BERJ British Educational Research Journal

BERA BRITISH EDUCATIONAL RESEARCH ASSOCIATION

Check for updates

*British Educational Research Journal*
*Vol. 47, No. 6, December 2021, pp. 1675–1699*

DOI: 10.1002/berj.3749

# Individual participant data meta-analysis of the impact of educational interventions on pupils eligible for Free School Meals

Bilal Ashraf*, Akansha Singh, Germaine Uwimpuhwe, Steven Higgins and Adetayo Kasim
*Durham University, UK*

Meta-analysis is the synthesis of findings from research projects, which enables an estimate of the average or pooled effect across various studies. This study presents findings from the intention to treat analysis for a series of educational evaluations in England using a two-stage meta-analysis with standardised outcome data and individual participant data meta-analyses. The research estimates the overall impact of educational trials on pupils eligible for Free School Meals (FSM) and the attainment gap in literacy and mathematics performance between FSM and non-FSM pupils based on analysis of 88 trials and data from over half a million pupils. For the meta-analyses, frequentist and Bayesian multilevel models were used to estimate the individual and pooled effect size across categories of explanatory variables such as age groups (key stages in England) and aspects of the type of interventions (one-to-one, small group, whole class). Results indicated that the overall impact of interventions on the literacy outcomes of FSM pupils was positive, with a pooled effect size of 0.06 (0.03, 0.08). However, for mathematics, no overall effect on FSM pupils was observed. Analysis of the attainment gap indicated that literacy outcomes for FSM pupils were improved by interventions marginally more than for non-FSM pupils (pooled attainment gap 0.01 (−0.01, 0.04)). The risk of bias assessment showed that estimates were consistent across different methodological approaches. Overall, evidence from this study can be used to identify, test and scale educational interventions in schools to improve educational outcomes for disadvantaged pupils.

**Keywords:** educational attainment gap; Free School Meals; individual participant data; meta-analysis

## Introduction

Educational attainment has become one of the clearest early indicators of life outcomes including employment, income and social status, and is a strong predictor of attitudes and wellbeing (Manstead, 2014). Marmot (2010) argued that there are particularly large gaps between extremes of the social hierarchy in the UK, with people from the highest social or economic background living longer and with a longer period of their life free from health issues. The impact of low levels of achievement in education is not restricted to adulthood, it is also a greater issue with school-aged

*Corresponding author. Durham Research Methods Centre, Durham University, Durham DH1 3LE, UK. Email: bilal.h.ashraf@durham.ac.uk

children. It is well known that children growing up in poorer families emerge from school with substantially lower levels of educational attainment (Chowdry *et al.*, 2010). Since 2011, 60% of children in absolute and relative poverty were eligible for Free School Meals (FSM) (DWP, 2013), which became mandatory for all pupils in Reception and Years 1 and 2 in England in 2014 (DfE, 2014). Pupils eligible for FSM are reported to make less progress on average compared to their peers (Humphrey *et al.*, 2013). The gap between disadvantaged pupils and their peers in England is equivalent to one whole General Certificate of Secondary Education (GCSE) grade for mathematics and 0.75 grade in reading. This gap is significantly higher than several other high-income countries in Europe and Asia (Jerrim *et al.*, 2018). The gap between disadvantaged pupils and their peers is evident even when children begin school at age 5, and increases at every stage of education afterwards (Education Endowment Foundation, 2019). In Scotland, children living in the most deprived areas are '6 to 13 months behind their peers in problem-solving at age 5; 11 to 18 months behind their peers in expressive vocabulary at age 5; and around two years of schooling behind their peers at age 15' (Scottish Government, 2014). By the time that children leave primary school, those in receipt of FSM are estimated to be significantly behind their more affluent peers (Spencer, 2015). This gap clearly indicates the need to focus on social deprivation to ameliorate the impact of poverty, and here schools have a pivotal role to play. High-quality education and better teaching methods can be important in reducing this attainment gap (Jerrim *et al.*, 2018). Improving the educational achievements of pupils eligible for FSM also has the potential to break the cycle of poverty, reduce health inequality, improve lifestyle choices and improve mental health (Hobbs & Vignoles, 2010).

The Education Endowment Foundation (EEF) is an independent charity dedicated to breaking the link between family income and educational achievement. More than 150 trials have been commissioned by the EEF to improve the academic attainment of children and reduce the attainment gap among deprived pupils as compared to their counterparts (Education Endowment Foundation, 2019). Subgroup analyses of pupils on FSM are usually reported in each trial report, but there is a need to synthesize evidence on the impact of EEF-funded interventions on FSM pupils across trials. The analysis of FSM pupils reported for each trial is a useful complement to the main findings from individual trials. However, it offers limited insights into how EEF-funded interventions as a whole affect FSM pupils. Are the interventions reducing attainment gaps between FSM pupils and their peers? And what types of intervention are likely to be more beneficial to FSM compared with their peers? These are some of the questions that need answers to improve the design or implementation of future interventions aiming to reduce the attainment gap (Schochet *et al.*, 2014).

The current COVID-19 closures of schools are predicted to reverse the progress made to close the attainment gap in the last decade (Coe *et al.*, 2020). Therefore, it is timely to highlight the characteristics of the most promising interventions that were effective in reducing the attainment gaps between recipients of FSM and their peers. This study provides a robust and independent assessment of how targeted interventions benefit FSM pupils and how they impacted on the attainment gap by synthesizing evidence from existing trials using individual participant data (IPD) meta-analysis methods. The traditional approach in meta-analysis relies on extracting effect sizes

from each trial (Burke *et al.*, 2017; Kontopantelis, 2018), but the use of summary statistics often suffers from loss of information and lack of consistency in the methods used to calculate individual effect sizes (Debray *et al.*, 2015). IPD meta-analysis is a more flexible approach to capture variability within and between trials by using data from the individual pupils who participated. This can also improve standardisation of outcomes; reduce publication, reporting and ecological biases; allow detailed checks of analysis assumptions and consideration of covariates and treatment–covariate interactions which are often lacking in traditional meta-analysis methods (Debray *et al.*, 2015).

This study meta-analysed evidence from randomised controlled trials (RCTs) commissioned by the EEF and reported between 2011 and 2019 and contained in their data archive, to assess the impact of EEF-funded interventions on FSM pupils. We defined FSM pupils as pupils who were ever eligible for FSM in the last 6 years in primary school. (This definition of educational disadvantage is not without its problems: see, for example, Gorard, 2012 and Taylor, 2018 using data from the millennium cohort study in Wales). We also aim to identify broad types of intervention with common pedagogical features (e.g. one-to-one tuition or small-group versus whole-class teaching approaches), which are more likely to improve the educational attainment of FSM pupils to support educational decision-making. We acknowledge that this is a somewhat simplistic characterisation of a range of very different and often complex interventions, but were looking to explore the value of the method and to see if any more general messages could be identified from the analysis. This research for the first time provided comparable individual and global pooled effect sizes for FSM pupils and the estimated attainment gap in their educational performances in literacy and mathematics. This article therefore seeks to add to what is known in this area, both in terms of the differences between disadvantaged pupils and their peers, and in terms of identifying successful interventions to address this challenge (e.g. Dietrichson *et al.*, 2017). Current approaches focus on identifying the gap and the nature of disadvantage. The approach described in this article is similar to using meta-synthesis (Higgins, 2016), where the results from meta-analysis are compared (e.g. Dietrichson *et al.*, 2017). However, instead of using summary statistics, individual attainment data from pupils was used across a similar set of outcomes [from England's National Pupil Database (NPD), which records all pupils' results from national tests and examinations].

## Materials and methods

### *Data and study design*

In this study, 82 EEF projects were available for meta-analysis with 4,396 schools and 525,534 pupils. Most of the EEF trials were either cluster randomised trials (CRT) or multi-site randomised trials (MST). In MST, randomisation was within a school, such that pupils in each school are involved in both the intervention and control groups (Xiao *et al.*, 2016). In CRT clusters, such as schools, classes or year groups, pupils were randomly assigned to either intervention or control groups. It is possible for both designs to be combined in a single trial, such as cluster

randomisation of classes within schools. Most of the MST and CRT trials in the EEF archive were two-armed trials, with an intervention and a business-as-usual control. A few trials with more than one treatment (more than two arms) were treated as separate trials for each treatment. Therefore, the total number of trials in each analysis reflects the number of trials including two or three treatment trials as separate comparisons. There were 76 trials in the data archive with one treatment trial, and six trials had more than one treatment giving, a total of 88 trials. All trials of the available trials in the EEF archive which met the criteria for IPD meta-analysis were included in this analysis. Trials with no literacy or mathematics outcome, or with IPD model computational inconsistencies, were mainly excluded from the analysis.

