UCL

CENTRE FOR EDUCATION POLICY & EQUALISING OPPORTUNITIES

Working Paper No. 20-16

Quantifying 'promising trials bias' in randomised controlled trials in education

Sam Sims
University College London

Jake Anders
University College London

Matthew Inglis
Loughborough University

Hugues Lortie-Forgues
Loughborough University

Randomized controlled trials have proliferated in education, in part because they provide an unbiased estimator for the causal impact of interventions. It is increasingly recognized that many such trials in education have low power to detect an effect, if indeed there is one. However, it is less well known that low powered trials tend to systematically exaggerate effect sizes among the subset of interventions that show promising results. We conduct a retrospective design analysis to quantify this bias across 23 promising trials, finding that the estimated effect sizes are exaggerated by an average of 52% or more. Promising trials bias can be reduced ex-ante by increasing the power of the trials that are commissioned and guarded against ex-post by including estimates of the exaggeration ratio when reporting trial findings. Our results also suggest that challenges around implementation fidelity are not the only reason that apparently successful interventions often fail to subsequently scale up. Instead, the findings from the initial promising trial may simply have been exaggerated.

VERSION: November 2020

Suggested citation: Sims, S., Anders, J., Inglis, M., & Lortie-Forgues, H. (2020). *Quantifying 'promising trials bias' in randomised controlled trials in education* (CEPEO Working Paper No. 20-16). Centre for Education Policy and Equalising Opportunities, UCL. https://EconPapers.repec.org/RePEc:ucl:cepeow:20-16.

Disclaimer

Any opinions expressed here are those of the author(s) and not those of the UCL Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

CEPEO Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Highlights

- Randomised controlled trials are widely used in education research to quantify the impact of programmes and interventions. ‘Warehouses’ and ‘toolkits’ summarising promising approaches are based on findings from such trials.
- It is increasingly recognized that many trials in education have low power to detect an effect. However, it is less well known that low powered trials tend to systematically exaggerate effect sizes among the subset of interventions that show promising results.
- Across 23 trials deemed to have shown promising findings, we estimate that effect sizes are exaggerated by an average of 52%, or more.
- Trial commissioners can mitigate this problem by increasing the power of the trials that are commissioned. Analysts can guard against this problem by including estimates of the exaggeration ratio when reporting trial findings.

Why does this matter?

Randomised controlled trials are one important way in which we attempt to discover ‘what works’ in education. Ensuring that such trials provide reliable findings is therefore central to the drive for evidence-based education.

Quantifying 'promising trials bias' in randomized controlled trials in education

Sam Sims[1]
Jake Anders[1]
Matthew Inglis[2]
Hugues Lortie-Forgues[2]

November 2020

Randomized controlled trials have proliferated in education, in part because they provide an unbiased estimator for the causal impact of interventions. It is increasingly recognized that many such trials in education have low power to detect an effect, if indeed there is one. However, it is less well known that low powered trials tend to systematically exaggerate effect sizes among the subset of interventions that show promising results. We conduct a retrospective design analysis to quantify this bias across 23 promising trials, finding that the estimated effect sizes are exaggerated by an average of 52% or more. Promising trials bias can be reduced ex-ante by increasing the power of the trials that are commissioned and guarded against ex-post by including estimates of the exaggeration ratio when reporting trial findings. Our results also suggest that challenges around implementation fidelity are not the only reason that apparently successful interventions often fail to subsequently scale up. Instead, the findings from the initial promising trial may simply have been exaggerated.

[1] Centre for Education Policy and Equalising Opportunities, UCL Institute of Education, 20 Bedford Way, London, UK, WC1H 0AL.
[2] Centre for Mathematical Cognition, Loughborough University, Epinal Way, Loughborough, UK, LE11 3TU.

1. Introduction

Randomized controlled trials (RCTs) have proliferated in education research in recent years (Connolly, Keenan, & Urbanska, 2018). Funders are attracted to descriptions of a 'gold standard' research design (Pocock, 1978), policy makers have emphasized the importance of subjecting educators' views to rigorous tests (e.g. Haynes, Service, Goldacre, & Torgerson, 2012), and researchers are drawn to an unbiased estimator for the causal impact of interventions (e.g. Torgerson & Torgerson, 2001). In England, this approach has been institutionalized through the creation of the publicly-funded Education Endowment Foundation (EEF), which has now completed over 100 RCTs (Dawson, Yeomans, & Brown, 2018). In the US, the Every Student Succeeds Act privileges RCTs as providing 'Tier I' evidence and the Institute of Education Science has now commissioned over 350 RCTs (Hedges & Schauer, 2018).

Despite the increasing popularity of RCTs in education - and the gradual accumulation of both null and positive results - an awareness of their limitations is growing. For example, researchers have shown that many education trials have statistical power well below 80% to detect effects of the magnitudes commonly found in the education literature (Cheung & Slavin, 2016; Spybrook, Shi, & Kelcey, 2016; Torgerson, Torgerson, Birks & Porthouse, 2005). This leads to trials that neither provide good evidence in support of an effect, nor evidence that there is no effect. Indeed, Lortie-Forgues & Inglis (2019) estimate that 40% of education RCTs are uninformative with respect to impact.

We address a related limitation of RCTs in education. Rather than focusing on when trials do not provide any evidence one way or another, we focus instead on when trials conclude that there is an effect of a given magnitude, but this finding is erroneous. More specifically, we ask: when a researcher concludes from an education RCT that an intervention does indeed have an effect, to what extent is the estimated effect an exaggeration of the true impact (Type-M, or magnitude error)? And how likely is the true effect to be in the other (negative) direction (Type-S, or sign error)? (Gelman & Carlin, 2014). Despite RCTs providing an unbiased estimator in general, we find that estimates from

education RCTs deemed to show promising results display surprisingly high levels of Type M error. By contrast, we find that Type S errors are unlikely in such trials.

It should be noted that the Type M error we identify is not the familiar random (mean zero) error that is inherent to all RCTs. Rather, this error constitutes systematic upward bias of effect sizes among trials that find statistically significant results. We therefore refer to it as *promising trials bias*. Given that this results from conditioning the set of trials under consideration on their p values, it can be thought of as the study-level analogue of survivor bias - when differential dropout of observations renders a sample unrepresentative of the population of interest. Analogously, while in general RCTs have mean zero error, the set of trials that identify statistically significant results are no longer representative of RCTs in general, and tend to overestimate effect sizes. This leads to 'warehouses' or 'toolkits' listing promising education interventions to contain inflated effect size estimates and means that promising trials are less likely to replicate. Indeed, we estimate that some of the interventions listed in such warehouses have a non-trivial probability of having a true effect size of zero.

This research is closely related to several strands of the recent academic literature. In particular, it builds on work by Gelman and Tuerlinckx (2000), Gelman and Carlin (2014) and Colquhoun (2019) concerned with the lack of information often conveyed by RCTs, even when $p<0.05$. We apply this thinking to the education setting, using empirical data to quantify the extent of this problem for the field. In that sense, our research is analogous to work by Button et al. (2013) in neuroscience. By estimating the extent to which an identical study of an apparently promising intervention would produce a different result, our analysis is also relevant to recent work on the relationship between *p* values and replicability of research (Anderson, 2020; Colquhoun, 2017; Makel & Plucker, 2014; Vasishtha, Mertzena, Jäger, & Gelman, 2018). Our suggestions for using external information (or priors) to interpret the results from education trials are also related to those recently made by Deke and Finucane (2019).

