1 Introduction

The Education Endowment Foundation (EEF) is an independent grant-making charity, which aims to address the challenge of disadvantage in educational achievement associated with family income and to help children from all backgrounds achieve academically. Established in 2011 with a £125 million endowment from the Department for Education, the EEF is dedicated to raising the educational attainment of disadvantaged children in primary and secondary schools in England using research and evidence in three ways. This is first by identifying and funding promising educational innovations that address the needs of children facing disadvantage; second by evaluating these innovations to extend the evidence on what is educationally effective and what can be made to work at scale; and third by encouraging schools, governments, charities, and others to apply evidence and adopt innovations found to be successful.

This paper focuses on the second of these approaches and presents a repeated analysis of evaluation results from 17 educational trials (see Table 1) which all reported findings publicly in 2014 –15. All EEF projects are independently evaluated by a number of evaluation teams which are from universities and independent research organisations. The data from these projects are deposited in an archive which will become a rich repository of findings from EEF interventions (over 100 have been commissioned so far involving over 650,000 pupils). One goal is to track the longer term impact of interventions as results from national tests become available where this is possible.

1.1 Rationale for the archive analysis

Andrew Gelman described statistics as "the science of defaults" (in Lin et al., 2014, p. 293), by which, he meant applied statisticians usually choose (and recommend) their default or preferred methods to solve problems in a wide range of settings, although these may not always be optimal in answering project-specific questions. In the EEF reports that are made publicly available, there are patterns of design and analysis associated with specific evaluation teams. As shown in the tables that follow, evaluators sometimes applied the same approach to different projects, even when the research designs and the quality of the data for causal inference varied. This arises, as Gelman noted, because there are competing philosophies, assumptions, and approaches to statistical analysis and inference, which makes consensus on *the* best approach difficult to achieve. This paper explores the differences these choices make in terms of the outcomes from different methods of analysis for each trial.

Archive analysis differs from replication studies in that the former does not require the collection of new data from the same population. Instead, it re-uses the data from original trials to reproduce the original results and/or answer new research questions. In this paper, our goal is mainly to answer a new question: how do effect size estimates and their uncertainties vary under different

project	archive label	full EEF title	evaluation report
1	ffe, ffm	Future Foundations	Gorard, Siddiqui, and See (2014)
2	sor	Switch-on Reading	Gorard, See, and Siddiqui (2014)
3	gfw	Grammar for Writing	D. Torgerson, Torgerson, Mitchell, et al. (2014)
4	rfr	Rhythm for Reading	Styles, Clarkson, and Fowler (2014b)
9	catchn, catcht	Catch Up Numeracy	Rutt (2014)
10	cbks+, cbks	Chatterbooks	Styles, Clarkson, and Fowler (2014a)
13	rp	Rapid Phonics	King and Kasim (2015)
14	ar	Accelerated Reader	Gorard et al. (2015a)
15	bp	Butterfly Phonics	Merrell and Kasim (2015)
16	iwq	Improving Writing Quality	D. Torgerson, Torgerson, Ainsworth, et al. (2014)
17	sar	Summer Active Reading	Maxwell et al. (2014a)
18	text	TextNow	Maxwell et al. (2014b)
21	uos	Units of Sound	Sheard et al. (2015)
22	ve	Vocabulary Enrichment	Styles, Stevens, et al. (2014)
31	fs	Fresh Start	Gorard et al. (2015b)
32	tfl	Talk for Literacy	Styles and Bradshaw (2015)
38	mms	Mathematics Mastery Secondary	Jerrim et al. (2015)

Table 1: Project information. The numbers 1 to 38 are EEF project numbers. We abbreviate full EEF titles to labels that mark each of the 20 outcomes for this study. The references to the 17 evaluation reports can also be used to identify evaluation teams.

model and design specifications? Unlike meta-analyses, which usually rely on summary statistics extracted from secondary sources that do not always report research in consistent and transparent ways to synthesise evidence, this archive analysis re-evaluates the evidence already found from EEF trials. In other words, it investigates how sensitive the findings are to design and model specifications, using full datasets from the aforementioned evaluation projects. It also aims to explain what causes any variation in impact and to support any subsequent comparison of impact between the studies examined.

The educational interventions included in this analysis all set out to improve educational attainment for school-age pupils and mainly targeted literacy and/or mathematics outcomes, with some focusing on phonics, vocabulary, grammar or other aspects of literacy, some through summer school interventions, others in schools as pedagogical interventions, such as those based on developing mastery or promoting learning through talk or thinking strategies. The samples varied in size from 178 to 5830 pupils, with numbers of schools (clusters) involved varying from three to 54 (see Table 2). Full details of the interventions and evaluations can be found in the individual evaluation reports which are listed in the references.

study	N	eval.n	n	n.t	n.c	n.sch
Future Foundations (English)	354	310	310	167	143	33
Future Foundations (Maths)	354	306	303	162	141	33
Switch-on Reading	308	308	308	155	153	19
Grammar for Writing	2500	1982	1367	667	700	50
Rhythm for Reading	419	355	355	175	180	6
Catch Up Numeracy (Catch)	318	216	216	108	108	54
Catch Up Numeracy (Time)	318	210	210	102	108	54
Chatterbooks Plus	577	295	303	154	149	12
Chatterbooks	577	304	311	162	149	12
Rapid Phonics	206	174	178	86	92	21
Accelerated Reader	349	339	326	167	159	4
Butterfly Phonics	302	310	302	159	143	6
Improving Writing Quality	920	261	265	144	121	22
Summer Active Reading	182	182	182	93	89	48
TextNow	391	391	391	199	192	54
Units of Sound	427	427	427	225	202	33
Vocabulary Enrichment	626	570	570	282	288	12
Fresh Start	423	419	419	215	204	10
Talk for Literacy	235	213	213	106	107	3
Mathematics Mastery Secondary	7712	5938	5830	3197	2633	44

Table 2: Summary statistics of 20 effect size estimates from 17 EEF projects. `N` is the total number of observations in the data we have access to. `eval.n` is the sample sizes evaluation teams reported for their analyses. `n` is the sample sizes used for the archive analysis, where `n.t` and `n.c` are sample sizes for treatment and control groups. `n.sch` is the number of schools for each study.

1.2 Effect size and p-value

A key concept in this paper is that of effect size, which, according to Borenstein (2009), is an index used to quantify the magnitude of relationship between two variables or the difference between two groups (p. 222). In theory, effect sizes from different studies, regardless of the design, should measure, approximately at least, the same relationship and be comparable. Like p-values, effect sizes are scale free (Hedges, 2008, p. 168). The two are certainly related to each other, but they are not the same – a significant p-value could be a function of a large effect or a small effect in a study with a large sample size, likewise, a big p-value could reflect a small effect or a large effect in a small study (Borenstein, 2009, p. 223). Effect size estimates are based on the samples studied, and the uncertainties surrounding those point estimates give us a range of possible effect sizes for the corresponding populations. While the calculation of effect size is a mathematical process, its interpretation involves judgement, and it is of little practical value to say an effect is large or small without comparing it with others in a specific context (Hedges, 2008, p. 170).

2 Methods

For better comparison, the 17 projects selected for this study are all Randomised Controlled Trials (RCTs), though with different specific designs. As Rubin (2008b) noted, RCTs and comparative observational studies should form a continuum rather than a dichotomy in terms of suitability for causal inference (p. 810), which means well designed observational studies may be better suited for causal inference than poorly implemented RCTs with systematic bias and/or serious attrition. Our focus is on the analytic strategies adopted and the variation in estimates of impact had other analytic models been adopted.