The outcomes in all the trials were literacy and mathematics, with attainment data either obtained from the NPD or collected directly by the evaluators' preferred measures of literacy and mathematics. Although this provided a consistent dataset, the differences in assessment across the complex domains of literacy and mathematics need to be borne in mind. It is also important to note that in the context of evidence synthesis, the false positives are implicitly controlled, since the inference is based on pooled evidence across the trials. Hence, adjustment for multiple testing is redundant and not undertaken (Brookes *et al.*, 2001).

### *Variables of interest*

Two major groups of variables (the ages of pupils or pupil 'key stages' in England) and the type of intervention were considered for the meta-analysis. The effect of EEF-funded interventions was assessed across the pupil key stages that are used to organise curriculum and assessment in England [KS1 (5–7 years old), KS2 (8–11 years old), KS3 (12–14 years old) and KS4 (15–16 years old)] separately for each key stage. The outcomes were also meta-analysed by type of intervention to determine which group of interventions was more beneficial for FSM pupils (this corresponds to pupils eligible for FSM in the previous 6 years or 'ever6 FSM', which is the EEF's preferred measure). Types of intervention were classified as one-to-one, small group, whole class or whole school. This classification was adopted from the EEF Evidence Database project. This was largely a pragmatic decision in identifying similar pedagogical features of the interventions which could be used to classify them. The interventions and approaches vary considerably in terms of their rationales, content and teaching and learning approaches, and the group size was a consistent variable which could be examined in all of the interventions.

### *Two-stage meta-analysis method*

A traditional meta-analysis approach typically aggregates the effect sizes from different studies by weighting them proportionally to study-specific variability and the variability between trials. The major drawback of this approach is the loss of information, which is typical of any summarised data (Debray *et al.*, 2015). Another limitation is that sometimes effect sizes are calculated differently using different statistical approaches and scaling factors. For example, the use of conditional or unconditional variance may result in different estimates of the magnitude of the effect, as well as different estimates

for the standard error (Singh *et al.*, 2021). Retaining the same framework for traditional meta-analysis methods, we proposed to re-estimate an effect size for all trials using the same, consistent methods and to compare this with the IPD approach. Although this approach does not correct for the loss of information, it reduces the variability between effect sizes attributable to the analytical approach (Xiao *et al.*, 2016). Our proposed two-stage meta-analysis involves two steps.

*Stage 1: Calculating effect size per trial.* Individual trials were analysed independently using the multilevel model (MLM) specified in Equation (1). Let $Y_{ijk}$ be the outcome data for pupil $i$ from school $j$ in trial $k$, then the two-level model for each trial is formulated as

$$Y_{ijk} = \beta_{0k} + \beta_{1k} Pret_{ijk} + \beta_{2k} T_{ijk} + b_{jk} + \varepsilon_{ijk} \quad (1)$$

where $\beta_{0k}$ is the overall intercept, $\beta_{1k}$ is the gradient between post- and pre-test scores, $\beta_{2k}$ is the adjusted difference between the intervention and control groups based on the indicator for intervention $T_{ijk}$, defined as $T_{ijk} = 1$ for intervention (treatment) group and $T_{ijk} = 0$ for comparison group for a two-arm trial. $b_{jk} \sim N(0, \omega_k{*}\omega_k)$ captured between-school variability and $\varepsilon_{ijk} \sim N(0, \sigma_k{*}\sigma_k)$. denotes residual variance. Furthermore, the effect size and its confidence intervals for each trial were calculated as

$$ES_k = \frac{\beta_{2k}}{\sqrt{\omega_k^2 + \sigma_k^2}},\ CI\_lower_k = \frac{Lower(\beta_{2k})}{\sqrt{\omega_k^2 + \sigma_k^2}},\ CI\_upper_k = \frac{Upper(\beta_{2k})}{\sqrt{\omega_k^2 + \sigma_k^2}}$$

where $Lower(\beta_2)$ and $Upper(\beta_2)$ are 95% confidence intervals for the adjusted difference between the intervention and comparison groups ($\beta_2$). Also note that the post-test scores for each trial were standardised pre-analysis by subtracting the mean score and then divided by the standard deviation of scores in the trial, $ES_k = \beta_{2k}$, $CI_{lower_k} = Lower(\beta_2)$ and $CI_{upper_k} = Upper(\beta_2)$. The lme4 package in R was used to fit the multilevelodel and to estimate all the parameters.

*Stage 2: Weighted average.* The standard error of the effect size from trial $k$ ($SE_k$) was calculated from the confidence interval ($CI\_upper_k$, $CI\_lower_k$) of $ES_k$, as shown in Equation (2) (Cochrane, 2019):

$$SE_k = \frac{CI\_upper_k - CI\_lower_k}{3.92} \quad (2)$$

Given that all EEF-funded interventions were not implemented in similar settings, both fixed-effect and random-effect meta-analyses were used to summarise the impact of EEF-funded interventions. The random-effects approach assumes that there is not one true effect size but a distribution of effects due to differing interventions. In this case, between-trials heterogeneity ($\tau^2$) has to be taken into account (Borenstein *et al.*, 2011), whilst the trials are assumed to be homogenous in fixed-effect meta-analysis.

Based on the estimated effect size ($ES_k$) in stage 1 and $\tau^2$, the weighted average effect size or pooled effect size was calculated as

$$Pooled\,ES = \frac{\sum_{k=1}^{K} W_k ES_k}{\sum_{k=1}^{K} W_k} \tag{3}$$

where $W_k = (SE_k^2 + \tau^2)^{-1}$ is the weight for the individual trial based on variability for each effect size and the heterogeneity between trials (Hedges & Olkin, 1985; Pigott, 2012). Specifically, in education trials, $SE_k$ also accounted for between-school variability when a multilevel model was used. Although this approach provides the global impact of the interventions, it suffers from loss due to the two-stage approach for obtaining the pooled effect size. This type of bias is called the ecological fallacy (Reade *et al.*, 2008), as it does not account for heterogeneity at the individual level (Debray *et al.*, 2015).

*IPD meta-analysis*

An IPD meta-analysis method offers a more flexible and pragmatic way to synthesise evidence from existing interventions (Burke *et al.*, 2017; Kontopantelis, 2018). It is a more powerful approach than traditional meta-analysis or a two-stage approach because of its ability to pool information across multiple trials, while also accounting for the different sources of variation (Debray *et al.*, 2015; Smith *et al.*, 2016). IPD meta-analysis allows important baseline data and trial-specific characteristics to be accounted for in the same model. IPD is more attractive because it fully exploits the available data of individual participants without having to perform additional transition steps (Fanshawe & Perera, 2019).

IPD meta-analysis can be considered an extension of a multilevel model, where two-level models are extended to incorporate a third level to capture heterogeneity between trials. Within a Bayesian framework (Burke *et al.*, 2017), pupils (level 1) were nested within schools (level 2) and schools were nested within trials (level 3). Let $Y_{ijk}$ be the outcome data for pupil $i$ from school $j$ who participated in trial $k$ as previously defined, a full IPD meta-analysis model can then be formulated as

$$Y_{ijk} = (b_{0k} + \varphi_0) + (b_{1k} + \varphi_1) Pret_{ijk} + (b_{2k} + \varphi_2) T_{ijk} + S_{jk} + \varepsilon_{ijk} \tag{4}$$

where $\varphi_0$, $\varphi_1$ and $\varphi_2$ were the pooled intercept, gradient between pre-test and post-test, and treatment effect across trials. Whilst $b_{0k} \sim N(0, \tau_k{*}\tau_k)$, $b_{1k} \sim N(0, \vartheta_k{*}\vartheta_k)$ and $b_{2k} \sim N(0, \delta_k{*}\delta_k)$ were the trial-specific deviations from the pooled intercept, gradient between pre-test and post-test, and the treatment effects. The additional sources of variation within each trial were captured by $S_{jk} \sim N(0, \omega_{sk}{*}\omega_{sk})$ and $\varepsilon_{ijk} \sim N(0, \sigma_k{*}\sigma_k)$, where $\omega_{sk}$ denoted heterogeneity between schools in trial $k$ and $\sigma_k$ captured between-pupil variability in trial $k$.

This model formulation highlights the first challenge with an IPD meta-analysis of evidence from educational trials. The pooled effect of the intervention $(\varphi_2)$ was only meaningful if the outcomes in each trial were on the same scale, which is often not the case in educational trials. A further challenge is that there was no single

measure of heterogeneity between schools ($\sigma_{sk}$) and within pupils ($\sigma_k$) per trial, except if one is willing to make unrealistic assumptions that $\omega_{sk}^2 = \omega_k^2$ and $\sigma_k^2 = \sigma^2$. Outcome measures in education trials are generally very variable between trials, even when measuring the same outcome, due to the fact that each education trial is typically based on a convenience sample of schools willing to take part in the trial. An even more complicated issue is that the outcome in each trial can be from a national test at any of the key stages, or from a bespoke test. Additional sources of variability typical in education trials are the nature of the pre-test scores and how strongly they are correlated with the outcome data. A further challenge is that one cannot safely assume that effect sizes from each trial are from a single distribution, or even driven by common underlying factors. This is partly the reason that IPD meta-analysis is not a common approach in education trials, despite the methodological advancements in health and clinical trials. Effect size, as a ratio measure, is a controversial metric, especially in education (Simpson, 2018). The distributions vary by age and subject (Bloom *et al.*, 2008) and may relate systematically to different features of interventions, such as sample size (however, they are the best measure we currently have to investigate effects across projects and to synthesise otherwise disparate findings; Higgins, 2018).