In Section 2 of the paper, we begin by setting out a conceptual framework for understanding erroneous findings in RCTs, grounded in potential outcomes notation (Rubin, 1974) and an adapted version of the diagrams developed by Gelman & Tuerlinckx (2000). Then in Section 3, we set out our

empirical approach to estimating the probability and severity of erroneous findings, paying particular attention to our use and selection of priors. In Section 4, we report our estimates of Type M and Type S errors in 23 'promising' RCTs (deemed to have provided evidence of efficacy) and in Section 5 we extend the analysis to the special case of Type M error in which the true effect is zero i.e. a false positive (Colquhoun, 2019). Finally, in Section 6, we conclude with a discussion of the implications of our findings for the design, analysis, interpretation, and commissioning of RCTs in education research.

2. Conceptualising erroneous findings in trials

2.1 Erroneous findings and potential outcomes

To bring clarity to the idea of erroneous findings, it is helpful to briefly restate the basis on which we would expect to obtain unbiased causal estimates of impact from randomized controlled trials. Individuals (i = 1, …, N) may be exposed to a policy or programme ($T_i = 1$) and have two potential outcomes, one which would be observed if they were exposed Y_{1i} to the treatment and one if they were not exposed to the treatment Y_{0i}. To quantify the effect of the intervention, we would like to calculate the average treatment effect, *D*:

$$D = \frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i}) \quad (1)$$

However, we can never observe both Y_{1i} and Y_{0i} for a given individual, i (Holland, 1986). Nevertheless, with random assignment of individuals to treatment and control in an RCT, the law of large numbers ensures that individuals assigned to the treatment group and individuals assigned to the control group have identical potential outcomes in expectation:

$$E(Y_{0i}|T_i = 1) = E(Y_{0i}|T_i = 0) \quad (2)$$

$$E(Y_{1i}|T_i = 0) = E(Y_{1i}|T_i = 1) \quad (3)$$

In each case, the terms on the left-hand side containing counterfactuals are now equivalent to the terms on right hand side containing observable quantities. This allows us to re-express our estimand in

(1) without the use of counterfactuals. We emphasise this henceforth by using the Y^{obs} term to denote the observed outcomes in a particular group:

$$D = E\left({Y^{obs}}_i | T_i = 1\right) - E\left({Y^{obs}}_i | T_i = 0\right) \quad (4)$$

It should be noted, however, that while 2) and 3) are true in expectation (across many hypothetical repetitions of the random allocation process), once random allocation has occurred in a given trial we observe the mean of Y in our data, rather than it's expectation. In practice, we therefore employ the following estimator:

$$\widehat{D} = \left(\overline{Y^{obs}} | T_i = 1\right) - \left(\overline{Y^{obs}} | T_i = 0\right) \quad (5)$$

where $\bar{Y}$ is the mean of Y_i for all is. However, while $E(\widehat{D}) = D$ across many trials, due to sampling variation, in a given trial it is very likely that:

$$E(Y_{0i} | T_i = 1) \neq \left(\overline{Y^{obs}} | T_i = 0\right) \quad (6)$$

and

$$E(Y_{1i} | T_i = 0) \neq \left(\overline{Y^{obs}} | T_i = 1\right) \quad (7)$$

In which case:

$$\frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i}) \neq \left(\overline{Y^{obs}} | T_i = 1\right) - \left(\overline{Y^{obs}} | T_i = 0\right) \quad (8)$$

and, equivalently:

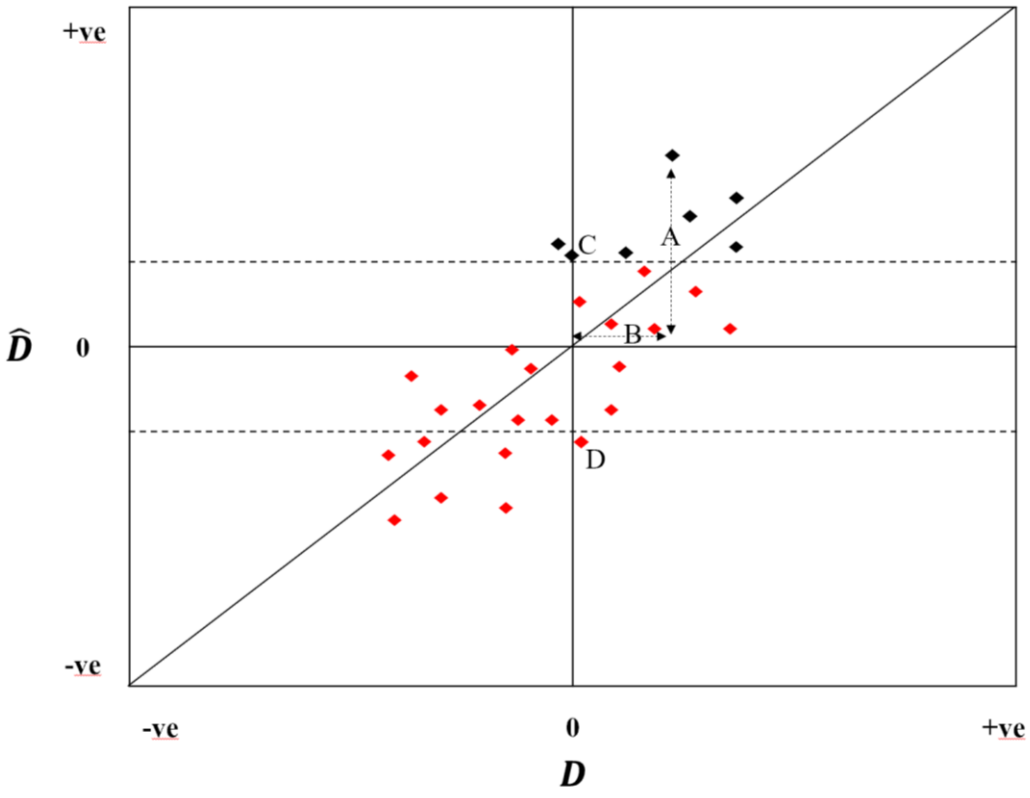
$$D \neq \widehat{D} \quad (9)$$

The inequality between D and $\widehat{D}$ in individual RCTs is the main focus of this paper. More specifically, we study two different types of $D \neq \widehat{D}$. First, Type M [magnitude] error, which can now be defined as cases in which $D \neq \widehat{D}$ and $\widehat{D}$ has a $p < \alpha$. The magnitude of Type M error (also known as the exaggeration ratio) is equivalent to $\widehat{D}/D$. Second, Type S [sign] error, which is present in all cases in which ($\widehat{D} > 0$ & $D < 0$) or ($\widehat{D} < 0$ & $D > 0$) and $\widehat{D}$ has $p < \alpha$. In an extension to our main analysis, we also look at the probability of false positives, which can be seen as a special case of Type M error in which $D = 0$.

2.2 Graphical representation of erroneous findings

The distinction between these types of error can be further clarified visually by extending the diagram originally used by Gelman & Tuerlinckx (2000) to illustrate Type S error. Figure 1 shows a hypothetical, purely illustrative plot of D and $\widehat{D}$ across many perfectly implemented RCTs of different interventions which, for simplicity, have been assumed to have identical standard errors. Points within the dotted lines indicate estimates that are not significantly different from zero, points outside the dotted lines indicate estimates that are. The trials in black might be considered to be 'promising' in that they have $\widehat{D} > 0$ and $p < \alpha$. By construction, all of the trials represented by black dots display some degree of Type M error, since all of them have $p < \alpha$ and none of them fall on the 45-degree line on which $D = \widehat{D}$. The trial with the highest value of $\widehat{D}$ shows Type M error of magnitude A/B, since $D = B$, $\widehat{D} = A$ and $p < \alpha$. The trial labelled D is a case of Type S error in that $D > 0$ (x axis) but $\widehat{D} < 0$ (y axis) and $p < \alpha$. The trial labelled C is a false positive, since $D = 0$, $\widehat{D} > 0$ and $p < \alpha$.

