association of cost effectiveness with geographical context, the only other contextual variable examined.

Implementation & fidelity

- Of the successful interventions, those having 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), but the association was not statistically significant.
- No evidence was found for an association between cost effectiveness and fidelity related to CPD, the intended fidelity (by the direct implementer), or the actual fidelity of intervention implementation.

Evaluation design

- Successful interventions evaluated using RCTs with pupil-level randomisation 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.
- Whilst the association is not statistically significant, successful shorter interventions were less cost effective than longer programmes, while interventions that engaged fewer pupils were less cost effective than those engaging larger numbers of pupils.
- Among the interventions included, commercial tests are associated with successful interventions that had a less cost effective impact (median = £89; mean = £204) compared with trials that used official/NPD data (median = £15; mean = £42).

Pupil-level attrition

Pupil attrition was also an evaluation/trial level variable reported for primary outcomes and obtained for 79 of the 82 evaluations in the review. The theoretical framework developed for the meta-analyses of primary ITT effect sizes was also adapted for these descriptive evaluation-level analyses of attrition. However, some themes did not align well with this attrition outcome. For example, whilst some themes focused on aspects of the intervention and implementation, the outcome had a broader focus of attrition across both intervention and control conditions. Future reviews may want to examine intervention and control group attrition more directly. Key findings are summarised below under the five overarching themes.

The intervention

- Pupil-level attrition rates were higher for interventions located at the Y6–Y7 primary to secondary transition stage compared to interventions located solely in secondary or primary schools. 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, these findings may well be related to the form of outcome measure used. The vast majority of KS3 interventions used a commercial test, as did the majority of KS2 interventions and Y6–Y7 transition interventions, but none of KS4 interventions used commercial tests. When comparisons are possible, the use of commercial tests results in higher attrition rates compared with the use of NPD outcomes.
- Whilst the overall median attrition rate for interventions that focused on maths (7.4%) was notably lower than for English (16.2%) or cross-curriculum (16.0%) interventions, this seems to relate largely to the type of primary outcome used (i.e., commercial or NPD).
- Interestingly, no evidence of an association between the intensity of an intervention and attrition was observed.

Theory & evidence

- No associations were observed between pupil attrition and variables in the theory & evidence theme.

Context

- No associations were observed between pupil attrition and variables in the context theme.

Implementation & fidelity

- No evidence was found for an association between attrition and explanatory variables included under the implementation & fidelity theme. This may relate to the issue of alignment: the attrition outcome has a broader focus (intervention and control samples), whilst these explanatory variables are specific to the intervention sample.

Evaluation design

- The number of schools in a trial has a complex but statistically significant association with pupil-level attrition. In general, there is a weak negative correlation between attrition and the number of schools in a trial.
- The association between testing burden and pupil-level attrition was statistically significant, with much lower attrition found for evaluations with no external tests.
- Unsurprisingly, trials that used commercial tests as primary outcome(s) had higher attrition than trials that used official/NPD data.

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, which are summarised below.

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 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 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

This 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 in 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 outcome analyses

Secondary outcomes for the review effect sizes for ITT analyses of trial secondary attainment outcomes, effect sizes for FSM subsample analyses of trial 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-level 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.

References

Demack, S., Maxwell, B., Coldwell, M., Stevens, A., Wolstenholme, C., Reaney-Wood, S., Stiell, B. and Lortie-Forges, H. (2021) *Review of EEF Projects*. Full report available at https://educationendowmentfoundation.org.uk/public/files/Publications/Review_of_EEF_Projects.pdf

Maxwell, B., Stiell, B., Stevens, A. Demack, S., Coldwell, M., Wolstenholme, C., Reaney-Wood, S. and 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

This document is available for download at https://educationendowmentfoundation.org.uk

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

https://educationendowmentfoundation.org.uk
@EducEndowFoundn
Facebook.com/EducEndowFoundn