### *Simplified IPD meta-analysis model*

The IPD meta-analysis model cannot therefore be directly applied to educational trials without further considerations. We propose to first eliminate heterogeneity between trials by scaling the post-test and pre-test outcome data to a unit variance of one per trial. This scaling approach is statistically not the ideal approach, but it offers the best trade-off in balancing between the challenges of the model and ensuring meaningful results.

The other issue that needs to be addressed is relaxing the assumption that the effects of the interventions are from a single distribution with common mean ($\varphi_2$), because the trial-specific impact ($b_{2k} + \varphi_2$) will shrink towards the pooled effects (Duchateau *et al.*, 1998; Lesaffre & Lawson, 2012; Kruschke, 2015). Depending on the shrinkage factor, these estimates may differ from the corresponding estimates from a two-stage meta-analysis approach and the individual effect size in the evaluation report of the different trials. The amount of shrinkage will depend on the extent of the variability [the between-trial variability ($\tau_k^2$), the within-trial variability ($\omega_{sk}^2 + \sigma_k^2$) and the number of schools and pupils in each trial; Laird, 2004]. Although the scaling of the post-test and pre-test outcome data removed the between-trial variability, within-trial variability may remain substantially different between the trials. Due to this within-trial variability, a less heterogeneous trial will be disadvantaged, because the lower the between-trial variance, the greater the shrinkage effect (Duchateau *et al.*, 1998).

To retain the power of an IPD meta-analysis and to ensure the meaning of the results in the context of educational interventions, we proposed a simplified IPD meta-analysis model as follows:

$$Y_{ijk}^s = \beta_{0k} + \beta_{1k} Pret_{ijk}^s + \beta_{2k} T_{ijk} + S_{jk} + \varepsilon_{ijk} \quad (5)$$

where $Y^{s}_{ijk}$ and $Pret^{s}_{ijk}$ are standardised post-test and pre-test scores. $\beta_{0k}$ is the fixed intercept, $\beta_{1k}$ is the fixed gradient between the standardised post-test and pre-test scores and $\beta_{2k}$ is the average effect of the intervention in trial $k$. However, $S_{jk} \sim N(0, \omega_{sk}{}^{*}\omega_{sk})$ and $\varepsilon_{ijk} \sim N(0, \sigma_{k}{}^{*}\sigma_{k})$ remain as random effects in the model. To obtain the pooled effect size, we use

$$\varphi_2 = \frac{\sum_{k=1}^{K} W_k \beta_{2k}}{\sum_{k=1}^{K} W_k}$$

where $W_k = (\omega_{sk}^2 + \sigma_k^2)^{-1}$ captures within-trial variability given that between-trial variability is pre-scaled to one. This simplified IPD model is expected to produce results consistent with the two-stage meta-analysis approach and the effect size from the evaluation report for each trial, where a multilevel model was used for effect size using conditional variance. Two-stage and IPD meta-analysis methods may produce different results when some studies have unbalanced sample sizes between the treatment and control groups (Burke *et al.*, 2017).

The proposed IPD meta-analysis method for educational trials was implemented within a Bayesian framework assuming vague normal priors for all fixed effects and vague inverse-gamma priors for all variance parameters. The use of non-informative or vague priors for a Bayesian evaluation of educational trials ensures that the conclusion is determined by the data instead of the researchers' previous knowledge (Uwimpuhwe *et al.*, 2020). The credible intervals for the pooled effect size and the trial-specific effect size were obtained as 2.5% and 97.5% quantiles from their posterior distributions. To ensure convergence of the parameters, we used three chains with 200,000 Markov chain Monte Carlo (MCMC) iterations. Further, the number of iterations necessary to obtain convergence depends on the analysis at hand; the more you increase this number, the greater the chance of sampling from the target distribution (Raftery & Lewis, 1995). The first half of each chain was discarded as the 'burn-in' part. The burn-in part is the number of iterations ignored since the beginning of an MCMC run, so that the posterior distribution can be independent of the initial values (Uwimpuhwe *et al.*, 2020). All results were reported after checking for convergence using Rhat and trace plots. The separate meta-analysis models were fitted for literacy and mathematics outcomes using all available data. Further meta-analyses were performed using different factors such as key stage and intervention types. We used the R2jags R software package in the Linux environment (high-performance computing) for the Bayesian IPD meta-analysis.

### *Attainment gaps*

The meta-analysis of effect sizes for only FSM pupils does not provide insight into whether EEF-funded interventions have reduced attainment gaps between them and their peers. It is possible that an intervention will have the same effect on FSM and non-FSM pupils and in such a situation, there may be a positive effect for FSM pupils but with no change in the attainment gap for the specific trial. Another possibility is that an intervention may have no or a lesser effect on FSM pupils, but a positive

effect on non-FSM pupils. In such a situation, the intervention is likely to widen the attainment gap. Lastly, an intervention may have a positive effect on FSM pupils and no effect or a lesser effect on non-FSM pupils. Such an intervention is likely to reduce the attainment gap, as more FSM pupils have improved their educational outcomes. Although this illustration is for an individual trial, it is also a possibility to consider for a pooled estimate of the impact of these interventions. To estimate the attainment gap between FSM and non-FSM pupils, the model specified in Equation (5) was extended with an interaction term between FSM and intervention groups (Kontopantelis, 2018) and using data for all pupils as follows:

$$Y^s_{ijk} = \beta_{0k} + \beta_{1k} Pret^s_{ijk} + \beta_{2k} T_{ijk} + \gamma_{1k} FSM_{ijk} + \gamma_{2k} T_{ijk} * FSM_{ijk} + S_{jk} + \varepsilon_{ijk} \qquad (6)$$

Parameter $\gamma_{2k}$ is the attainment gap (i.e. the difference in average effect of the interventions between FSM pupils and their peers in trial $k$ and the impact of the intervention on FSM pupils in trial $k$), $\beta_{2k}$ is the impact of the intervention on non-FSM pupils in trial $k$, and the impact of the intervention on FSM pupils in trial $k$ is $\beta_{2k} + \gamma_{2k}$. To estimate the pooled effect of the intervention on attainment gap, the model is further specified as

$$\text{Attainment gap } (\eta) = \frac{\sum_{k=1}^{K} V_k \gamma_{2k}}{\sum_{k=1}^{K} V_k}$$

where $V_k = (\omega^2_{sk} + \sigma^2_k)^{-1}$. The model was fitted within a Bayesian framework using the same sets of priors as previously defined. The attainment gap was also estimated using the two-stage meta-analytic approach by simply adding an interaction between treatment and FSM variables in the model defined in Equation (2), estimating the attainment gap from each trial and pooling the attainment gap estimates together using the Cochrane method (Cochrane, 2019).

### *Heterogeneity*

We measured heterogeneity using the statistical test usually applied in meta-analyses for determining whether there is true heterogeneity among the studies' effects, adopting the $Q$-test proposed by Cochran (1954) and also described in Bowden *et al.*, (2011). The $Q$-statistics used in this study are defined as

$$Q = \begin{cases} \sum_{k=1}^{K} W_k (\varphi_2 - \beta_{2k})^2 \text{ for FSM subgroup} \\ \sum_{k=1}^{K} V_k (\eta - \gamma_{2k})^2 \text{ for attainment gap} \end{cases}$$

Further, the $I^2$ index proposed by Higgins & Thompson (2002) was also estimated. This index quantifies the extent of heterogeneity from a collection of effect sizes by comparing the $Q$ value to its expected value assuming homogeneity, that is, to its degrees of freedom (df = $k - 1$).

## Results

Tables 1 and 2 provide a summary of the trial outcomes and the number of pupils, schools and FSM (non-FSM) eligible pupils by key stage and type of intervention. For literacy, among the 81 trials, 13 trials assessed KS1, 33 trials assessed KS2, 29 trials assessed KS3 and 6 trials assessed KS4. Similarly, 9, 24, 9 and 6 trials assessed mathematics in KS1–4, respectively. Furthermore, for literacy, 24, 17, 30 and 10 trials assessed one-to-one, small-group, whole-class and whole-school interventions. There were also 10, 7, 23 and 8 one-to-one, small-group, whole-class and whole-school interventions for mathematics performance, respectively.

Overall, there were 211,920 instances of FSM pupils from 4,000 instances of schools with literacy outcomes and 217,728 instances of FSM pupils from 3,178 instances of schools with mathematics outcomes. We had reported on instances of pupils and schools because there was no indicator to uniquely identify the schools and pupils across the trials.