2.1 Pluralists' dilemma

As Table 3 indicates, the analytic models evaluators employed differ not only in regression forms and the number of covariates, variables that are not affected by treatments (Rubin, 2007, p. 33), but also in their choice of raw or transformed outcomes, which itself can be primary or secondary (for another description of the phenomenon, see Olken, 2015, p. 62). This diversity in model construction might be problematic for the reasons outlined below.

project	methods	covariates
1	Δ	`t + pret (r) + pret (w) + pret (m) + fsm + sex + sen + eth`
2	Δ	`t + pret + age + eth + sex + fsm + sen + eal + dosage`
3	MLM	`t + pret + sex + fsm + eal + age`, random = $\sim 1 \mid$ `(sch + class)`
4	OLS	`t + pret + sch + sex + fsm + age`
9	MLM	`t + pret`, random = $\sim 1 \mid$ `sch`
10	RMM	`t + pret + sex + age + fsm + time*t`, random = $\sim 1 \mid$ `(time + sch)`
13	MLM	`t + pret`, random = $\sim 1 \mid$ `sch`
14	Δ	`t + pret + age + fsm + sen + eal + eth`
15	MLM	`t + pret`, random = $\sim 1 \mid$ `sch`
16	MLM	`t + pret + sex + eal + fsm + age`, random = $\sim 1 \mid$ `(class + sch)`
17	MLM	`t + pret + sex + fsm`, random = $\sim 1 \mid$ `sch`
18	MLM	`t + pret + sex + fsm`, random = $\sim 1 \mid$ `sch`
21	MLM	`t + pret + phase`, random = $\sim 1 \mid$ `sch`
22	OLS	`t + pret + sch + sex + fsm + age`
31	Δ	`t + pret + age + sex + fsm + eth + sen + eal + dosage`
32	OLS	`t + pret + sch + sex + fsm + age`
38	OLS	`t + pret + sex + fsm + eth`

Table 3: Regression models and covariates evaluation teams used in their primary analysis. When the primary method is Δ, either a post-test only comparison or difference in average gain scores, the original evaluation team also used multiple regression with the above-listed covariates to check the main results. As shown above, apart from treatment indicator (`t`) and pre-test scores (`pret`), the number and type of covariates added vary a lot. For instance, Project 1 has three pre-test scores in reading, writing, and maths. Some have dosage and time effects, whereas others include interaction terms.

First, as the number of covariates increases, so does the pool of potential models for analysis, which may encourage successful data dredging, particularly when sample size is small (Humphreys, Sanchez de la Sierra, & van der Windt, 2013, p. 7). Second, when the outcome is used to identify significant covariates, the resulting model with those covariates is no longer objective (Rubin, 2008b, p. 837). That is to say, any "significant" findings after seeing the outcome lack statistical rigour and should only be considered as "exploratory" to guide further research (Olken, 2015, p. 62).

Second, the effect size for a given predictor depends on what other covariates are also included in a regression model. If the number of covariates differs, as Nakagawa and Cuthill (2007) argued, it could be inappropriate to compare effect sizes from those models, because the total variance "to be explained" (p. 597) is no longer the same. Hedges (2008) also stressed that effect size estimations that involve statistical controls "depend on what is being controlled" (p. 170), and it is crucial to note that estimates coming from studies that control for different covariates "may not be comparable" (p. 170).

Among the models evaluators employed, Δ is the difference-in-means of either post-test or gain scores. OLS and MLM represent Ordinary Least Squares and Multilevel Modelling respectively. RMM in Project 10 refers to Repeated Measures Model in `MLwiN`. Evaluators for Project 16 employed a cross-classified MLM to estimate the effect size, because the number of secondary schools is so small that they had to choose secondary school classes and primary schools to model the random effects. Project 38 has schools randomly assigned to treatment arms and the analytic model is OLS. However, the evaluators adjusted the standard errors using the Huber-White correction.

2.2 Study designs

Depending on treatment assignment mechanism, the 17 projects can be grouped or re-grouped into three categories, namely, Simple Randomised Trial (SRT), Multi-Site Trial (MST), or Cluster Randomised Trial (CRT). SRT, by definition, is the simplest form of randomisation, where individuals are randomly assigned to intervention or control group. This design is known for its simplicity, but often generates imbalanced groups when sample size is small and even in large samples, it can produce imbalance in key variables (D. J. Torgerson & Torgerson, 2008, pp. 30-31).

Usually, methodological decisions surrounding RCTs are made as a trade-off between internal and external validities. With individuals randomly assigned to treatment arms within each cluster or block, MST can help achieve the former without sacrificing the latter, because randomisation is performed in multiple sites and has standardised protocols for data collection, management, and analysis. In addition, samples can be more rapidly accrued in MST than in SRT. This enhances the "*timeliness*" of scientific evidence needed for decision-making (Weinberger et al., 2001, p. 628, original emphasis). However, it is worth noting that most EEF trials with pupil-level randomisation involve multiple schools, meaning they can be classified as MSTs, even though they were described as simple RCTs. Therefore, the differences between SRT and MST are not always straightforward

when stratification is not clearly specified in evaluation reports.

In CRT, clusters, such as schools or classes, are randomly allocated to treatment arms. This design has the potential to avoid contamination that could be epidemic in SRT and MST, but individuals within the same clusters are usually correlated, which violates the independence assumption of standard statistical methods. Also, the number of clusters is usually small, but cluster sizes are often large. To achieve balance between treatment arms, strategies such as matching or stratification in randomisation are often necessary (Hayes & Moulton, 2009, p. 5).

2.3 Analytic models for the archive analysis

While evaluators sought to construct optimal models for their unique evaluation projects, we are devising an approach that facilitates assessment of impact across all the studies. To balance the two purposes, we only focus on evaluators' primary models for the analyses of raw primary outcomes. When a study has two primary outcomes, we report both.

Regarding covariates, pre-test is envisioned as important in the design of EEF projects as their main focus is educational improvement, so we include in the models as our predictors only treatment indicator and pre-test scores. Also, the inclusion of pre-test scores is likely to have "substantial positive" impact on the estimated precision of point estimates (Rubin, 2008a, p. 1352).

In the original evaluations, one team used gain scores as their dependent variable for most of their projects, including where pre and post-tests are not comparable. While there is no straightforward answer to whether one should subtract pre-test from post-test or use pre-test as a covariate (Simmons, Nelson, & Simonsohn, 2011, p. 1363) in RCTs, we believe an extra layer of uncertainty might be introduced when a different distribution is imposed by transforming test results which are not directly comparable into z-scores so the difference can be found. Paterson and Goldstein (1991) cautioned against the transformation of data, as they contended, "slight perturbations to the data or to the model can produce markedly different results" (p. 389). We also hold the view that analyses involving transformed data (e.g., aggregation or standardisation) for causal inference "are remote from the social and educational processes that are of interest" (Paterson & Goldstein, 1991, p. 389). For instance, it would be more difficult for the lay public to understand derivatives, namely, change, gain, or loss, than it is for them to comprehend post-test scores (for a similar argument, see Wainer, 2009, p. 33). Given that all other teams used post-test results as their dependent variable and pre-test scores as a covariate, we used the post-ANCOVA model even when gains were used in a few of the evaluations.