Figure 1: Illustrating Type M, false positive and Type S errors in trials



2.3 Erroneous findings, statistical power and the statistical significance filter

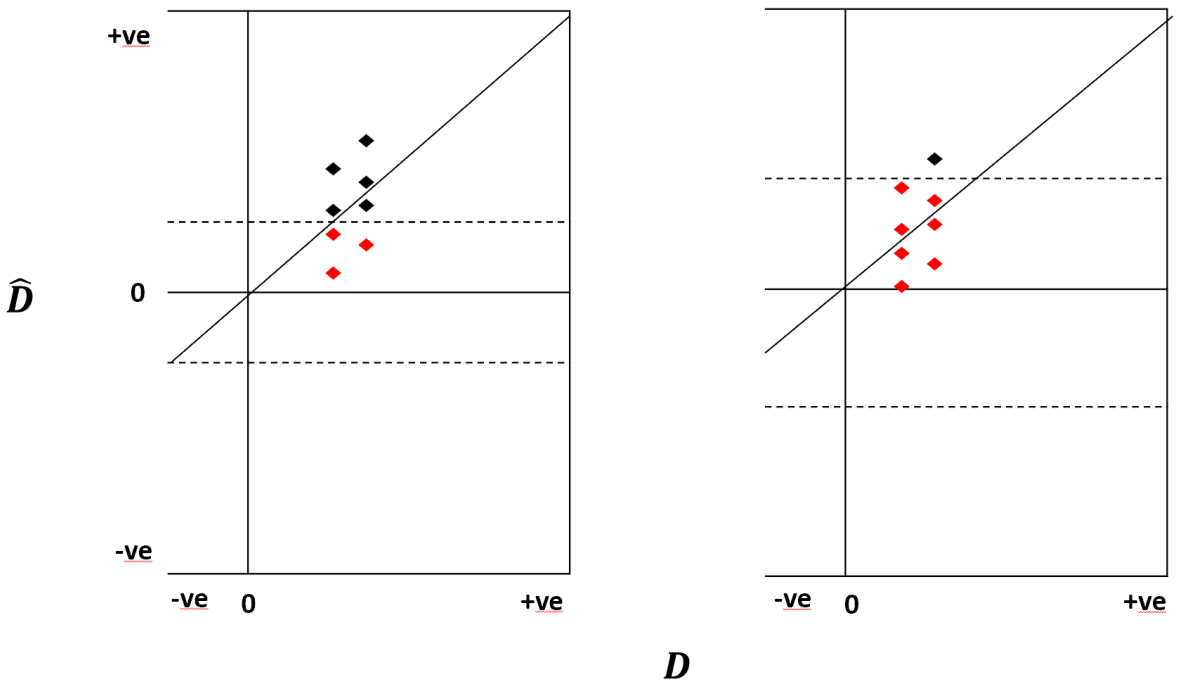
The prior subsections illustrate how impact estimates from specific RCTs will in general diverge from the true effect, given finite sample size. This is not particularly problematic in and of itself since these errors have a mean of zero across trials. Statistical inference therefore helps to account for the random error inherent to trials. However, our focus in this paper is on a specific subset of RCTs: those that have been deemed to show evidence of impact based on $p < \alpha$. Amongst this group of trials, we would expect Type M errors to systematically inflate effect sizes, particularly when those studies have lower power (Ioannidis, 2008; Gelman & Carlin, 2014). To see why, consider that when $\alpha = 0.05$ an estimate must be 1.96 standard errors away from zero to be declared a discovery, and the most exaggerated estimates are systematically more likely to clear this threshold. In effect, the requirement for $p < \alpha$ means that, the less exaggerated a result is, the more likely it is to be filtered out from being a declared a discovery.

This 'statistical significance filter' can also be seen in Figure 2, which shows the same set of RCT point estimates in both panels, but with estimates in the right-hand panel assumed to come from trials, with lower power and larger standard errors. This results in only one of the five point estimates with $p < 0.05$ in the left-hand panel remaining so in the right-hand panel, and this is the estimate with the largest Type M error. The relationship between power (or sample size) and Type M error can also be verified analytically using Gelman and Carlin's (2014) retrodesign() function. In sum, when trials have lower power, effect size estimates for promising interventions tend to be more inflated – a problem we refer to as *promising trials bias*.

Where decisions about commissioning scale-up or replication trials are based on lower-powered initial trials showing promising results, promising trials bias will create difficulties with replication since the most exaggerated trials are also the least likely to replicate (Button et al., 2013; Vasishth et al., 2018). Consider, for example, the Education Endowment Foundation's (EEF) approach to commissioning. They conduct initial 'efficacy' trials (with relatively small samples, under ideal conditions) and then deem a subset of these interventions to be 'promising' based on the results. Promising interventions are then often tested in subsequent 'effectiveness' trials (with larger samples

and under everyday conditions). Promising trials bias means we would *expect* impact estimates from promising efficacy trials to reduce in size in subsequent effectiveness trials due to a process of mean reversion. In the rest of the paper, we aim to quantify this effect.

Figure 2: Lower powered studies inflate Type M error



3. Method

3.1 Estimating Type-M and Type-S error

In section 2.1 we set out how RCTs provide an unbiased estimator for the causal impact of an intervention but, when we focus on a specific estimate from a single trial, sampling variation means that the estimate will very likely contain error of some magnitude. The retrospective design analysis (Gelman & Carlin, 2014) that we conduct here aims to move back in the other direction (from the specific trial to the general) by asking: for each published RCT, what results would we be likely to obtain under hypothetical replications of the study? Doing so requires us to adopt a probability model, which specifies the assumed probability that different effect sizes would be obtained in a hypothetical replication of the trial (D^{rep}).

The probability model for D^{rep} is defined by three quantities. The first is an assumption about the true effect size. This centres the probability model by determining the most likely effect size to be obtained in a hypothetical replication. This assumption plays a critical role in design analysis and we hence discuss it at some length in section 3.3. The second quantity that defines the probability model for D^{rep} is the standard error of the estimated effect in the original trial. This determines the variability in the effect sizes that would be obtained under hypothetical replications. The third quantity is the statistical significance threshold α, which determines the region of the probability model for D^{rep} that is far enough from zero that it would be declared a discovery (Gelman & Tuerlinckx, 2000; Gelman & Carlin, 2014). If $\alpha = 0.05$, we then estimate the expected magnitude of Type M error by repeatedly simulating draws from this probability model and averaging the differences between D and D^{rep} for all draws of D^{rep} more than 1.96×SE way from zero. We estimate the probability of Type S error by calculating the proportion of the probability model that protrudes more than 1.96×SE in the opposite direction (+/-) to the true effect, D. In practice, we implement this using the R function retrodesign() (Gelman & Carlin, 2014), but note that this can also be done using the Stata module rdesigni (Klein, 2017).