### *Heterogeneity between trials*

An important consideration in the meta-analysis of existing evidence is how comparable are the measures of treatment or intervention effects. Variability between trials due to different participating populations, different outcomes with respect to scale or underlying constructs, differences in methods of how the effect size were calculated and differences in quality of the trials play a significant role in estimating pooled effects across trials (Brookes *et al.*, 2001). There is a consensus that variable measures of intervention effects are likely to produce unreliable evidence of the average effects of the interventions across trials (Thomas *et al.*, 2014), although some of the variability between trials can be accounted for in a random effects meta-analysis.

The level of variability between trials is particularly important in IPD meta-analysis because the data will be analysed on the original scales, which are likely to be different between trials. An important example in EEF trials is with respect to the different key stage results. It is also well known that schools and pupils participating in educational trials are rarely representative of the wider population of

Table 1. Overview of literacy trials by outcome type, study design and types of intervention

| | | No. trials | No. schools | No. pupils | No. FSM pupils | No. non-FSM pupils |
|---|---|---|---|---|---|---|
| | Overall | 81 | 4,000 | 302,138 | 90,218 | 211,920 |
| Key stage outcome | KS1 | 13 | 529 | 19,905 | 4,444 | 15,461 |
| | KS2 | 33 | 2,265 | 102,835 | 34,085 | 68,750 |
| | KS3 | 29 | 552 | 39,297 | 10,108 | 29,189 |
| | KS4 | 6 | 654 | 140,101 | 41,581 | 98,520 |
| Type of intervention | One-to-one | 24 | 1,358 | 97,368 | 28,194 | 69,174 |
| | Small group | 17 | 503 | 22,451 | 6,914 | 15,537 |
| | Whole class | 30 | 1,339 | 83,550 | 29,774 | 53,776 |
| | Whole school | 10 | 800 | 98,769 | 25,336 | 73,433 |

Table 2. Overview of mathematics trials by outcome type, study design and type of intervention

| | | No. trials | No. schools | No. pupils | No. FSM pupils | No. non-FSM pupils |
|---|---|---|---|---|---|---|
| | Overall | 48 | 3,178 | 306,975 | 89,247 | 217,728 |
| Key stage Outcome | KS1 | 9 | 639 | 18,718 | 4,394 | 14,324 |
| | KS2 | 24 | 1,577 | 79,671 | 25,946 | 53,725 |
| | KS3 | 9 | 269 | 30,434 | 6,667 | 23,767 |
| | KS4 | 6 | 693 | 178,152 | 52,240 | 125,912 |
| Type of intervention | One-to-one | 10 | 857 | 117,290 | 33,754 | 83,536 |
| | Small group | 7 | 496 | 18,391 | 5,032 | 13,359 |
| | Whole class | 23 | 1,210 | 75,525 | 26,632 | 48,893 |
| | Whole school | 8 | 615 | 95,769 | 23,829 | 71,940 |

schools and pupils (Weiss *et al.*, 2017). This has implications for how trial findings should be interpreted in terms of how they might apply in other settings. We can only infer that they might be applicable to other similar schools. The percentage of variability explained by the differences between trials, differences between schools and residual variance (pupils) for literacy and mathematics outcomes is presented in Table 3. The differences between trials accounted for 86% of the variability in literacy outcomes across trials and 87% of the variability in mathematics outcomes when raw data was used. (Please note, this could also result from scores of the outcome measures being on different scales.) However, standardised scores of post-test and pre-test outcomes show consistent patterns as normally observed in education trials.

Most of the variability in the outcomes was due to the differences between pupils and then due to the differences between schools. The difference in effect sizes between trials is negligible. We share the view that IPD meta-analysis of educational trials without properly accounting for the huge heterogeneity between trials will be prone to misleading conclusions. The rescaling of post-test and pre-test scores in each trial will reduce the variability between the trials, as shown in Table 3. This approach is not without its own limitations, as it may distort the distributions of the outcomes, particularly if the outcomes do not come from a common underlying construct. It should therefore be noted that this kind of comparison is vulnerable to uncertainty, which might, for example, go some way to explaining the strange effect on literacy of Shared Maths, though it could also be argued that the impact derives from the regular shared reading of mathematical word problems, which was the main shared activity.

The heterogeneity measures $Q$ and $I^2$ index, which are usually provided in the meta-analysis studies, are also reported. These measures were estimated using the variance and trial-specific effect size obtained from the Bayesian IPD model and the formula provided in the Methods section. All heterogeneity estimates are reported for overall analysis in Figures 2 and 3 later, while estimates for each subgroup are provided in Figures S2–S9 in the online Supplementary Material. However, please note that these results need to be carefully interpreted, since rescaling pre-test and post-test data has reduced the between-trial variability significantly, as discussed in Table 3.

Table 3. Percentage of total variability in literacy and mathematics outcomes explained by differences between trials, differences between schools and residual variance (pupils)

| | Pupils (%) | School (%) | Trial (%) |
|---|---|---|---|
| *Literacy* | | | |
| Raw | 12 | 2 | 86 |
| Standardised | 82 | 13 | 5 |
| *Mathematics* | | | |
| Raw | 11 | 2 | 87 |
| Standardised | 75 | 13 | 12 |

*Simplified IPD model versus two-stage models*

We present the comparison of our proposed simplified IPD meta-analysis model and two-stage methods in the online Supplementary Material (Tables S1–S4). Figure S1 shows the individual trial effect size for FSM subgroup literacy, mathematics, literacy attainment gap and mathematics attainment gap outcomes using IPD and two-stage fixed effect (FE) and random effect (RE) meta-analysis methods with raw and standardised scores. Most of the two-stage and one-stage IPD individual trial and pooled estimates corresponded well in terms of direction and magnitude. However, the IPD model produced a greater effect size for literacy outcome than the two-stage model. One of the reasons why the IPD model resulted in greater effect than a two-stage model may be because of how the weights were defined. The weights in the two-stage models were defined using standard errors approximated from confidence intervals, whilst the IPD model directly used estimated variance from the data. Figure 1

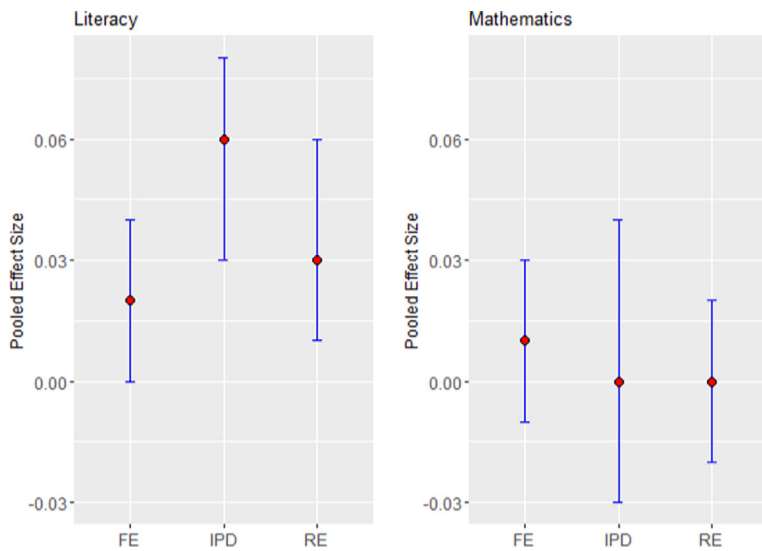


Figure 1. Overview of pooled effect size from IPD meta-analysis and two-stage fixed effect (FE) and random effect (RE) models using standardised outcome data [Colour figure can be viewed at wileyonlinelibrary.com]

Table 4. Pooled ES and credible intervals for FSM subgroup literacy and mathematics outcomes

| Outcome | No. trials | No. schools | No. FSM pupils | Pooled ES | Range (min, max ES) |
|---|---|---|---|---|---|
| Literacy | 81 | 3,804 | 90,218 | 0.06 (0.03, 0.08) | −0.20 (−0.44, 0.04) 0.42 (−0.07, 0.93) |
| Mathematics | 48 | 3,006 | 89,247 | 0.00 (−0.03, 0.04) | −0.18 (−0.36, 0.01) 0.31 (−0.25, 0.98) |

presents an overview of the pooled effect size from IPD meta-analysis and two-stage meta-analysis using standardised outcome data.

*Did pupils eligible for FSM benefit from EEF-funded interventions?*

The pooled effect size for literacy as either primary or secondary outcome across 81 trials was 0.06 (0.03, 0.08). This means on average that EEF-funded interventions had positive benefits on the literacy outcomes of FSM pupils who participated in the trials, equivalent to about 1 month's progress. However, there was no evidence from the 48 trials analysed that EEF-funded interventions had positive effects on the mathematics outcomes of FSM pupils, with an effect size of 0.00 (−0.03, 0.04). It is important to note that there was also no evidence that the interventions on average were worsening their mathematics outcomes. The estimated pooled effect sizes across all 81 trials are presented in Table 4.