Having specified the outcome and control variables, we applied some of the common analytic models used by EEF evaluators to each of the 17 datasets. These models are difference-in-means of post-test results only (δ), classical linear (OLS), and multi-level models (MLM). For MLMs, we use only within and total variances, because effect sizes based on between-cluster variance are much larger than those derived from the other two sources of variation (Hedges, 2007, p. 345). Since

there are concerns about the validity of standard errors, as reflected in the four missing confidence intervals from an evaluation team in Figures 2 and 3, we also compare, using the equations in Hedges (2007), results from frequentist MLM based on standard errors to their Bayesian counterparts with vague priors, which are the same, thus comparable, across the studies examined. In the next section, we detail the equation or logic for each of the above-mentioned models.

2.3.1 Difference-in-Means (δ)

To calculate Hedges' g for the difference-in-means model, which does not control for any pre-test imbalance in the archive analysis, we first find the raw mean difference $\bar{Y}_t - \bar{Y}_c$, where $\bar{Y}_t$ and $\bar{Y}_c$ are the sample means of post-test scores for the intervention and control groups. We let S_t and S_c be the sample standard deviations, and n_t and n_c their sample sizes. We assume the two populations share a common variance and calculate the pooled variance as:

$$S^2_{pooled} = \frac{(n_t - 1)S_t^2 + (n_c - 1)S_c^2}{n_t + n_c - 2}.$$

Knowing the above information allows us to calculate Cohen's d,

$$d = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{S^2_{pooled}}}, \tag{1}$$

and its variance,

$$V_d = \frac{n_t + n_c}{n_t n_c} + \frac{d^2}{2(n_t + n_c)}. \tag{2}$$

We then convert Cohen's d into Hedges' g using the following corrector (see Borenstein, 2009, p. 226):

$$J(df) = 1 - \frac{3}{4df - 1},$$

where df is the degrees of freedom used to estimate the pooled standard deviation, so that

$$g = J(df)d, \tag{3}$$

and its variance

$$V_g = [J(df)]^2 V_d. \tag{4}$$

Note that since the multiplicative correction factor *J(df)* approaches one as sample size increases, Hedges' g, Cohen's d, and their variances converge for moderate to large sample sizes. This is supported empirically in Figure 1. As such, we report g only from all the models considered and for all the studies included.

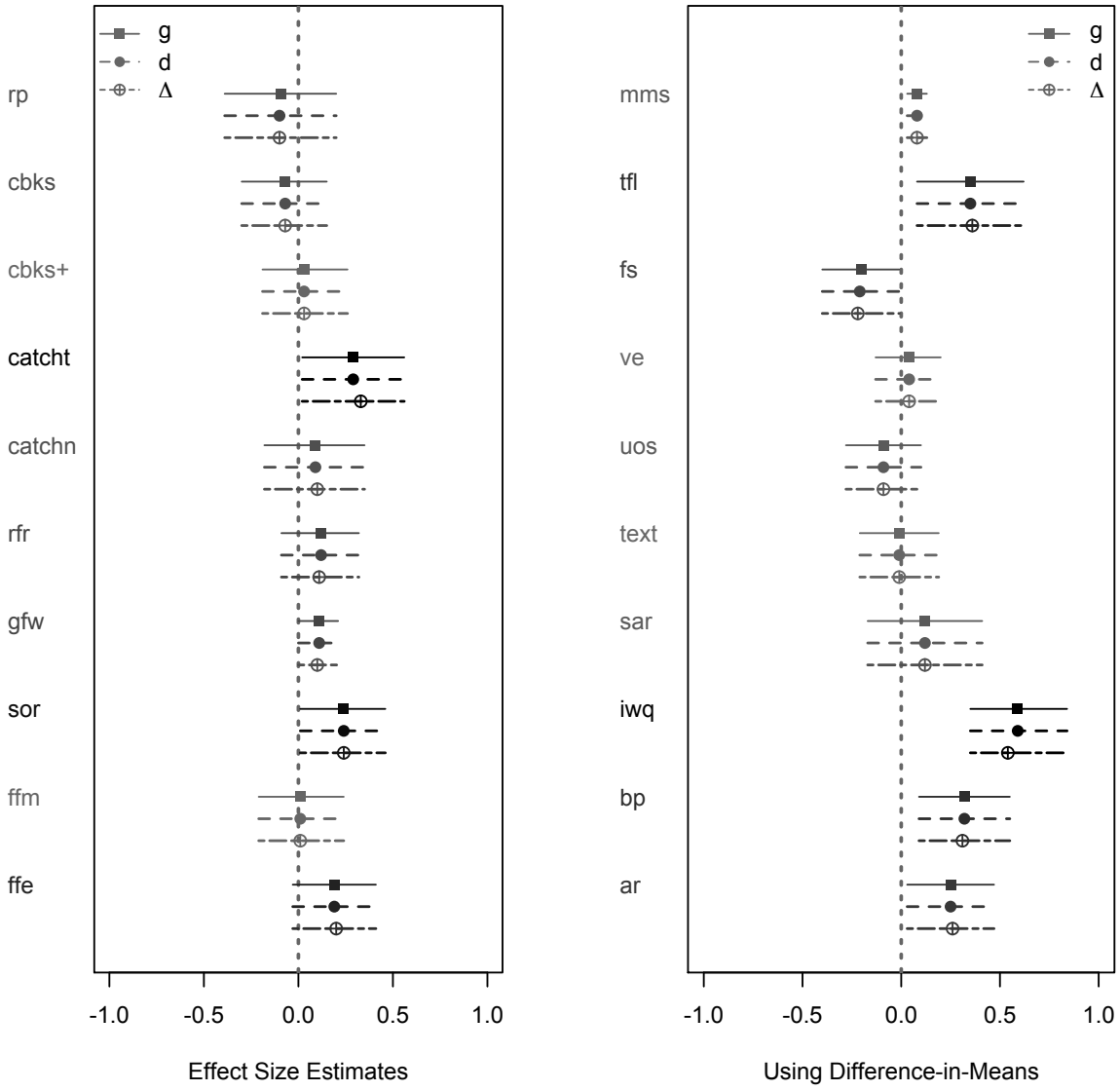

Figure 1: Effect size estimates in Hedges' g, Cohen's d, and Glass' Δ. As shown in the plot, there is almost no difference in the results from the three metrics. Note the larger the effect estimates, the darker the colors.

2.3.2 Ordinary Least Squares (OLS)

Like δ, the classical linear regression is appropriate when clustering of data is absent and sample sizes per school are equal. Unlike δ, the OLS used in this study controls for pre-test scores. The model's functional form is rather straightforward, as reflected in the equation below (Ames, 2013,

p. 597):

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 pret_i + \epsilon_i,$$

where Y_i is post-test results for individual pupils. T_i is treatment indicator with a value of 0 for control or 1 for intervention in a two-arm trial, and $pret_i$ is pre-test scores. $\epsilon_i \sim N(0, \sigma^2)$ is the residual component of the model. The estimated effect size in Cohen's d for the OLS model is calculated as below:

$$\frac{\beta_1}{\sqrt{\sigma^2}}. \tag{5}$$

The equations used to convert d to g and to calculate the variance remain the same as those defined for δ.