3.2 Sample of RCTs

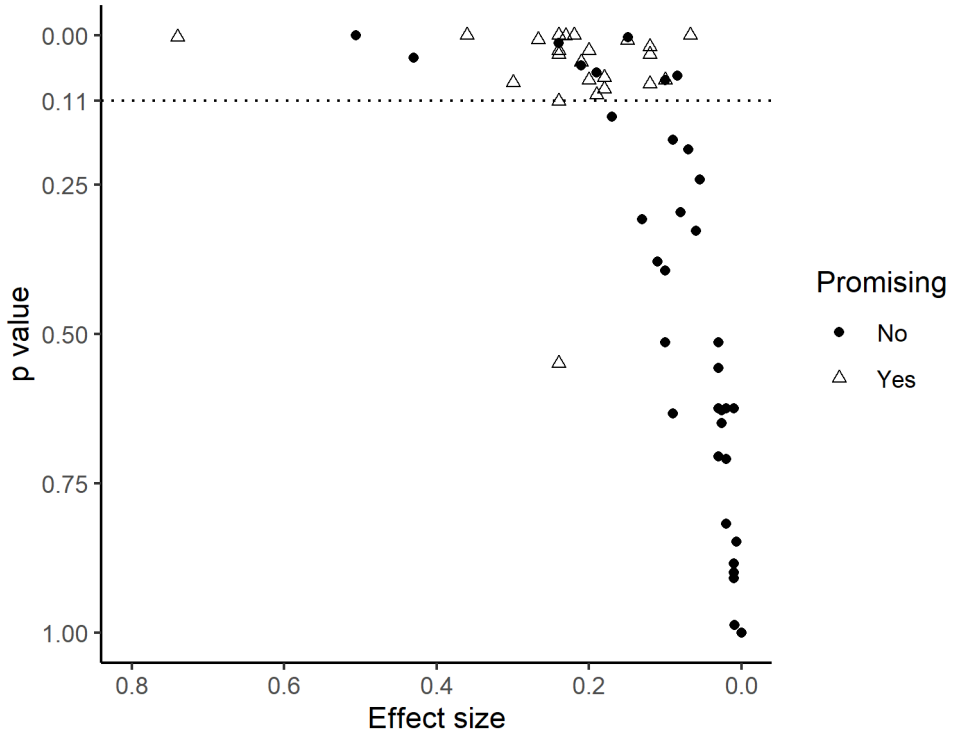
Recall that our motivation is to understand erroneous findings in trials that have been deemed to show promising results. Our empirical analysis therefore employs the largest set of published education trials that are systematically replicated when they show evidence of impact – those commissioned by the Education Endowment Foundation. This constitutes over 100 published trials, 23 of which have been deemed to show 'promising' results by the EEF at some point (marked with an * in the reference list). In cases where trials are deemed to have shown promising results, and where the intervention developer is happy to cooperate on further research, the EEF commissions a further trial of the original intervention. Of the 23 trials in our sample, 8 have so far been replicated and had their results published (marked with an † in the reference list).

The EEF maintain a list of interventions deemed to have shown promising results on their website.[1] Their most recent guidance (EEF, 2020) explicitly rejects the use of $p < 0.05$ as a cut-off for claiming a discovery, stating that they rely instead on a combination of the estimated effect size and a "continuous" (i.e. not dichotomized around 0.05) *p* value (EEF, 2020). Since our estimates of Type M and Type S error are a function of the α level, this poses something of a challenge.

Figure 3 shows a scatter plot of all first-stage EEF trials that yielded a positive effect size estimate, with the *p* value on the x axis (reversed) and the effect size on the y axis. Where trials analyze multiple outcomes and associated *p* values, we use the lowest *p* value. Our justification for this choice is that these are the only result that we know unambiguously to be deemed as 'promising' by the EEF. Triangles are those initial (or 'efficacy') trials for which a replication (or 'effectiveness') trial was subsequently commissioned. It is clear from the graph that there is a sharp increase in the probability of trials being recommissioned when the lowest *p* value is below 0.11 (dotted horizontal line). In addition, when *p* is less than 0.11, the probability of a trial being recommissioned seems unrelated to effect size. We therefore adopt 0.11 as EEF's implicit α in our calculation of Type M and Type S errors. Having said that, we acknowledge that this implicit α is somewhat out of line with the social scientific literature, where $p < 0.05$ remains the dominant convention for rejecting the null and declaring a discovery. Accordingly, in Appendix Table A2, we report a parallel set of results for trials that reported $p < 0.05$ using $\alpha = 0.05$ in our calculations of Type M and Type S error. This set of results is more in line with what might be expected from the academic literature in general.

[1] The list can be found here. Please note that interventions are also removed from this list if subsequent trials of the same intervention to do not find an effect. https://educationendowmentfoundation.org.uk/tools/promising

Figure 3: Scatterplot of EEF first-stage trials



Notes: Vertical axis shows the lowest p value in each EEF trial and horizontal axis shows the effect size associated with that p value. Number of trials = 59. Triangles are trials of interventions deemed to have been 'promising' by EEF at any point.

3.3. Selection and justification of priors

As noted above, calculation of a trial's Type M/S error relies on a *prior* assumption about the true effect size. Gelman & Carlin (2014, p. 642) recommend thinking about the true effect as "that which would be observed in a hypothetical infinitely large sample". This makes sense in that, in a well-implemented RCT, bias approaches zero as the sample grows toward infinity (Imai, King, & Stuart, 2008). While this provides a precise way of conceptualising a prior about the true effect, it leaves open the question of how to determine the value of the prior in each case.

What would constitute an ideal source of evidence for our prior assumptions about the true effect size? First, our prior would ideally be derived from a separate experimental study using standardized tests as outcome measures (similar to those used in EEF trials) or, where available, meta-analysis of

all relevant studies using such methods (Gelman & Carlin, 2014). Second, the external studies should be of a similar intervention, targeting a similar subject, in a similar age/grade (Cheung & Slavin, 2016; Kraft, 2020). Third, the external studies should be free from publication bias (Gage, Cook, & Reichow, 2017). The next three paragraphs describe the three different sources we use for our priors and briefly discuss their strengths and weaknesses, which are also summarized in Table 1.

Our first source of priors is a recent meta-review of 750 education RCTs that used standardized tests as outcome measures (Kraft, 2020). In particular, we use the median effect size for the grade of the pupils enrolled in each of the trials in our sample, for the subject (mathematics or English) targeted by the intervention (see Appendix Table B1). The strength of this source of priors is that it is summarizes information from a very wide range of studies and tailors these to both grade and subject for each of the trials in our sample. The corresponding downside is that it is potentially infected by publication bias.

Our second source of priors is the National Centre for Education Evaluation (NCEE) database of trials. The NCEE is similar to the EEF in that it uses government funding to run RCTs of interventions that have previously shown promise and then publishes the results in a transparent way, regardless of the outcome. From this, we derived our second set of priors based on the average NCEE effect size in the subject area being targeted by each of the trials in our sample. In cases where the EEF trial targeted multiple subjects, we use the outcome associated with the lowest *p* value (for the same reasons given above). An important strength of this source is that it constitutes a highly analogous, out-of-sample set of trial results, which is free from publication bias and can be applied to all of the trials in our sample. However, while this source tailors the prior to the subject domain being targeted by each of the interventions in our sample trials, it does not tailor the prior to the age/grade or the specifics of the intervention. Nevertheless, the trials are all of interventions that have shown promising evidence in previous evaluations, suggesting they will be broadly comparable to those in EEF trials.

Our third set of priors are specific to each EEF intervention and are derived from evaluations of similar interventions published in the academic literature. In order to maximize comparability of

effect sizes, we restrict these to evaluations using experimental designs and employing standardized tests as outcome measures (Gelman & Carlin, 2014; Cheung & Slavin, 2016). Where possible, we relied on the results of meta-analysis using these same criteria to select studies, in order to get the most complete picture of the existing literature. To ensure a thorough search was conducted, two team members independently searched for suitable studies from which to derive priors for all 23 interventions in our sample. Using this approach, we were able to identify appropriate intervention-specific priors for 8 of the 23 trials in our sample (the sources and justification for each of these are set out in Appendix A Table A1). For the other 15 trials, we could not find any suitable study for a similar intervention from which to derive intervention specific priors. The advantage of this final source of priors is that they are tailored directly to each intervention. The disadvantage is that they may be subject to publication bias and are not available for 15 of our 23 trials.