Figure 2 shows the individual trial and the pooled effect size with their credible intervals. The most beneficial interventions for FSM pupils with positive effects on their literacy outcomes were Shared Maths, Graduate Coaching Programme, Accelerated Reader, Online Reading Programme (ABRA), Butterfly Phonics, Response to Intervention and Nuffield Early Language Intervention 1. The individual trial-specific effect size ranged from −0.20 to 0.42. However, it was surprising that Shared Maths was one of the most effective interventions for literacy, since it was primarily intended to improve attainment in mathematics. Although there was no evidence of overall effects on mathematics outcomes, there were promising interventions with positive effect size, such as Dialogue Teaching, Powerful Learning Conversations, Improving Numeracy and Literacy, and Act, Sing and Play. The trial-specific effect size for mathematics outcomes ranged from −0.18 to 0.31. The reports by the independent evaluators for all of these trials are available from the EEF's website.

*By key stages*

Table 5 provides the estimate of pooled effect size for the literacy and mathematics outcomes across four key stages. The maximum pooled effect size was observed for KS1 [pooled ES = 0.09 (0.02, 0.16)] and KS3 literacy outcomes [pooled ES = 0.08 (0.03, 0.13)], followed by the KS2 literacy outcome [pooled ES = 0.03 (−0.01, 0.07)]. These results clearly suggest that EEF-funded interventions were beneficial for FSM pupils in KS1 and KS3.

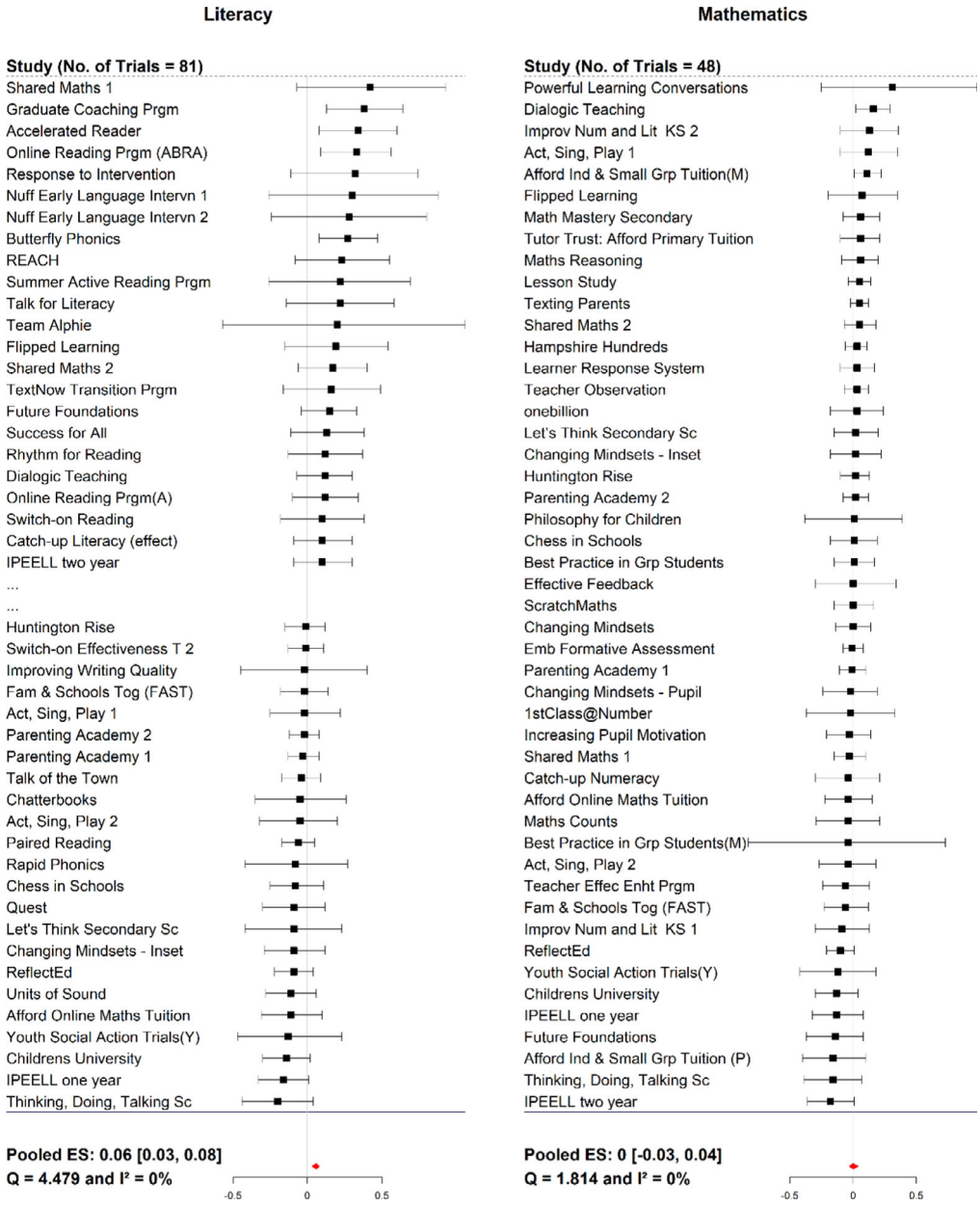


Figure 2. Forest plot of effect sizes for literacy and mathematics outcomes from FSM pupils [Colour figure can be viewed at wileyonlinelibrary.com]

The effect sizes for individual trials in KS1 were mostly positive, with the highest effect size estimate of 0.34 for the Online Reading Programme (ABRA). Few trials in KS2 had an effect size more than 0.30 SD, such as Response to Intervention and Shared Maths (Figure S2), with the maximum effect size estimate of 0.42. Most trials in KS3 also had positive effects. Accelerated Reader and Graduate Coaching Programme were the trials most beneficial for FSM pupils in KS3. Individual trial effect

Table 5. Pooled ES and credible interval for FSM subgroup literacy and mathematics outcomes by key stage

| Key stage | No. trials | No. schools | No. FSM pupils | Pooled ES | Range (min, max ES) |
|---|---|---|---|---|---|
| *Literacy* | | | | | |
| KS1 | 13 | 481 | 4,444 | 0.09 (0.02, 0.16) | −0.05 (−0.29, 0.21) 0.34 (0.09, 0.57) |
| KS2 | 33 | 2,175 | 34,085 | 0.03 (−0.01, 0.07) | −0.20 (−0.43, 0.04) 0.42 (−0.07, 0.94) |
| KS3 | 29 | 507 | 10,108 | 0.08 (0.03, 0.13) | −0.13 (−0.48, 0.22) 0.39 (0.13, 0.64) |
| KS4 | 6 | 641 | 41,581 | 0.02 (−0.05, 0.08) | −0.01 (−0.15, 0.12) 0.08 (−0.03, 0.18) |
| *Mathematics* | | | | | |
| KS1 | 9 | 540 | 4,394 | 0.02 (−0.07, 0.11) | −0.09 (−0.30, 0.13) 0.13 (−0.10, 0.35) |
| KS2 | 24 | 1,524 | 25,946 | −0.01 (−0.04, 0.03) | −0.18 (−0.37, 0.01) 0.16 (0.03, 0.29) |
| KS3 | 9 | 261 | 6,667 | 0.01 (−0.09, 0.12) | −0.16 (−0.41, 0.10 )0.31 (−0.36, 0.98) |
| KS4 | 6 | 681 | 52,240 | 0.02 (−0.03, 0.07) | −0.06 (−0.25, 0.12) 0.11 (0.00, 0.22) |

size for KS4 ranged from −0.01 to 0.08 (Table 5), and Teacher Observation and Affordable Individual and Small Group Tuition (E) were the most beneficial KS4 trials (Figure S2).

As Table 5 reveals, the pooled estimate of effect size for the mathematics outcome was about 0.02 SD for KS1 and KS4. From both the literacy and mathematics outcome analysis, EEF-funded interventions improved the literacy and mathematics scores in most key stages. In KS1, Act, Sing and Play and Improving Numeracy and Literacy were the most beneficial trials, where the effect size was more than 0.10 SD (Figure S3). In KS2, there were few trials that had a positive impact on the FSM pupils' scores, and the Dialogue Teaching trial had a maximum effect size of 0.16. The individual trial effect size ranged from −0.18 to 0.16 in KS2 (Table 5). Though it is worth noting that the larger trials in KS2 had mostly positive effect sizes. Powerful Learning Conversations and Math Mastery Secondary were the most beneficial interventions for KS3 FSM pupils (Figure S3). The Affordable Individual and Small Group Tuition trial in KS4 was the most beneficial, and improved the literacy outcome of FSM pupils by more than 0.10 SD (Figure S3).

### *By types of intervention*

The effects of one-to-one and small-group interventions on literacy outcomes were greater than whole-class or whole-school interventions. Small-group interventions had a pooled effect size of 0.14 (0.06, 0.22), whilst one-to-one interventions had a pooled effect size of 0.08 (0.04, 0.13). Both types of intervention improved the literacy of FSM pupils by an equivalent of more than 1 month's progress according to the EEF scale (Table 6).