2.3.3 Frequentist Multi-Level Modelling (MLM)

As in the OLS defined above, both frequentist and Bayesian MLMs in the study account for pre-test imbalance. But when average pupil attainment varies considerably from school to school or there are unequal samples per school, MLMs are necessary to estimate the weighted average across schools as well as the variances that occur across pupils (level 1) and schools (level 2) (Peugh, 2010, pp. 88-89). At level one, the frequentist MLM is built as:

$$Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + \beta_{2j} pret_{ij} + r_{ij},$$

the continuous outcome variable Y_{ij} represents post-test result of student i in school j, where $j = 1, 2, \cdots, M$ and $i = 1, 2, \cdots, n_j$. M is the number of schools, n_j is the number of pupils per school, and $r_{ij} \sim N(0, \sigma^2)$ captures individual pupil differences in post-test results around their school means. T_{ij} and $pret_{ij}$ are the reported treatment status and pre-test score for student i in school j.

At level two, the model is constructed as:

$$\beta_{0j} = \gamma_{00} + u_{0j},\ \beta_{1j} = \gamma_{10},\ \text{and}\ \beta_{2j} = \gamma_{20},$$

where school average, β_{0j}, is a function of the grand-mean in post-test, γ_{00}, plus school-level residuals around that grand-mean, which are modelled as $u_{0j} \sim N(0, \sigma^2_{Sch})$. The three level two equations assume average attainment varies across schools, but consider average treatment and pre-test effects constant or fixed at γ_{10} and γ_{20} respectively, hence the absence of u_{1j} and u_{2j}.

Substituting level two equations into level one results in a MLM that simultaneously estimates the weighted effect of intervention across schools as well as the different sources of variability. The combined model is

$$Y_{ij} = \gamma_{00} + \gamma_{10} T_{ij} + \gamma_{20} pret_{ij} + r_{ij} + u_{0j}. \tag{6}$$

Where γ_{10} is the estimated mean difference in intervention effect between the intervention and control schools in a CRT, T_{ij} is a dummy variable coded as 1 for pupils in the intervention schools and 0 for pupils in the control schools, and $u_{0j} \sim N(0, \sigma^2_{Sch})$ captures the variation between schools. r_{ij} and u_{0j} are assumed to be independent and respectively capture within and between school variances.

Effect size estimation under the MLM depends on which source of variation is used. Using within-school variance, it can be estimated as:

$$d_W = \frac{\gamma_{10}}{\sqrt{\sigma^2}}.$$

However, effect size based on within-school variance may be inflated if there is substantial heterogeneity between schools. Assuming equal sample size per school, effect size and Intra-Cluster Correlation (ICC) based on total variance can be calculated as:

$$d_T = \frac{\gamma_{10}}{\sqrt{\sigma^2+\sigma^2_{Sch}}} \qquad \text{and} \qquad ICC = \frac{\sigma^2_{Sch}}{\sigma^2+\sigma^2_{Sch}}.$$

ICC measures how much pupil attainment variation occurs at school level. In other words, ICC is the proportion of student attainment variation that can be explained by school differences in their average attainment scores (Peugh, 2010, p. 89). Unlike in CRT, MST has an additional source of variability due to intervention by school interaction. Its effect size and ICC can be calculated as:

$$d_T = \frac{\gamma_{10}}{\sqrt{\sigma^2+\sigma^2_{Sch}+\sigma^2_{Sch*Trt}}} \qquad \text{and} \qquad ICC = \frac{\sigma^2_{Sch}+\sigma^2_{Sch*Trt}}{\sigma^2+\sigma^2_{Sch}+\sigma^2_{Sch*Trt}}.$$

Where $\sigma^2_{Sch*Trt}$ captures the variability of the intervention effects across schools. Hedges (2007, p. 350) provides a detailed discussion of frequentist effect size calculation for multi-level models, particularly how to calculate the associated standard errors.

2.3.4 Bayesian Multi-Level Modelling

Unlike the frequentist MLM, where the parameters of interest are treated as fixed unknowns, the Bayesian MLM regards them as random variables. To estimate them, we use vague priors and employ an approach as described in Wang, Rutledge, and Gianola (1993). To differentiate the frequentist and Bayesian MLMs, we use slightly different notations to represent the fixed effects of $\{\beta_0, \beta_1, \beta_2\}$ and the random intercept b_j.

First, a Scale-Inverse-χ^2 distribution, $p(\sigma^2) \propto (\sigma^2)^{-1}$, was specified as prior to compute within-school or residual variance, which has the following full conditional posterior distribution:

$$p(\sigma^2|\beta, \sigma^2_{Sch}, b, y) \propto \text{Scale-Inverse-}\chi^2(N, S^2_e),$$

where

$$S_e^2 = \frac{\sum_{j=1}^{M}\sum_{i=1}^{n_j}(y_{ij}-\beta_0-\beta_1 T_{ij}-\beta_2 pret_{ij}-b_j)}{N} \quad \text{and} \quad N = \sum_{j=1}^{M} n_j.$$

The same distribution, $p(\sigma^2_{Sch}) \propto (\sigma^2_{Sch})^{-1}$, was then used as prior to simulate between-school variance, for which the full conditional posterior distribution is as follows:

$$p(\sigma^2_{Sch}|\beta, \sigma^2, b, y) \propto \text{Scale-Inverse-}\chi^2(M, S^2_{b_j}),$$

where

$$S^2_{b_j} = \frac{\sum_{j=1}^{M} b_j^2}{M},$$

and M is the number of clusters.

The prior for the fixed effects, $\beta = \{\beta_0, \beta_1, \beta_2\}$, is independently and identically distributed as $N(0, 1000)$. The joint posterior distribution of the fixed effects is

$$p(\beta|\sigma^2, \sigma^2_{Sch}, b, y) \propto N(\tilde{\beta}, (X'X)^{-1}\sigma^2),$$

with

$$\tilde{\beta} = (X'X)^{-1}X'(y-b),$$

where $\mathbf{X}$ is the design matrix for the fixed effects.

The prior for the random effect of schools in the MLM is given as $b_j \sim N(0, \sigma^2_{Sch})$. The random effect has the following posterior distribution:

$$p(b_j|\sigma^2, \sigma^2_{Sch}, \beta, b_{j'}, y) \propto N\left(\tilde{b_j}, \left(n_j + \frac{\sigma^2}{\sigma^2_{Sch}}\right)^{-1}\sigma^2\right),$$

with

$$\tilde{b_j} = \left(n_j^2 + \frac{n_j\sigma^2}{\sigma^2_{Sch}}\right)^{-1}\left(\sum_{i=1, j'\neq j}^{n_{j'}} y_{ij'} - \beta_0 - \beta_1 T_{ij'} - \beta_2 pret_{ij'} - b_{j'}\right),$$

where $j' \neq j$ means excluding the corresponding data for school j when the posterior sample for the random effect of school j is drawn.

Using Gibbs sampler, we generate posterior distributions for the parameters by iteratively sampling from their respective full conditional distributions. We also monitor the effect size at each iteration and calculate the posterior point estimate and the credible intervals after discarding the burn-in part (see Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012, for an introduction to the algorithm).

In addition, we introduce the concept of **Minimum Expected Effect Size** (MEES) by calculating the probability for a specified effect size (ϕ) given the the data. Mathematically, MEES can be expressed as:

$$p(\phi|\sigma^2, \sigma^2_{Sch}, \beta, b, y) = \frac{|d_T \geq \phi|}{|d_T|},$$

where $|d_T \geq \phi|$ is the number of posterior samples of d_T that is at least as extreme as a pre-specified effect size (ϕ) and $|d_T|$ is the length of the Markov Chain Monte Carlo (MCMC) simulation (Gelfand & Smith, 1990) after discarding the burn-in part. The Bayesian estimation as described above was implemented as part of the *eefAnalytic* package in `R`.