- **Table 1: Comparing the three sources of priors for the true effect size**

	Kraft	NCEE	Specific
Experimental design	✓	✓	✓
Standardised tests	✓	✓	✓
Summarising multiple studies	✓	✓	Varies
Tailored to the subject targeted	✓	✓	✓
Tailored to pupils' grade/year	✓		
No publication bias		✓	
Tailored to the intervention			✓

4. Results: Type M and Type S error

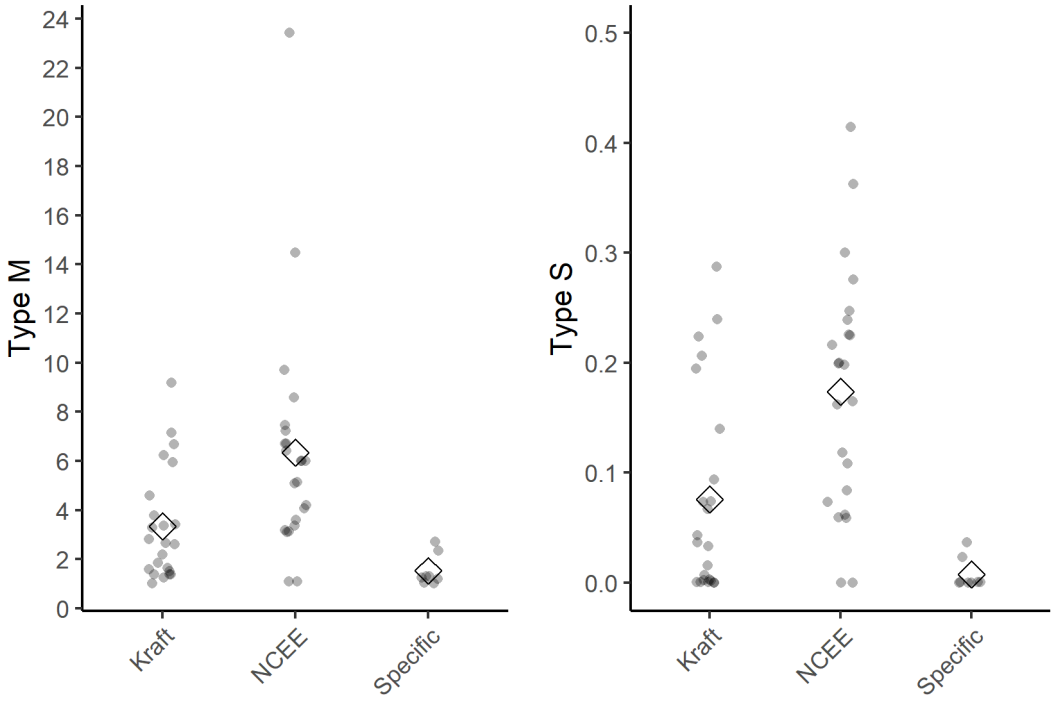
In Figure 4, we present estimates of the Type M error (left panel) and Type S error (right panel) for each trial, for each of the three sources of priors. The white diamond indicate the mean Type M error reported as a ratio (left hand panel) and the mean Type S error reported as a probability (right hand panel). The NCEE priors yield the largest and most dispersed estimates of both types of error, followed by the Kraft priors and then the intervention-specific priors. This in part reflects the smaller

average effect sizes found in NCEE trials, which might be expected, given that they employ standardized test as outcome measures and publish trial findings regardless of the results. Table 2 shows a numerical summary of the results.

Our estimates of mean Type M error range from 1.52 (intervention specific priors) through 3.33 (Kraft priors) up to 6.35 (NCEE priors). While these results are clearly somewhat different in magnitude, the findings suggest that trials show large Type M error across all three sets of assumptions we make about D. Even when we use our intervention specific priors - which lead to the smallest estimates of Type M error - we find that the effect sizes in this sample of trials are on average 52% larger than the true effect size. Further, as Figure 4 makes clear, these means conceal wide variation across trials. Again, even when we use our intervention specific priors, very few individual trials have a Type M error estimate close to one. In Appendix Table A2, we show the equivalent results for the subset of trials with $p < 0.05$ and assuming $\alpha = 0.05$. The mean Type M error falls to 1.20 when we use our intervention specific priors, although this is based on just four interventions. The mean Type M on the Kraft and NCEE assumptions fall slightly to 2.85 and 5.85, respectively. In sum, across all our analyses, we find consistent evidence of substantial promising trial bias in this sample of RCTs.

Our estimates of mean Type S error range from 1% (intervention specific priors) through 8% (Kraft priors) up to 17% (NCEE priors). Whilst the latter two results arguably represent sizable probabilities that a positive finding in fact reflects negative impact, our intervention specific priors suggests a very low mean probability of Type S error. By contrast with our findings relating to Type M error then, our findings cannot be said to be robust to our choice of prior assumptions about D. For the Kraft and NCEE priors, there is again wide variation at the level of individual trials, with e.g. half of trials having an estimated Type S above 20% based on the NCEE priors. When we restrict the sample to trials with $p < 0.05$ and assume $\alpha = 0.05$ the mean Type S is even smaller. In sum, when an intervention has been tested in a trial with promising results, the probability that its true effect is in fact negative is low.

Figure 4: Type M and Type S error by source of prior



Notes: Vertical axis in the left-hand panel shows the exaggeration ratio. Vertical axis in the right-hand panel shows the probability. Each dot is a trial. Diamonds show the mean.

Table 2: Type S and Type M error estimates under different prior assumptions

	Type M			Type S		
	Kraft	**NCEE**	**Specific**	**Kraft**	**NCEE**	**Specific**
Median	2.70	5.92	1.27	0.04	0.20	<0.01
Mean	3.33	6.35	1.52	0.08	0.17	0.01
SD	2.25	4.81	0.63	0.09	0.11	0.01
N (Trials)	23	23	8	23	23	8

Notes: Using all trials ever deemed 'promising' by the EEF and an alpha value of 0.11.

5. False positive risk

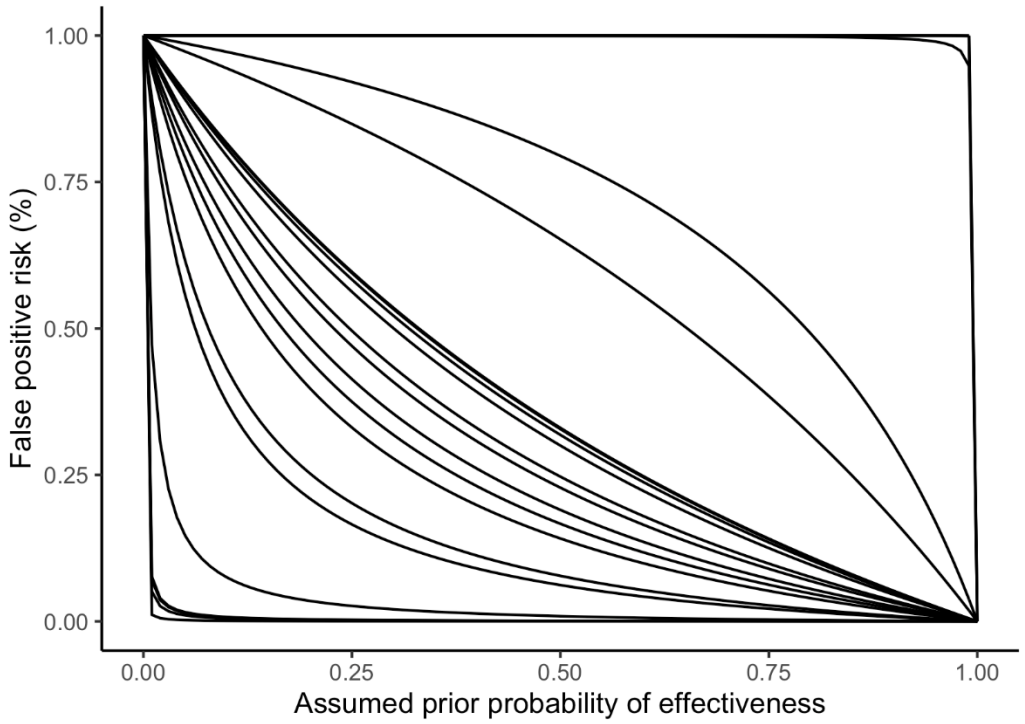
Recall that Type M error occurs when $D \neq \widehat{D}$ and $\widehat{D}$ has $p < \alpha$. A special case of Type M error is when $D = 0$. This is often referred to as a 'false positive' since the null hypothesis of no effect has been rejected when there is in fact no effect. In this case, the estimated effect is composed entirely of error. A common misconception about *p* values is that they give the probability that a result is in fact such a false positive (Greenland et al., 2016). Motivated in part by a pragmatic desire to provide a statistic that does what some people mistakenly think *p* values do, Colquhoun (2017, 2019) has developed a method for estimating the false positive risk for a given study. In this section, we use

Colquhoun's method to quantify the probability that the exaggeration is relative to a true effect size of zero. It is important to note that we are calculating the False Positive Risk (for a single study) not the False Positive Rate (across many studies), the latter of which has been criticized for conflating power and sensitivity (Mayo & Morey, 2017).