One-to-one interventions, namely Graduate Coaching Programme and Accelerated Readers, were most beneficial. Small-group interventions such as Shared Maths followed by Butterfly Phonics benefitted FSM pupils the most. Flipped Learning was the most beneficial whole-class and Success for All the most beneficial whole-school intervention for FSM pupils (Figure S4).

Table 6. Pooled ES and credible intervals for FSM subgroup literacy and mathematics outcomes by type of intervention

| Type of intervention | No. trials | No. schools | No. FSM pupils | Pooled ES | Range (min, max ES) |
|---|---|---|---|---|---|
| *Literacy* | | | | | |
| One-to-one | 24 | 1,260 | 28,194 | 0.08 (0.04, 0.13) | −0.11 (−0.28, 0.05) 0.38 (0.14, 0.64) |
| Small group | 17 | 463 | 6,914 | 0.14 (0.06, 0.22) | −0.12 (−0.47, 0.24) 0.42 (−0.06, 0.92) |
| Whole class | 30 | 1,286 | 29,774 | 0.01 (−0.04, 0.05) | −0.20 (−0.43, 0.04) 0.18 (−0.15, 0.53) |
| Whole school | 10 | 795 | 25,336 | 0.02 (−0.02, 0.06) | −0.04 (−0.17, 0.10) 0.14 (−0.10, 0.38) |
| *Mathematics* | | | | | |
| One-to-one | 10 | 777 | 33,754 | 0.04 (−0.04, 0.12) | −0.04 (−0.22, 0.15) 0.30 (−0.29, 0.95) |
| Small group | 7 | 452 | 5,032 | −0.04 (−0.11, 0.03) | −0.15 (−0.40, 0.09) 0.05 (−0.08, 0.18) |
| Whole class | 23 | 1,163 | 26,632 | −0.01 (−0.06, 0.05) | −0.18 (−0.37, 0.02) 0.15 (0.02, 0.29) |
| Whole school | 8 | 614 | 23,829 | 0.02 (−0.02, 0.07) | −0.06 (−0.25, 0.13) 0.07 (−0.08, 0.22) |

Table 6 also shows that one-to-one and whole-school interventions had a positive effect on mathematics outcomes of FSM pupils. In contrast, small-group or whole-class interventions had a negative impact. However, it should be noted that the number of FSM pupils in small-group interventions was much lower than in the other types of intervention.

Powerful Learning Conversations and Affordable Tuition projects were the most beneficial one-to-one interventions. Shared Maths and OneBillion were the most beneficial small-group interventions (Figure S5). Even though the pooled effect of the class-level intervention was negative, trials such as Dialogue Teaching and Act, Sing and Play improved FSM pupils' scores by more than 0.10 SD.

*Are the interventions reducing attainment gaps between FSM pupils and their peers?*

In literacy, the reduction in the attainment gap between FSM and non-FSM pupils was close to zero, but positive. This seems to suggest that on average, EEF-funded interventions had similar effects for both FSM and non-FSM pupils across all trials. There is no evidence to suggest that EEF-funded interventions had widened attainment gaps in literacy between FSM and non-FSM pupils (Table 7). This is important because of the so-called 'Matthew effect', where interventions tend to widen the spread of attainment as more successful pupils may benefit more from additional

Table 7. Pooled attainment gap and credible interval for the study outcomes

| Outcome | No. trials | No. schools | No. pupils | Pooled attainment gap | Range (min, max attainment gap) |
|---|---|---|---|---|---|
| Literacy | 81 | 4,000 | 302,138 | 0.01 (−0.01, 0.04) | −0.27 (−0.53, 0.00) 0.42 (−0.22, 1.06) |
| Mathematics | 48 | 3,178 | 306,975 | −0.01 (−0.04, 0.02) | −0.43 (−0.78, −0.06) 0.20 (0.04, 0.35) |

support (Pfost *et al.*, 2012). Similarly, the attainment gap in mathematics was also closer to zero. It can therefore be argued that there was no evidence of the attainment gaps widening for mathematics.

Figure 3 shows the individual trial and the average attainment gap for both the literacy and mathematics outcomes. More than half of the trials had positive attainment gaps in literacy scores, which means that on average FSM pupils were more likely to benefit than their peers. The attainment gap in literacy scores between FSM and non-FSM pupils was more than 0.20 SD for trials such as Text Now Transition Programme, Affordable Individual and Small Group Tuition Programme, Nuffield Early Language Intervention, Improving Numeracy and Literacy, and Best Practice in Grouping Students. However, the attainment gap in mathematics between FSM and non-FSM pupils was closer to 0.0 SD.

### *Attainment gaps by key stages*

The attainment gap in literacy between FSM and non-FSM pupils appears to be decreasing linearly with key stages (Table 8). KS1 had the pooled estimate of 0.07 (0.00, 0.14), whilst KS4 had the negative pooled attainment gap and interventions in favour of the non-FSM pupils.

In KS1, individual trial effect size ranged from −0.11 to 0.43 (Table 8), with the maximum positive attainment gap for Nuffield Early Language Intervention (Figure S6). The individual trial attainment gap in KS2 ranged from −0.24 to 0.16. Shared Maths and Response to Intervention trials in KS2 had benefitted FSM pupils more than their counterparts. More than two-thirds of the trials in KS3 had positive attainment gap. In KS4, four trials had benefitted FSM pupils more than non-FSM pupils (Figure S6).

The pooled attainment gaps for all key stage mathematics outcomes was zero, except for KS2 mathematics. However, the attainment gap between FSM and non-FSM pupils is very low for all key stages (Table 8). In KS1, two trials (Act, Sing and Play, and Improving Numeracy and Literacy) benefitted FSM pupils the most, with an attainment gap of more than 0.10 SD. The individual trial effect size in KS2 ranged from −0.43 to 0.10. The attainment gap for the Affordable Maths trial in KS2 was 0.10 SD. Let's Think Secondary Science and Changing Mindsets—Pupil were the two most beneficial trials for KS3 FSM pupils (Figure S7). There were few trials in KS4, such as Affordable Individual and Small Group Tuition (M) and Teacher Observation, with positive attainment gaps in favour of FSM pupils, though the overall pooled attainment gap was zero.

Overall, comparison of the attainment gaps across the key stages was positive for literacy and mostly negative for mathematics. KS3 was the only subgroup where the attainment gap was positive for both literacy and mathematics. This indicates that FSM pupils in KS3 tended to benefit more than non-FSM pupils.

### *Attainment gaps by types of intervention*

The attainment gap between FSM and non-FSM pupils' literacy outcomes was higher for one-to-one and small-group interventions than class or whole-


Figure 3. Forest plot with attainment gap between FSM and their peers by study outcomes [Colour figure can be viewed at wileyonlinelibrary.com]

school interventions (Table 9). On average, FSM pupils were 0.02 and 0.05 SD better than non-FSM pupils for one-to-one and small-group interventions, respectively.

Table 8. Pooled attainment gap and credible interval for literacy and mathematics outcomes by key stage

| Key stage | No. trials | No. schools | No. pupils | Pooled attainment gap | Range (min, max attainment gap) |
|---|---|---|---|---|---|
| *Literacy* | | | | | |
| KS1 | 13 | 529 | 19,905 | 0.07 (0.00, 0.14) | −0.11 (−0.38, 0.15) 0.43 (−0.19, 1.05) |
| KS2 | 33 | 2,265 | 102,835 | 0.00 (−0.03, 0.03) | −0.24 (−0.44, −0.05) 0.16 (−0.03, 0.36) |
| KS3 | 29 | 552 | 39,297 | 0.01 (−0.05, 0.07) | −0.27 (−0.53, 0.00) 0.34 (−0.01, 0.68) |
| KS4 | 6 | 654 | 140,101 | 0.00 (−0.03, 0.03) | −0.07 (−0.13, −0.01) 0.09 (−0.06, 0.23) |
| *Mathematics* | | | | | |
| KS1 | 9 | 639 | 18,718 | 0.00 (−0.06, 0.07) | −0.26 (−0.48, −0.03) 0.18 (−0.07, 0.42) |
| KS2 | 24 | 1,577 | 79,671 | −0.02 (−0.06, 0.00) | −0.43 (−0.79, −0.08) 0.10 (−0.01, 0.20) |
| KS3 | 9 | 269 | 30,434 | 0.02 (−0.07, 0.10) | −0.16 (−0.45, 0.13) 0.20 (0.05, 0.34) |
| KS4 | 6 | 693 | 178,152 | 0.00 (−0.02, 0.02) | −0.05 (−0.12, 0.03) 0.08 (−0.01, 0.17) |