3 Results

As shown in Table 4, most point estimates of effect size match well across the different models, except where the outcome variable is different (e.g., `fs`), or the ICC is high (e.g., `catcht`), or the number of clusters is small (e.g., `rfr`), and/or there were serious problems in implementation or evaluation such as attrition (e.g., `bp`). Also, the majority of the studies do not show any change of sign in the point estimates, which is an indication of "Type S" error (when the sign changes according to the methods, see Gelman, Hill, & Yajima, 2012, p. 194). In three cases (`ffm`, `cbks+`, and `fs`) where the sign does change, the difference in the first two is too small to be meaningful, and the third occurs mainly due to the choice of outcome, where the evaluators reported an effect size based on gain scores, whereas we selected a post-test comparison, as all other evaluation teams did.

It is worth noting that point estimates from δ are most likely to differ from regression-based estimates, when the research design is CRT (e.g., `iwq`) or MST (e.g., `catchn`) and the number of schools is small (e.g., `tfl`). This is not surprising, because δ ignores the clustering structure of the data and assumes pupils are independent from one another, even when they come from the same schools, which may differ in average attainment or other characteristics. Point estimates based on within-cluster variances of MLMs are either equal to or greater than those derived from total-variances, the latter occurs when trials are either MST or CRT and their ICCs are high (e.g., `catcht`). However, the difference in effect size estimates based on within and total variances of the MLM is smaller than that in those derived from δ and MLM, suggesting model misspecification might lead to greater biases. When ICCs are low, point estimates from OLS and MLMs are almost identical, as in `sor`, `rp`, `ar` and a few others where ICC is zero. The next section looks in more detail at the forest plots of effect estimates by project.

3.1 Forest plot: picturing estimation and its uncertainties

To understand the forest plots, we suggest readers first check for any change in sign of the point estimates before examining any difference in magnitude. Where there is a sign change, it is useful to assess how meaningful that change is. As we can see in Figure 2, there are two sign changes in `ffm` and `cbks+`. However, the differences are too small to warrant any further investigation. What

study	eval	δ	ols	wth	tt	bs.wth	bs.tt	n	n.sch	icc	lock	type
ffe	0.17	0.19	0.18	0.19	0.14	0.19	0.14	310	33	0.44	2	mst
ffm	0.00	0.01	-0.02	-0.05	-0.04	-0.04	-0.04	303	33	0.13	2	mst
sor	0.24	0.24	0.27	0.27	0.27	0.27	0.27	308	19	0.00	3	srt
gfw	0.10	0.11	0.07	0.10	0.09	0.10	0.08	1367	50	0.23	3	crt
rfr	0.03	0.12	0.04	0.04	0.03	0.03	0.03	355	6	0.18	3	srt
catchn	0.21	0.09	0.27	0.31	0.28	0.30	0.28	216	54	0.12	3	mst
catcht	0.27	0.29	0.32	0.40	0.32	0.40	0.32	210	54	0.36	2	mst
cbks+	-0.01	0.03	0.03	0.04	0.04	0.04	0.04	303	12	0.09	3	srt
cbks	-0.14	-0.07	-0.07	-0.08	-0.08	-0.08	-0.07	311	12	0.06	3	srt
rp	-0.05	-0.09	-0.05	-0.05	-0.05	-0.05	-0.05	178	21	0.00	3	srt
ar	0.24	0.25	0.31	0.33	0.32	0.31	0.31	326	4	0.00	3	mst
bp	0.43	0.32	0.48	0.49	0.47	0.48	0.48	302	6	0.00	0	mst
iwq	0.74	0.59	0.68	0.81	0.67	0.81	0.68	265	22	0.29	2	crt
sar	0.13	0.12	0.13	0.14	0.14	0.14	0.14	182	48	0.11	3	srt
text	-0.06	-0.01	-0.06	-0.07	-0.07	-0.07	-0.07	391	54	0.19	3	mst
uos	-0.08	-0.09	-0.02	-0.04	-0.04	-0.04	-0.04	427	33	0.10	1	mst
ve	0.01	0.04	0.08	0.08	0.08	0.08	0.08	570	12	0.00	4	mst
fs	0.24	-0.20	0.08	0.08	0.07	0.07	0.07	419	10	0.01	3	mst
tfl	0.20	0.35	0.24	0.25	0.25	0.24	0.24	213	3	0.00	4	mst
mms	0.06	0.08	0.08	0.08	0.08	0.09	0.08	5830	44	0.07	4	crt

Table 4: Headline effect sizes from evaluation teams (`eval`) and other analytic models employed in this study. Note that `wth` and `bs.wth` are effect size estimates derived from within-cluster variances of the frequentist and Bayesian MLMs. `tt` and `bs.tt` are estimates based on their total variances. `lock` refers to the number of padlocks awarded to a project. For convenience of comparison, we reproduce here `n` and `n.sch`, which represent pupil and school numbers used in the analytic models. `icc` is intra-cluster correlation. `type` indicates the type of experimental design.

is worth noting is the point estimate of 0.09 from δ for `catchn`, which is up to 0.22 lower than the estimates from other models. The point estimates based on total variances of the MLMs are the same, but smaller than those derived from within variances. In `catcht`, which has an ICC as high as 0.36, the within variance based estimates are much higher than those from other models, and their intervals wider than others' too.

Second, we suggest checking if the intervals are of similar width and whether they contain the vertical zero line or not. Again in Figure 2, plots for projects such as `ffe`, `gfw`, and `rfr` have intervals that deviate considerably in length, which signalls that there are greater uncertainties associated with some estimators and that the choice of analytic models matters. Conventionally speaking, when an interval contains zero, it indicates that an intervention is not statistically significant. When some intervals contain zero and others do not in a given study, the inconsistency might lead to different conclusions for different evaluators. So identifying such inconsistencies is another way of reading the plot. As shown in Figure 2, the interval produced by δ in `gfw` is $[0, 0.21]$, whereas all others contain zero. In `catchn`, four intervals contain zero, but three do not. The divergence is