For a given trial, the probability of a false positive depends on the Bayes Factor (the ratio of how well the alternative and null hypotheses predict the data) and the prior odds of the intervention being effective (Colquhoun, 2017, 2019). Colquhoun has provided an R programme (2017) and a web-based application (Colquhoun & Longstaff, 2017) to calculate the False Positive Risk in simple trial designs. However, these do not allow for the clustered designs that are more common in education trials and that make up our sample of trials. We therefore set out how we have adapted the approach for cluster randomized trials in Appendix B. We can extract most of the quantities needed for this calculation from the EEF trial reports but still need to make one assumption: the prior probability that the intervention would be effective. Unlike our priors about the true effect size, we are not aware of any credible source of empirical evidence that could help us calibrate this assumption. Hence, we take an agnostic approach, and plot the estimated false positive risk across the full potential range (between 0 and 1) of the prior probability of there being an effect.

Figure 5 shows the results. The key finding is that, for the majority of trials, even small deviations from prior certainty that the intervention is effective lead to large increases in the false positive risk. An alternative way of interpreting these results is that only ten trials have an estimated false positive risk greater than 50% when the prior probability of effectiveness is 50%. As such, they do not have the evidential weight to sway somebody who was equally disposed to believe the intervention effective/ineffective before the trial. This is despite the fact that the results have all been deemed promising. At the beginning of this section, we referenced the misconception that $p < 0.05$ means there is less than a 5% risk of a false positive. A third and final way of looking at these results is then to note that, with a 50% assumed prior probability of effectiveness, only five trials in this sample have an estimated false positive risk below 5%. Indeed, across the trials, the median false positive risk with a 50% assumed prior probability of effectiveness is 22.8%.

Figure 5: False positive risk



Notes: Vertical axis reports the estimated false positive risk (as a percentage) for each trial analyzed, given the assumed prior probability of effectiveness on the horizontal axis, which is allowed to range across the full probability space. All other aspects of the calculation are based on elements of trial design and results extracted from relevant reports on EEF website.

6. Discussion

We set out to investigate the prevalence and magnitude of promising trials bias in education RCTs. Even when using our intervention specific priors – which result in our lowest estimates of Type M error, we find that the average promising trial in our sample exaggerates the true effect size by 52%. Arguably the most reasonable priors to use in such calculations are those derived from NCEE trials, since this constitutes a highly analogous out-of-sample benchmark, which is not affected by publication bias. Using the NCEE assumptions, the average trial finding in our sample exaggerates the true effect size by 535%. Indeed, regardless of which of the three sets of assumptions that we employ, it is clear that we would expect promising trials to substantially overestimate effect sizes and that replications of such trials would likely find substantially lower effects. Further, when we plot the probability of promising findings representing a false positive, we find that many do not contain sufficient evidential weight to convince somebody who already thought it 50% likely that the

intervention would work prior to the trial. These results graphically illustrate the very limited information contained within point estimates of 'promising' findings from low-powered education trials .

6.1 Limitations

These findings should, of course, be interpreted in light of the limitations of this study. Since we set out to make claims about the probability of hypotheses given data, it was necessary for us to invoke priors in our analysis. The true value of D cannot be known with certainty and our conclusions depend on the accuracy of the assumptions we have made. Having said that, we believe our setting is unusually well suited to providing good empirical benchmarks to inform these priors. First, we were able to draw on three different empirical sources in order to calibrate our priors. Though none of these is ideal (see Table 1) they collectively provide good evidence about the effect sizes that might be expected. Most importantly, our finding of large Type M errors is insensitive to which of the three priors we use. Another limitation relates to our sample. Our focus on promising trials limited us to just 23 EEF RCTs. Future work should therefore look to expand this research, as and when more trials are published.

6.2 Implications

Given the magnitude of the Type M errors we observed across our sample of deemed-to-be-promising trials, it is clear that promising trials bias is an important issue for trial commissioners. We suggest that funders should supplement effect sizes and p values with retrospective design analyses of the sort conducted here when deciding whether to commission trials of previously-evaluated interventions. Some trials with $p < .05$ are very unlikely to replicate on any plausible prior and this should be taken into account when making commissioning decisions. In addition, funders should increase the power of the trials that they do commission. Funding fewer effectiveness trials that are unlikely to yield results consistent with promising prior efficacy trials would help to cover the costs of funding larger efficacy trials to begin with.

The findings also have implications for educators looking to understand the costs and benefits of different approaches. In particular, it would be prudent for them to expect smaller effects than would be suggested by single trial results listed in 'warehouses' or 'toolkits' of promising interventions.

In addition, these findings have implications for researchers who design and analyze RCTs. We believe that researchers should consider using prospective design analysis when planning RCTs in education. In particular, the information regarding the potential for Type M error in trials of different sizes complements calculations regarding the minimum detectable effect size. Gelman and Carlin summarize this as moving from asking only "What is the power of a test?" to also asking "What might be expected to happen in [future] studies of this size?" (2014, p. 649). Trial analysts should also report retrospective estimates of Type M errors in order to give readers the answer to the question: how likely is it that an exact replication of this trial would come to a different conclusion? The two general sources of empirical evidence for priors that we set out in this paper provide a useful starting point for researchers looking to conduct either prospective or retrospective design analysis.

Our findings also prompt two further reflections about impact evaluation more generally. First, our analysis raises questions about the appropriateness of elevating RCTs above other evaluation designs in education. RCTs sidestep the fundamental problem of causal inference at the cost of replacing it with the challenge of randomisation inference. In education, where effect sizes tend to be small (Kraft, 2020), this is problematic, since many trials end up having low power and yielding uninformative findings (Lortie-Forgues & Inglis, 2019). Observational evaluation designs such as matching and comparative interrupted time series do not constitute unbiased estimators in the way that RCTS do. However, empirical research shows that such non-experimental designs come close to reproducing those from education trials on average, without the attendant problems with power (Clair, Hallberg, & Cook, 2016; Cook, 2017; Weidmann & Miratrix, 2019). At the very least, this suggests that evidence from a range of rigorous large-scale observational evaluations should be seen as complementary, rather than inferior, to RCTs.

Second, our research raises questions about the significance of implementation fidelity in challenges around "scaling up" interventions. Many other researchers have noticed that scale-up or effectiveness

trials often find smaller effects than are found in initial or efficacy trials of the same intervention (e.g., Protzo & Schooler, 2017). In education, researchers almost always account for this with reference to the difficulties of maintaining fidelity when an intervention is implemented at scale (e.g., Elmore, 1996; Fletcher-Wood & Zuccollo, 2020; Honig, 2006). Our analysis suggests a different, or at least supplementary, account. In a context where initial efficacy trials tend to have low power, the statistical significance filters means that we would expect smaller – perhaps much smaller – effects in subsequent effectiveness trials, even if implementation fidelity was perfectly maintained. Thus, researchers cannot simply point to difficulties in implementation when scale-up trials fail to replicate initially promising results.