Table 9. Pooled attainment gaps and credible interval for literacy and mathematics outcomes by type of intervention

| Intervention type | No. trials | No. schools | No. pupils | Pooled attainment gap | Range (min, max attainment gap) |
|---|---|---|---|---|---|
| *Literacy* | | | | | |
| One-to-one | 24 | 1,358 | 97,368 | 0.02 (−0.04, 0.07) | −0.27 (−0.54, 0.00) 0.34 (0.00, 0.68) |
| small group | 17 | 503 | 22,451 | 0.05 (−0.04, 0.14) | −0.23 (−0.86, 0.44 ) 0.42 (−0.23, 1.07) |
| Whole class | 30 | 1,339 | 83,550 | 0.00 (−0.04, 0.04) | −0.24 (−0.43, −0.04) 0.30 (−0.18, 0.80) |
| Whole school | 10 | 800 | 98,769 | 0.00 (−0.03, 0.03) | −0.09 (−0.22, 0.03) 0.03 (−0.02, 0.07) |
| *Mathematics* | | | | | |
| One-to-one | 10 | 857 | 117,290 | −0.05 (−0.12, 0.01) | −0.44 (−0.80, −0.09) 0.10 (−0.02, 0.21) |
| small group | 7 | 496 | 18,391 | −0.06 (−0.14, 0.02) | −0.26 (−0.49, −0.04) 0.03 (−0.07, 0.13) |
| Whole class | 23 | 1,210 | 75,525 | 0.02 (−0.03, 0.06) | −0.19 (−0.38, −0.01) 0.20 (0.05, 0.34) |
| Whole school | 8 | 615 | 95,769 | 0.00 (−0.02, 0.03) | −0.04 (−0.09, 0.01) 0.02 (−0.05, 0.09) |

Half of one-to-one and small-group interventions had positive attainment gaps, suggesting that FSM pupils were more likely to benefit from these interventions. Text Now Transition Programme (one-to-one) and Nuffield Early Language Intervention (small group), Best Practice in Grouping Students (whole class) and Lesson Study trial (whole school) interventions are the most beneficial for FSM pupils (Figure S8). Table 9 also shows that the pooled attainment gaps in mathematics scores between FSM and non-FSM pupils was positive for whole-class interventions. One-to-one

and small-group interventions were the least beneficial for FSM pupils. This is contradictory to the pattern for literacy performance, where one-to-one and small-group interventions were the most beneficial for FSM pupils. The trial-specific attainment gap in one-to-one intervention varied from −0.44 to 0.10, small group varied from −0.26 to 0.03, whole class varied from −0.19 to 0.20 and whole school varied from −0.04 to 0.02 (Table 9). The Affordable Online Maths Tuition one-to-one intervention had attainment gap of more than 0.10 SD (Figure S9). Overall, one-to-one or small-group interventions were more effective for literacy, while whole-class and whole-school interventions appeared to be more beneficial and reduced attainment gaps in mathematics.

*Risk of bias assessment*

Flaws in the study design and reporting of randomised trials can lead to under-rated or over-rated impact of interventions. The risk of bias of the pooled effect size for the security or 'padlock' rating of trials was assessed by excluding the trials with lower than three padlocks. As part of the evaluation process, the EEF classifies the security of its trials with a rating system based on key threats to internal validity (EEF, 2019). Although the specific wording of the classification framework has developed over time, the broad categories and elements for classification have remained consistent. The security ratings of the EEF's educational trials varies from low (padlock = 0) to the best type of evidence that could be expected from a study (padlock = 5) (EEF, 2019). Table 10 presents the results of the sensitivity analysis alongside the main analysis for all trials. There was no evidence to suggest that padlock ratings were substantially related to the average effect of the interventions or the average attainment gaps between FSM and non-FSM pupils and attainment gaps from the trials.

## Discussion and conclusions

Effective practices or interventions need to be developed for FSM pupils in order to reduce the attainment gap between FSM pupils and their peers. With this aim, an IPD meta-analysis was conducted to synthesise evidence of the overall impact of EEF-funded education interventions on FSM pupils and quantify the effect of the interventions on the gaps between FSM pupils and their peers. Meta-analysis helps to

Table 10. Sensitivity analysis for literacy and mathematics outcomes by excluding trials with less than three padlocks

| Outcome | Effect | No. schools | No. pupils | All | No. schools | No. pupils | Padlocks ≥ 3 |
|---|---|---|---|---|---|---|---|
| Literacy | FSM | 3,804 | 90,218 | 0.06 (0.03, 0.08) | 2,337 | 48,216 | 0.06 (0.03, 0.10) |
| | Gap | 4,000 | 302,138 | 0.01 (−0.01, 0.04) | 2,436 | 156,004 | 0.02 (−0.02, 0.06) |
| Mathematics | FSM | 3,006 | 89,247 | 0.00 (−0.03, 0.04) | 1,990 | 49,783 | 0.01 (−0.02, 0.05) |
| | Gap | 3,178 | 306,975 | −0.01 (−0.03, 0.02) | 2,115 | 165,735 | −0.00 (−0.04, 0.03) |

counteract the risk that individual studies may be underpowered due to the smaller sample size (Moher & Olkin, 1995), which is often a concern for FSM pupils in education trials. There has been no previous attempt in education research to systematically review such a large archive of individual pupil data in education trials and provide reliable individual and pooled estimates of effect size and attainment gap for the key study outcomes of FSM pupils—describing these outcomes by a range of important factors such as type of intervention and key stage of pupils using a robust approach of research synthesis through IPD meta-analysis. The approach is not without its limitations and challenges. Some of these relate to the use of effect sizes, as noted above, challenges associated with summarising trials' pooled effects due to large heterogeneity in education trial outcomes and relaxing this assumption that the effects in each trial are from a single distribution or related to common underlying factors. Identifying patterns or characteristics of successful interventions also loses the causal warrant from the RCT design. Common features are associations rather than causal mechanisms.

Overall, EEF-funded interventions had beneficial impacts on the literacy performance of pupils eligible for FSM, compared to mathematics performance, which showed no overall effect. Attainment gap estimates showed that literacy outcomes for FSM pupils had improved more than those for non-FSM pupils with EEF-funded interventions. Mathematics performance was affected in a similar way for both FSM pupils and their non-FSM peers. In the last decade, several programmes were developed to assist children with mathematics attainment in England (See *et al.*, 2019) and worldwide. However, there is clearly a need to identify mathematics interventions which can benefit FSM pupils and other young people in most need of such interventions. Act, Sing and Play, Improving Numeracy and Literacy, and Affordable Maths were some of the most promising interventions as observed in this study for FSM pupils.

Across key stages, we observed that FSM pupils benefitted from EEF-funded interventions in all key stages except KS2 for mathematics. However, when comparing FSM pupils with non-FSM pupils, EEF-funded interventions helped FSM pupils across key stages to perform better than others, as observed from the positive but low attainment gap estimates. Although another interpretation of this finding can be that FSM pupils' academic performance has not fallen behind non-FSM pupils who participated in EEF trials. The attainment gap estimate for KS3 was positive for both mathematics and literacy outcomes, indicating that EEF-funded interventions improved both mathematics and literacy performances of FSM pupils slightly more than non-FSM pupils in KS3.

By type of intervention, individual or small-group interventions improved the literacy outcome of FSM pupils considerably, while interventions with a focus on the whole class were beneficial for mathematics performances. Evidence from previous meta-analyses also suggested that small-group or individual interventions are beneficial for children's educational outcomes (e.g. Lou *et al.*, 2001).

Reliability of the estimates of pooled effect size and pooled attainment gap was assessed by risk of bias assessment, which shows that our estimates were consistent across different methodological approaches, even after excluding a few trials which were less robust.

Overall, evidence from this study demonstrates the value of IPD meta-analysis as an approach to identify and understand educational interventions with positive impacts which can be implemented in schools to improve the educational attainment of FSM children. Using IPD meta-analysis provided a better understanding of the effects of different interventions, which can inform decisions about specific interventions to target disadvantaged pupils and can be used to suggest ways to improve the design or implementation of the tested interventions among FSM children. The key emphasis of this article was mainly on the feasibility of the Bayesian IPD meta-analysis approach in education trials, rather than showing its superiority over two-stage aggregate meta-analyses. Future work using synthetic data will aim to establish the superiority of the method over other methods for meta-analysis.

The analysis also highlights the challenge of addressing disadvantage through educational intervention and using evidence from research to improve outcomes for FSM pupils. It certainly indicates the extent of the challenge of identifying and scaling possible solutions to reducing educational inequity in schools. Decisions about specific interventions require the careful accumulation of evidence in different forms over time, including RCTs and meta-analysis for that specific intervention. Without replication studies, this challenge is greater as it will only be by looking at similarities between approaches that successful factors can be identified both in terms of interventions and approaches, but also by investigating which approaches are successful for different groups of pupils and accumulating that evidence through approaches such as IPD meta-analysis and other forms of synthesis.

## Author contributions

A. K., B. A, A. S. and G. U. designed the model and the computational framework, and analysed the data. B. A., A. S. and G. U. carried out the implementation and performed the analysis. A. K., B. A, A. S., G. U. and S. H. wrote the manuscript with input from all authors. S. H. and A. K. conceived the study and were in charge of overall direction and planning.