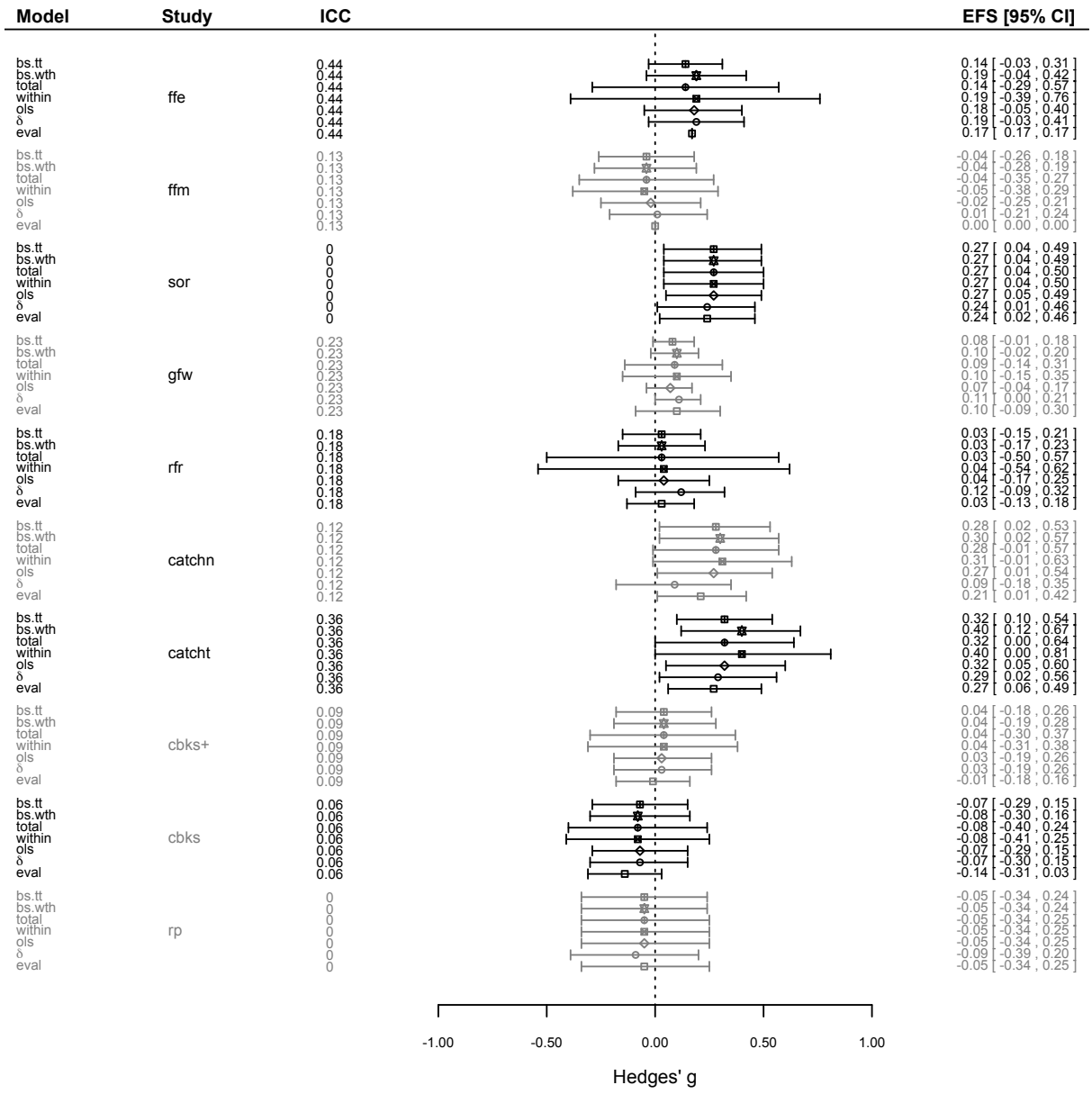

Figure 2: Forest plot of estimated effect sizes and their 95% confidence or credible intervals for the first ten outcomes. Note that, for this plot and others that follow, Where confidence intervals were not reported, the upper and lower bounds are set at the point estimate for comparison.

greater than that in `gfw`, suggesting the choice in analytic models for `catchn` is again important.

In Figure 3, `fs` immediately draws attention, because the point estimate that the evaluators reported is not only different in sign, but also in magnitude. This difference mainly arises from the fact that post-test results are used as the outcome variable in this study, whereas the evaluators

transformed both pre and post-test scores into z-scores first and then calculated gains for the groups to estimate the effect. However, the evaluators did examine post-test results and reported in the text of their report, an estimated effect size of -0.19, which is similar to the one from δ for the archive analysis. Other analytic models in this study use pre-test scores as a covariate, and they produce almost identical results at 0.08. The discordance in results again demonstrates that both the choice of outcome variable and the selection of covariates make important differences in point estimates as well as estimation uncertainties.

For `ar` in Figure 3, the frequentist MLM produces intervals that are not consistent with those from other models. Project `bp` shows much smaller point estimates from δ than those from other models. The ICC is 0, point estimates are similar for all models other than δ, but their intervals differ remarkably. However, the intervention and evaluation encountered some serious problems such as attrition and was not regarded as a secure finding.

Also worth noting in the plot is **mms**, which has a sample of 5830 pupils in 44 schools. The ICC is low and the evaluation was regarded as secure (awarded four out of five padlocks). It turns out that all models produce similar point estimates and quite narrow intervals. Since the study is a CRT, the results also show that models that do not reflect the research design, namely, δ and OLS, would be prone to false positives. Estimate from δ in `tfl` also produces a statistically significant result when other models suggest otherwise. The project is a SRT and has an ICC of zero, the results from δ and OLS should not be of serious concern, particularly when the number of schools is three. The statistically significant result from δ thus shows that failing to control for pre-test scores has an impact on the point estimate of effect which jeopardises the comparability of the effect sizes computed from the models under examination.

The combined information in Figures 2 and 3 illustrates that the most conservative, or the the widest, intervals are frequentist MLM's within and total variance bands, particularly when ICCs are high and the number of clusters is small. Bayesian MLM produces the shortest of all intervals. When the intervals from δ and OLS are short, we might attribute it to their inability to take into account between-school variances, which over-inflates the level of certainty associated with them. However, when interval bounds from MLMs are narrow, they increase our confidence in estimation, because the models have accounted for all sources of variation and controlled for exactly the same covariates.

3.2 Minimum expected effect sizes based on three sources of variation

Figure 4 presents the estimated probabilities of observing effect sizes that are at least as extreme as the specified ones on the x-axis. The calculations are based on Bayesian MLM using three sources of variation. As we can see, effect sizes based on between-cluster variances are usually greater than those derived from within and total variances. When ICCs are zero, which means the schools do not account for any variation in outcomes, it is not meaningful to talk about between-variance

based effect sizes, so the lines that represent them reduce to a flat one. When ICCs are non-zero, the greater the ICCs, the more they differentiate within and total-variance based effect sizes. It is therefore important to choose which effect size to report when between-school variances are large. As demonstrated in the plots, total-variance based effect sizes are the most conservative, and those based on within-variances either converge with the total ones or stay between the total and between curves. It seems that, when ICC is below 0.2, there is not much difference in within and total-variance based effect size estimations. However, when ICC is bigger than 0.2, the difference becomes more observable and important to consider.

4 Discussion and Conclusion

As we have shown, point estimates of effect are similar from across the models in well-powered studies without serious implementation problems. That is to say, if randomisation is well performed and sample size large enough, point estimates tend to remain stable across the models. This is because the distributions of a multitude of covariates, either observed or not, are usually, but not guaranteed, to be similar. However, the standard errors would vary from model to model, and become smaller and smaller as more and more covariates are adjusted for (Rubin, 2007, p. 32). This is reflected in the shorter confidence intervals we often see in evaluation reports, because evaluators normally control for other covariates than pre-test scores and treatment indication. Nevertheless, when ICC is high, sample size is small, and serious attrition is reported, both point estimates and the precision of estimation can vary when different analytic models are used.

Although covariates are important for improved estimates of precision (Rubin, 2008a, p. 1353), adding them to analytic models would reduce some of the total variances "to be explained" (Nakagawa & Cuthill, 2007, p. 597) in well-implemented experiments. So for the comparability of EEF trials, we think future evaluations should only use the pre-test scores and the group status as predictors, except where there are clear reasons otherwise, such as when randomisation is stratified on other covariates than pre-test when these should also be included in the analysis. Of critical importance, we suggest outcome variables be pre-specified and analytic models reflect study designs and the nested structure of educational data, otherwise, we cannot rule out the possibility that some statistically significant effects are misleading and that some point estimates result from "blind luck" (Rubin, 2008b, p. 818).

Since MLMs take into account all sources of variation, outcome estimates from this approach provide more confidence in the estimation of effect. In particular, we think evaluators should use total variance of MLMs to estimate effect sizes, because those derived from between and within variances can be biased, particularly when ICC is high. For higher precision in estimation, Bayesian MLMs with vague priors are recommended.