References

Anderson, S. F. (2020). Misinterpreting p: The discrepancy between p values and the probability the null hypothesis is true, the influence of multiple testing, and implications for the replication crisis. *Psychological Methods*, *25*(5), 596-609.

Bullock, J. C. (2005). Effects of the Accelerated Reader on reading performance of third, fourth, and fifth-grade students in one western Oregon elementary school (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3181085)

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscienc*e, 14(5), 365– 376.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283-292.

Clair, T., Hallberg, K., & Cook, T. D. (2016). The validity and precision of the comparative interrupted time-series design: three within-study comparisons. *Journal of Educational and Behavioral Statistics*, *41*(3), 269-299.

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, *4*(12), 171085.

Colquhoun, D., & Longstaff, C. (2017). False Positive Risk Calculator. Retrieved from: http://fpr-calc.ucl.ac.uk/

Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about p-values. *The American Statistician*, *73*(1), 192-201.

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, *60*(3), 276-291.

Cook, T. (2017). Empirical demonstrations of the internal validity of certain quasi-experimental designs. Retrieved from: https://cspv.colorado.edu/blueprints/tom-cook-talks-october-13-2017/tom_cook_technical_talk_presentation_2017-10-13.pptx

Dawson, A., Yeomans, E., & Brown, E. R. (2018). Methodological challenges in education RCTs: reflections from England's Education Endowment Foundation. *Educational Research*, *60*(3), 292-310.

Deke, J., & Finucane, M. (2019). *Moving beyond statistical significance: the BASIE (BAyeSian Interpretation of Estimates) framework for interpreting findings from impact evaluations*. U.S. Department of Health and Human Services.

Education Endowment Foundation [EEF] (2020). Statement on statistical significance and uncertainty of impact estimates for EEF evaluations. Retrieved from: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Research_Report/Statement_on_statistical_significance_and_uncertainty_of_impact_estimates_for_EEF_evaluations

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

Fletcher-Wood, H. & Zuccollo, J. (2020). *The effects of high-quality professional development on teachers and students: A rapid review and meta-analysis*. Education Policy Institute.

Foliano, F., Rolfe, H., Buzzeo, J., Runge, J., Wilkinson, D. (2019). Changing Mindsets Effectiveness Trial*: Evaluation report and executive summary*. Education Endowment Foundation.†

Klein, D. (2017). RDESIGNI: Stata module to perform design analysis, Statistical Software Components S458423, Boston College Department of Economics.

Gage, N. A., Cook, B. G., & Reichow, B. (2017). Publication bias in special education meta-analyses. *Exceptional Children*, *83*(4), 428-445.

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6), 641–651.

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, *15*(3), 373-390.

Gorard, S., Siddiqui, N., & Huat See, B. (2014). *Switch-on Reading: Evaluation report and executive summary*. Education Endowment Foundation.*†
Gorard, S., Siddiqui, N., & Huat See, B. (2015). *Philosophy for Children: Evaluation report and executive summary*. Education Endowment Foundation.*
Gorard, S., Siddiqui, N., Huat See, B. (2015). *Accelerated Reader: Evaluation report and executive summary*. Education Endowment Foundation.*
Gorard, S., Siddiqui, N., Huat See, B., Smith, E., White, P. (2017). *Children's University: Evaluation report and executive summary*. Education Endowment Foundation.*
Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337-350.
Haynes, L., Service, O., Goldacre, B. & Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. UK Cabinet Office.
Hedges, L. V., & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research*, *60*(3), 265-275.
Hodgen, J., Adkins, M., Ainsworth, S., & Evans, S. (2019). *Catch Up Numeracy: Evaluation report and executive summary*. Education Endowment Foundation.†
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960.
Honig, M. I. (2006). *New directions in education policy implementation*. State University of New York Press.
Hume, S., O'Reilly, F., Groot, B., Chande, R., Sanders, M., Hollingsworth, A., Ter Meer, J., Barnes, J., Booth, S., Kozman, E., & Soon, X. (2018). *Improving engagement and attainment in maths and English courses: insights from behavioural research.* Department for Education.
Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(2), 481-502.
Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640-648.
Jay, T., Willis, B., Thomas, P., Taylor, R., Moore, N., Burnett, C., Merchant, G., Stevens, A. (2017). *Dialogic Teaching: Evaluation report and executive summary*. Education Endowment Foundation.*
Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253.
Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547-588.
Kraft, M. A., & Monti-Nussbaum, M. (2017). Can schools enable parents to prevent summer learning loss? A text-messaging field experiment to promote literacy skills. *The ANNALS of the American Academy of Political and Social Science*, *674*(1), 85-112.
Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, 31, 107–112.
Lord, P., Bradshaw, S., Stevens, E., & Styles, B. (2015). *Perry Beaches Coaching Programme: Evaluation report and executive summary*. Education Endowment Foundation.*
Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, 48(3), 158–166.
Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260-293.
Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304-316.
Mayo, D., & Morey, R. D. (2017). A poor prognosis for the diagnostic screening critique of statistical tests. Working paper.

McNally, S., Ruiz-Valenzuela, J., & Rolfe, H. (2016). *ABRA Online Reading Support: Evaluation report and executive summary*. Education Endowment Foundation.*

Miller, S., Davison, J., Yohanis, J., Sloan, S., Gildea, A., & Thurston, A. (2016). *Texting Parents: Evaluation report and executive summary*. Education Endowment Foundation.*

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

NfER (2014). *Catch Up Numeracy: Evaluation report and executive summary*. Education Endowment Foundation.*†

Nunes, T., Barros, R., Evangelou, M., Strand, S., Mathers, S., Sanders-Ellis, D. (2018). *FirstClass@Number: Evaluation report and executive summary*. Education Endowment Foundation.*

Nunes, T., Malmberg, L., Evans, D., Sanders-Ellis, D., Baker, S., Barros, R., Bryant, P., Evangelou, M. (2019). *onebillion: Evaluation report and executive summary*. Education Endowment Foundation.*

Outhwaite, L. A., Faulder, M., Gulliford, A., & Pitchford, N. J. (2019). Raising early achievement in math with interactive apps: A randomized control trial. *Journal of Educational Psychology, 111*(2), 284–298.

Patel, R., Jabin, N., Bussard, L., Cartagena, J., Haywood, S., & Lumpkin, M. (2017). *Switch-on Effectiveness Trial: Evaluation report and executive summary*. Education Endowment Foundation.†

Pocock, S. J. (1982), Statistical aspects of clinical trial design. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 31, 1-18.

Protzko, J., & Schooler, J. W. (2017). Decline effects: Types, mechanisms, and personal reflections. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (p. 85–107). Wiley-Blackwell.

Ritter, G. W., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The effectiveness of volunteer tutoring programs for elementary and middle school students: A meta-analysis. *Review of Educational Research, 79*(1), 3-38.

Rienzo, C., Rolfe, H., & Wilkinson, D. (2015). *Changing Mindsets: Evaluation report and executive summary*. Education Endowment Foundation.*†

Sibieta, L., Kotecha, M., Skipp, A. (2016). *Nuffield Early Language Intervention: Evaluation report and executive summary*. Education Endowment Foundation.*

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science, 29*(4), 549-571.

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.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research, 79*(4), 1391-1466.

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

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the US Institute of Education Sciences. *International Journal of Research & Method in Education, 39*(3), 255-267.

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

Styles, B., & Bradshaw, S. (2015). *Talk for Literacy: Evaluation report and executive summary*. Education Endowment Foundation.*

Torgerson, C. J., & Torgerson, D. J. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies, 49*(3), 316-328.

Torgerson, C. J., Torgerson, D. J., Birks, Y. F., & Porthouse, J. (2005). A comparison of randomised controlled trials in health and education. *British Educational Research Journal*, *31*(6), 761-785.