## Conflict of interest

The authors declare no conflict of interest that they are aware of.

## Funding statement

We are grateful to the Education Endowment Foundation for providing a grant to Durham University, UK to conduct this research work.

## Data availability statement

The data is not open access to ensure anonymity and confidentiality, given the complexity of de-identification.

## References


Bloom, H.S., Hill, C.J., Black, A.R. & Lipsey, M.W. (2008) Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions, *Journal of Research on Educational Effectiveness*, 1(4), 289–328.

Borenstein, M., Hedges, L.V., Higgins, J.P. & Rothstein, H.R. (2011) *Introduction to meta-analysis* (Chichester, Wiley).

Bowden, J., Tierney, J.F., Copas, A.J. & Burdett, S. (2011) Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics, *BMC Medical Research Methodology*, 11(1), 1–12.

Brookes, S.T., Whitley, E., Peters, T.J., Mulheran, P.A., Egger, M. & Davey Smith, G. (2001) Subgroup analysis in randomised controlled trials: Quantifying the risks of false-positives and false-negatives, *Health Technology Assessment*, 5(33), 1–56.

Burke, D.L., Ensor, J. & Riley, R.D. (2017) Meta-analysis using individual participant data: One-stage and two-stage approaches, and why they may differ, *Statistics in Medicine*, 36(5), 855–875.

Chowdry, H., Crawford, C., Dearden, L., Joyce, R., Sibieta, L., Sylva, K. *et al* (2010) *Poorer children's educational attainment: How important are attitudes and behaviour* (York, Joseph Rowntree Foundation), 1–72.

Cochran, W.G. (1954) The combination of estimates from different experiments, *Biometrics*, 10, 101–129.

Cochrane (2019) *Standard errors from confidence intervals and P values: Difference measures*. Available online at: https://handbook-5-1.cochrane.org (accessed 9 August 2019).

Coe, R., Weidmann, B., Coleman, R. & Kay, J. (2020) *Impact of school closures on the attainment gap: Rapid evidence assessment* (London, Education Endowment Foundation).

Debray, T.P., Moons, K.G., van Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R.H. *et al* (2015) Get real in individual participant data (IPD) meta-analysis: A review of the methodology, *Research Synthesis Methods*, 6(4), 293–309.

DfE (2014) *Children and Families Act* (London, HMSO).

Dietrichson, J., Bøg, M., Filges, T. & Klint Jørgensen, A.-M. (2017) Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis, *Review of Educational Research*, 87(2), 243–282.

Duchateau, L., Janssen, P. & Rowlands, J. (1998) Linear mixed models: An introduction with applications in veterinary research. ILRI (aka ILCA and ILRAD).

DWP (2013) *Free school meal entitlement and child poverty in England*. Available online at: https://dera.ioe.ac.uk/19084/1/ (accessed 1 June 2020).

Education Endowment Foundation (2019) Available online at: https://educationendowmentfoundation.org.uk/ (accessed 1 June 2019).

Fanshawe, T.R. & Perera, R. (2019) Conducting one-stage IPD meta-analysis: Which approach should I choose?, *BMJ Evidence-based Medicine*, 24(5), 190–190.

Gorard, S. (2012) Who is eligible for free school meals? Characterising free school meals as a measure of disadvantage in England, *British Educational Research Journal*, 38(6), 1003–1017.

Hedges, L. & Olkin, I. (1985) *Statistical methods for meta-analysis* (New York, Academic Press).

Higgins, S. (2016) Meta-synthesis and comparative meta-analysis of education research findings: some risks and benefits. *Review of Education*, 4(1), 31–53.

Higgins, S. (2018) *Improving learning: Meta-analysis of intervention research in education* (Cambridge, Cambridge University Press).

Higgins, J.P.T. & Thompson, S.G. (2002) Quantifying heterogeneity in a meta-analysis, *Statistics in Medicine*, 21, 1539–1558.

Hobbs, G. & Vignoles, A. (2010) Is children's free school meal 'eligibility' a good proxy for family income?, *British Educational Research Journal*, 36(4), 673–690.

Humphrey, N., Wigelsworth, M., Barlow, A. & Squires, G. (2013) The role of school and individual differences in the academic attainment of learners with special educational needs and disabilities: A multi-level analysis, *International Journal of Inclusive Education*, 17(9), 909–931.

Jerrim, J., Greany, T. & Perera, N. (2018) *Educational disadvantage: How does England compare?* (London, Education Policy Institute).

Kontopantelis, E. (2018) A comparison of one-stage vs two-stage individual patient data meta-analysis methods: A simulation study, *Research Synthesis Methods*, 9(3), 417–430.

Kruschke, J.K. (2015) *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd edn) (Waltham, MA, Academic Press).

Laird, N. (2004) *Analysis of longitudinal and cluster-correlated data* (Beachwood, OH: Institute of Mathematical Statistics.

Lesaffre, E. & Lawson, A.B. (2012) *Bayesian biostatistics* (Chichester, Wiley).

Lou, Y., Abrami, P.C. & d'Apollonia, S. (2001) Small group and individual learning with technology: A meta-analysis, *Review of Educational Research*, 71(3), 449–521.

Manstead, A. (2014) *The wellbeing effect of education: Evidence briefing* (Swindon, Economic and Social Research Council).

Marmot, M. (2010) *Fair society, healthy lives: The Marmot Review* (London, University College).

Moher, D. & Olkin, I. (1995) Meta-analysis of randomized controlled trials: A concern for standards, *JAMA*, 274(24), 1962–1964.

Pfost, M., Dörfler, T. & Artelt, C. (2012) Reading competence development of poor readers in a German elementary school sample: An empirical examination of the Matthew effect model, *Journal of Research in Reading*, 35(4), 411–426.

Pigott, T. (2012) *Advances in meta-analysis* (New York, Springer).

Raftery, A.E. & Lewis, S.M. (1995) The number of iterations, convergence diagnostics and generic Metropolis algorithms, *Practical Markov Chain Monte Carlo*, 7(98), 763–773.

Reade, M.C., Delaney, A., Bailey, M.J. & Angus, D.C. (2008) Bench-to-bedside review: Avoiding pitfalls in critical care meta-analysis – funnel plots, risk estimates, types of heterogeneity, baseline risk and the ecologic fallacy, *Critical Care*, 12(4), 220.

Schochet, P.Z., Puma, M. & Deke, J. (2014) *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods*. Report No. NCEE, 4017 (Washington, DC, US Department of Education).

Scottish Government (2014) *Raising attainment for all – Scotland: The best place in the world to go to school* (Edinburgh, Scottish Government).

See, B.H., Morris, R., Gorard, S. & Siddiqui, N. (2019) Evaluation of the impact of Maths Counts delivered by teaching assistants on primary school pupils' attainment in maths, *Educational Research and Evaluation*, 25(3–4), 203–224.

Simpson, A. (2018) Princesses are bigger than elephants: Effect size as a category error in evidence-based education, *British Educational Research Journal*, 44, 897–913.

Singh, A., Uwimpuhwe, G., Li, M., Einbeck, J., Higgins, S. & Kasim, A. (2021) Multisite educational trials: Estimating the effect size and its confidence intervals, *International Journal of Research & Method in Education*. https://doi.org/10.1080/1743727X.2021.1882416

Smith, C.T., Marcucci, M., Nolan, S.J., Iorio, A., Sudell, M., Riley, R., Williamson, P.R. *et al* (2016) Individual participant data meta-analyses compared with meta-analyses based on aggregate data. *Cochrane Database of Systematic Reviews* (9).

Spencer, S. (2015) *The cost of the school day*. Report (Glasgow, Child Poverty Action Group in Scotland).

Taylor, C. (2018) The reliability of free school meal eligibility as a measure of socio-economic disadvantage: Evidence from the Millennium Cohort Study in Wales, *British Journal of Educational Studies*, 66(1), 29–51.

Thomas, D., Radji, S. & Benedetti, A. (2014) Systematic review of methods for individual patient data meta-analysis with binary outcomes, *BMC Medical Research Methodology*, 14(1), 79.

Uwimpuhwe, G., Singh, A., Higgins, S. & Kasim, A. (2020) Application of Bayesian posterior probabilistic inference in educational trials, *International Journal of Research & Method in Education*. 10.1080/1743727X.2020.1856067

Weiss, M.J., Bloom, H.S., Verbitsky-Savitz, N., Gupta, H., Vigil, A.E. & Cullinan, D.N. (2017) How much do the effects of education and training programs vary across sites? Evidence from

past multisite randomized trials, *Journal of Research on Educational Effectiveness*, 10(4), 843–876.

Xiao, Z., Kasim, A. & Higgins, S. (2016) Same difference? Understanding variation in the estimation of effect sizes from educational trials, *International Journal of Educational Research*, 77, 1–14.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Supplementary Material