The models we have tested, except for δ, might be sensitive to non-linearities of the relationship between the outcome variable and covariates. If the differences in intervention group means are too

large, or their variance ratios too great, linear models might be unreliable (Rubin, 2007, pp. 30-31). Given the purpose of the archive analysis, we did not investigate this sensitivity problem. If the analytic models are indeed very sensitive to non-linearity, this limitation would be a good example that, "All models are wrong, but some are useful" (Box, 2013, p. xii). In any case, it implies that, as Rubin put it, "Running regression is no substitute for careful thinking" (2008b, p. 816) in design.

We also argue for pre-specified analysis plans, to prevent data dredging. Without such a plan, it is all too easy for "researcher degrees of freedom" to grow to such a level that "it is unacceptably easy" to produce "statistically significant" results that are "consistent with *any* hypothesis" (Simmons et al., 2011, p. 1359, original emphasis; see also Olken, 2015, p. 67). It is not surprising that many psychology findings could not be replicated in a recent study featured in the journal *Science* (Open Science Collaboration, 2015). In education, similar problems might exist as well, although they are beyond the scope of this investigation.

Even when analysis plans are pre-specified, the data that arrives at an analyst's computer may not be the raw data, and the persons who pre-process the data are not always the ones who analyse it. Consequently, the model an evaluator eventually chooses, however statistically sound, has already been simplified (as Meng described in Lin et al., 2014, p. 545). In the EEF-funded projects, the samples we have access to in the archive sometimes vary from the ones used and reported by the evaluators for their primary analyses (see Table 2). In some projects, the difference was so great that archive analysis as described above would not have been possible without the support from original evaluation teams.

Diversity in analytic models of educational trials poses challenges at two levels at least. First to the EEF which seeks to assess the impact of the interventions they have funded and decide which ones are worth scaling up or investigating further. If estimates of impact from different models are not comparable, it is hard to tell what has really worked and what might be effective at scale to address the needs of disadvantaged children in primary and secondary schools in England. Second to the aggregation of impact in meta-analysis, which depends on the comparability of the effect sizes which are synthesised. This paper shows that there are limits to the precision which is achievable in a pooled estimate which may be related to the analytic choices of the included studies. In particular, the clustered nature of educational data is not always taken into account. This may both inflate estimates, and imply greater precision than is warranted.

Acknowledgement

This research was funded by a grant to Durham University from the Education Endowment Foundation. We would also like to acknowledge our thanks to Gary King for his comment and Larry Hedges, Tom Crook, and Kathy Sylva who provided feedback on a poster version of this work, and to a number of the evaluators who were generous with their time in answering questions about the data deposited in the EEF archive.

References

Ames, A. J. (2013). Accuracy and Precision of an Effect Size and Its Variance From a Multilevel Model for Cluster Randomized Trials: A Simulation Study. *Multivariate Behavioral Research*, *48*(4), 592–618. doi: 10.1080/00273171.2013.802978

Borenstein, M. (2009). Effect Sizes for Continuous Data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–235). New York: Russell Sage Foundation.

Box, G. E. P. (2013). *An Accidental Statistician: The Life and Memories of George E.P. Box.* Hoboken, New Jersey: Wiley. doi: 10.1002/9781118514948

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. doi: 10.1080/19345747.2011.618213

Gorard, S., See, B. H., & Siddiqui, N. (2014). *Switch-on Reading: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Gorard, S., Siddiqui, N., & See, B. H. (2014). *Future Foundations: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Gorard, S., Siddiqui, N., & See, B. H. (2015a). *Accelerated Reader: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Gorard, S., Siddiqui, N., & See, B. H. (2015b). *Fresh start: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Hayes, R., & Moulton, L. (2009). *Cluster Randomised Trials.* London: Chapman and Hall/CRC.

Hedges, L. V. (2007). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370.

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*(3), 167–171. doi: 10.1111/j.1750-8606.2008.00060.x

Humphreys, M., Sanchez de la Sierra, R., & van der Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, *21*(1), 1–20. doi: 10.1093/pan/mps021

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

King, B., & Kasim, A. (2015). *Rapid Phonics: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., & Wang, J.-L. (2014). *Past, Present, and Future of Statistical Science.* London: CRC Press.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Maxwell, B., Connolly, P., Demack, S., O'Hare, L., Stevens, A., & Clague, L. (2014a). *Summer Active Reading Programme: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Maxwell, B., Connolly, P., Demack, S., O'Hare, L., Stevens, A., & Clague, L. (2014b). *TextNow Transition Programme: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Merrell, C., & Kasim, A. (2015). *Butterfly Phonics: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*(4), 591–605. doi: 10.1111/j.1469-185X .2007.00027.x

Olken, B. A. (2015). Promises and Perils of Pre-Analysis Plans. *Journal of Economic Perspectives*, *29*(3), 61–80. doi: 10.1257/jep.29.3.61

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/science.aac4716

Paterson, L., & Goldstein, H. (1991). New Statistical Methods for Analysing Social Structures: An Introduction to Multilevel Models. *British Educational Research Journal*, *17*(4), 387–393.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85–112. doi: 10.1016/j.jsp.2009.09.002

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*, 20–36. doi: 10.1002/ sim.2739

Rubin, D. B. (2008a). Comment: The Design and Analysis of Gold Standard Randomized Experiments. *Journal of the American Statistical Association*, *103*(484), 1350–1353. doi: 10.1198/016214508000001011

Rubin, D. B. (2008b). For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics*, *2*(3), 808–840.

Rutt, S. (2014). *Catch Up Numeracy: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Sheard, M., Chambers, B., & Elliott, L. (2015). *Units of Sound: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. doi: 10.1177/0956797611417632

Styles, B., & Bradshaw, S. (2015). *Talk for Literacy: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Styles, B., Clarkson, R., & Fowler, K. (2014a). *Chatterbooks: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Styles, B., Clarkson, R., & Fowler, K. (2014b). *Rhythm for Reading: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

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

Torgerson, D., Torgerson, C., Ainsworth, H., Buckley, H., Heaps, C., Hewitt, C., & Mitchell, N. (2014). *Improving Writing Quality: Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

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

Torgerson, D. J., & Torgerson, C. J. (2008). *Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction.* London: Palgrave Macmillan.

Wainer, H. (2009). *Picturing the Uncertain World: How to Understand, Communicate, and Control Uncertainty through Graphical Display.* Princeton: Princeton University Press.

Wang, C., Rutledge, J., & Gianola, D. (1993). Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics Selection Evolution*, *25*, 41–62.

Weinberger, M., Oddone, E., Henderson, W., Smith, D., Huey, J., Giobbie-Hurder, A., & Feussner, J. (2001). Multisite Randomized Controlled Trials in Health Services Research: Scientific Challenges and Operational Issues. *Medical Care*, *39*(6), 627–634.