Torgerson, D., Torgerson, C., Mitchell, N., Buckley, H., Ainsworth, H., Heaps, C., & Jefferson, L. (2014). *Grammar for Writing: Evaluation report and executive summary*. Education Endowment Foundation.*†

Torgerson, C., Bell, K., Coleman, E., Elliott, L., Fairhurst, C., Gascoine, L., Hewitt, C., & Torgerson, D. (2018). *Tutor Trust Affordable Primary Tuition: Evaluation report and executive summary*. Education Endowment Foundation.*

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688-701.

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151-175.

Weidmann, B., & Miratrix, L. (2020). Lurking inferential monsters? Quantifying selection bias in non-experimental evaluations of school programs. *Journal of Policy Analysis & Management*, DOI:10.1002/pam.22236.

What Works Clearinghouse (2016). Accelerated Reader: A summary of findings from a systematic review of the evidence. Retrieved from: https://files.eric.ed.gov/fulltext/ED566522.pdf

Worth, J., Sizmur, J., Ager, R., & Styles, B. (2015). *Improving Numeracy and Literacy: Evaluation report and executive summary*. Education Endowment Foundation.*†

Wright, H., Dorsett, R., Anders, J., Buzzeo, J., Runge, J., & Sanders, M. (2019). *Improving Working Memory: Evaluation report and executive summary*. Education Endowment Foundation.*

Appendix A

Table A1: Intervention-specific priors for the effect size

Subject	Trial	Source of prior	Justification	Prior
English	Nuffield Early Language	Slavin, Lake, Chambers, Cheung, & Davis (2009)	Meta-analysis of non-remedial reading programmes using RCTs or quasi-experiments using tests that are not tailored to the intervention. Effect size is the mean for phonological awareness interventions that (like Nuffield Early Language) focus on talking to support early reading development, as reported on page 1406.	0.22
	Talk for Literacy	Slavin, Lake, Chambers, Cheung, & Davis (2009)	Meta-analysis of non-remedial reading programmes using pre-post observationally equivalent control group designs or RCTs and using tests that are not tailored to the intervention. Effect size is the mean for phonological awareness interventions that (like Talk for Literacy) focus on talking to support early reading development, as reported on page 1406.	0.22
Mathematics	1st Class at Number	See, Morris, Gorard, & Siddiqui (2019)	RCT of a similar intervention providing one-to-one mathematics tutoring led by teaching assistants with outcomes measured using the standardised InCAS assessment. Effect size is taken from Table 6.	0.12
	onebillion	Outhwaite, Faulder, Gulliford, & Pitchford (2019)	Prior RCT of the same intervention using the standardised PTM5 mathematics tests as the outcome measure. Effect size is from the comparison of the control group and the time-equivalent treatment group.	0.21
Science	TDT Science	Lynch, Hill, Gonzales, & Pollard (2019)	Meta-analysis of experimental and quasi-experimental evaluations of professional development programmes for science and mathematics teachers. Effect size is calculated for an RCT using state-standardised tests, based on Table 3.	0.11
General	Affordable Tuition	Ritter, Barnett, Deny, & Albin (2009)	Meta-analysis of RCTs of volunteer (adult non-professional) one-to-one tutoring programmes using standardised test outcome measures. The effect size is that for 'global reading' outcome measures on page 16.	0.26
	Changing Mindsets	Sisk, Burgoyne, Sun, Butler, & Macnamara (2018)	Meta-analysis of RCTs or quasi-experimental evaluations of growth mindset interventions using academic achievement outcome measures. Effect size is the average from meta-analysis 2, as reported on page 565.	0.08
	Graduate Coaching Prog.	Ritter, Barnett, Deny, & Albin (2009)	Meta-analysis of RCTs of volunteer (adult non-professional) one-to-one tutoring programmes using standardised test outcome measures. The effect size is that for 'global reading' outcome measures on page 16.	0.26

Table A2: Summarising Type S and Type M error estimates for trials with p<0.05

	Type M			Type S		
	Kraft	**NCEE**	**Specific**	**Kraft**	**NCEE**	**Specific**
Median	1.95	5.23	1.19	<0.01	0.11	<0.01
Mean	2.85	5.85	1.20	0.03	0.12	<0.01
SD	1.97	3.67	0.17	0.06	0.09	<0.01
N (Trials)	14	14	4	14	14	4

Notes: Using all EEF trials with p<0.05 and an alpha level of 0.05.

Appendix B: Calculating False Positive Risk

To calculate the false positive risk on each for each of our trials, we follow the method laid out by Colquhoun (2019), starting by using Bayes' theorem to observe that:

$$\frac{P(H_1|data)}{P(H_0|data)} = \frac{P(data|H_1)}{P(data|H_0)} * \frac{P(H_1)}{P(H_0)} \tag{10}$$

where H_0 is the null hypothesis and $H_1$is the alternative hypothesis; in other words that the posterior odds of the alternative hypothesis are equal to the Bayes factor multiplied by the prior odds of the alternative hypothesis. By using a simple hypothesis test of the null hypothesis that $H_0: \theta = 0$ against the alternative hypothesis of $H_1: \theta \neq 0$ the Bayes factor in the equation above becomes a simple likelihood ratio:

$$L_{10} = \frac{P(\text{data}|H_1)}{P(\text{data}|H_0)} \tag{11}$$

which can be calculated, at least approximately, from information usually reported in trials i.e. the sample size, the *p* value, and the degrees of freedom of the t test (further details below). Beyond this, we need only specify a simple prior probability $P(H_1) = 1 - P(H_0)$, rather than a prior distribution, allowing us to simplify Equation 12 to the following:

$$\frac{P(H_1|\text{data})}{1 - P(\text{data}|H_1)} = L_{10} * \frac{P(H_1)}{1 - P(H_1)} \tag{12}$$

By rearranging for $P(H_1|\text{data})$ i.e. the False Positive Risk we can estimate this quantity as follows:

$$FPR = \frac{1}{1 + L_{10}\frac{P(H_1)}{1 - P(H_1)}} \tag{13}$$

To estimate the likelihood ratio L_{10} we start with the trial's reported *p* value and use this to calculate the critical value of the t statistic for the test of the reported effect size:

$$t_{crit} = qt\left(1 - \frac{p}{2}, df, ncp = 0\right) \tag{14}$$

where $qt()$ is the inverse cumulative t distribution, p is the observed *p* value, df is the number of degrees of freedom of the test, and ncp is the non-centrality parameter (which is zero under the null hypothesis).

We then use this to calculate the probability density under the null hypothesis:

$$y_0 = dt(t_{crit}, df, ncp = 0) \tag{15}$$

where $dt()$ is the probability density function of the t distribution with df degrees of freedom and a non-centrality parameter ncp (again zero for the null hypothesis).

Likewise, the probability density under the alternative hypothesis can be calculated, although here we need to use a noncentral t distribution with non-centrality parameter $ncp = \frac{E}{S_E}$ where E is the difference between means (we use the reported effect size) and S_E is the standard deviation of the difference between means (we have to estimate this assuming independence of the samples and that the variance of the outcome measure, having been standardised to calculate the effect size, will be approximately 1 in both groups):

$$y_0 = dt(t_{crit}, df, ncp) \tag{16}$$

Assuming the reported *p* values come from two-sided tests, under the null hypothesis the probability occurs at $\pm t_{crit}$ and, as such, we calculate:

$$L_{10} = \frac{P(\text{data}|H_1)}{P(\text{data}|H_0)} = \frac{y_1}{2y_0} \tag{17}$$

Since we can generally recover all other quantities needed to calculate (or at least estimate) the inputs to equation 15 from the trial reports, we need only make one assumption: $P(H_1)$ i.e. the prior probability that the intervention would be effective.