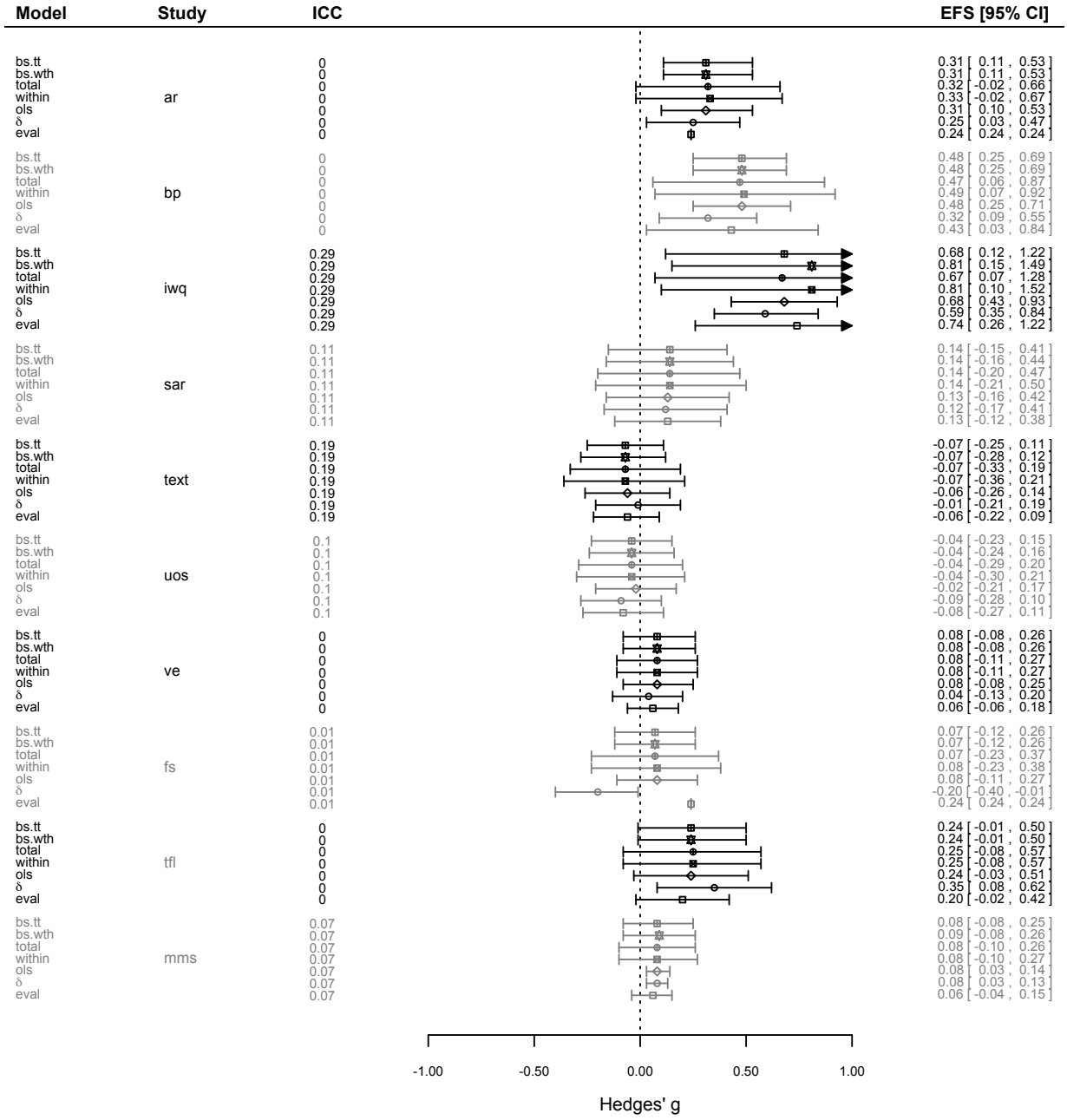

Figure 3: Forest plot of estimated effect sizes and their 95% confidence or credible intervals for the second ten outcomes.

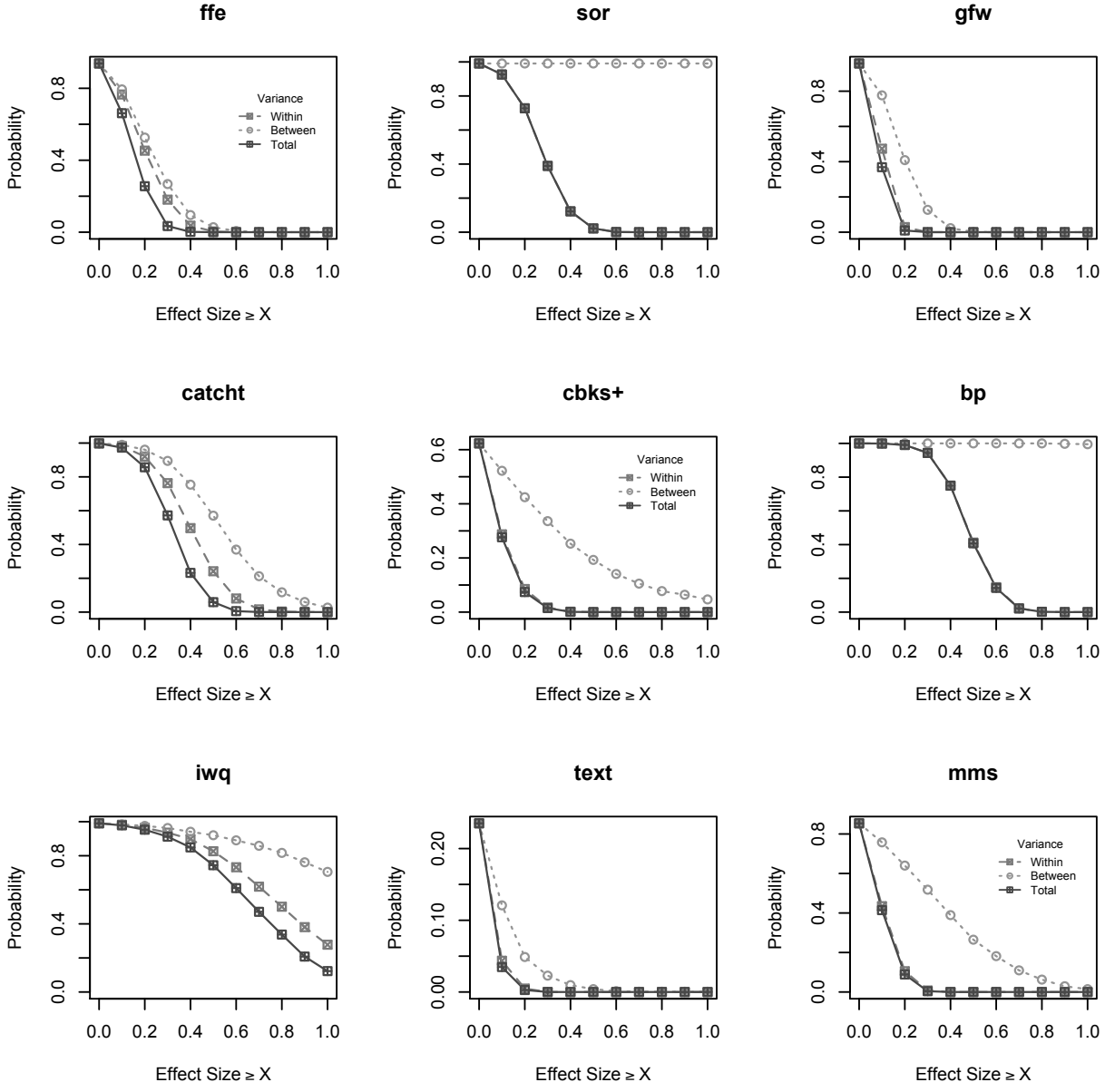

Figure 4: Minimum expected effect sizes based on within, between, and total variances for nine representative projects.