
Article:
Lortie-Forgues, Hugues and Inglis, Matthew (2019) Rigorous Large-Scale Educational RCTs are Often Uninformative:Should We Be Concerned? Educational Researcher. pp. 158-166. ISSN: 0013-189X

https://doi.org/10.3102/0013189X19832850

Rigorous Large-Scale Educational RCTs are Often Uninformative: Should We Be Concerned?

Hugues Lortie-Forgues
University of York, UK

Matthew Inglis
Loughborough University, UK

Department of Education
University of York
York
YO10 5DD
United Kingdom
Email: hugues.lortie-forgues@york.ac.uk

Please cite as:

Lortie-Forgues, H., & Inglis, M. (in press). Rigorous Large-Scale Educational RCTs are Often Uninformative: Should We Be Concerned? *Educational Researcher.*

Acknowledgements

We are grateful to Adrian Simpson for his suggestions and for alerting us to a problem with an R function used to calculate results in an earlier version of this manuscript. We thank David W. Braithwaite, Steve Higgins, Robert M. Klassen, Michael Schneider, and ZhiMin Xiao for their suggestions; Erin Pollard (IES) for assistance with NCEE trials, and Anh Nguyen Van Pham for assistance with coding.

Abstract

There are a growing number of large-scale educational Randomized Controlled Trials (RCTs). Considering their expense, it is important to reflect on the effectiveness of this approach. We assessed the magnitude and precision of effects found in those large-scale RCTs commissioned by the EEF (UK) and the NCEE (US) which evaluated interventions aimed at improving academic achievement in K-12 (141 RCTs; 1,222,024 students). The mean effect size was 0.06 standard deviations (SDs). These sat within relatively large confidence intervals (mean width 0.30 SDs) which meant that the results were often uninformative (the median Bayes factor was 0.56). We argue that our field needs, as a priority, to understand why educational RCTs often find small and uninformative effects.

Rigorous Large-Scale Educational RCTs are Often Uninformative: Should We Be Concerned?

Large-scale Randomized Controlled Trials (RCTs) are now regularly used to evaluate educational interventions. For example, the US-based National Center for Educational Evaluation and Regional Assistance (NCEE) started funding large-scale RCTs in 2002, and the UK-based Education Endowment Foundation (EEF) has funded more than 160 since 2012. This trend is not limited to these two countries: in recent years funding organizations in the European Union (e.g., European Schoolnet), Japan (e.g., Nippon Foundation), Australia (e.g., Social Ventures), Switzerland (e.g., Jacob's Foundation), Brazil (e.g., Lemann Foundation) and Bangladesh (e.g., BRAC) have also prioritized RCTs in education.

Evaluating the efficacy of educational programs before implementation is important to avoid wasting resources. In medicine, there are many instances where RCTs have shown that promising treatments were ineffective or harmful (Sibbald & Roland, 1998). However, conducting large-scale RCTs is expensive. For example, the EEF spends around £500,000 per trial (EEF, 2015a). Given the growing number of large-scale RCTs in education, and their expense, it is important to reflect on how informative this new research focus has been. To our knowledge, no study has systematically evaluated this recent trend. In this paper we use empirical data from two prominent educational funding bodies to evaluate the typical effects produced by large-scale educational RCTs. Our aim is to provide an empirical basis for discussions of the field's efforts to build rigorous scientific evidence.

Randomized Control Trials

RCTs are widely regarded as the 'gold standard' for measuring the efficacy of interventions (Pocock, 1983). In their simplest form, participants are randomly assigned to an experimental group which receives the intervention, or a control group that receives an alternative treatment or possibly no treatment. The effectiveness of the intervention is then determined by comparing the outcomes between groups. RCTs are highly regarded because, compared with other types of studies (e.g., case studies), they ensure that the groups are probabilistically identical at the outset and that any difference in outcome are therefore *caused* by the intervention (assuming that the probability of the difference occurring by chance is sufficiently low).

Unfortunately, not all RCTs are of the same quality (e.g., Higgins et al., 2011). The conclusions of an RCT can be distorted or of limited use if, for example, the sample is too small or not representative, if the allocation of the participants is compromised, if the outcomes are selectively reported, if attrition is ignored, or if the outcome measure provides

an unfair advantage to the intervention group (by including, for example, material that is taught to the intervention group but not the control group).

In the paper, we focus on RCTs commissioned by the EEF and NCEE. Both organizations commission trials that involve large numbers of participants, often more than a thousand per trial. Moreover, to ensure the quality of their trials, both organizations follow strict methodological guidelines that include comparing the intervention to an active control group, using reliable and valid outcome measures that are not excessively aligned with the intervention, preregistering measures and analyses, commissioning independent evaluators to randomize the participants and analyse the data, and publishing the findings regardless of outcome (EEF, 2017; NCEE, 2017).

The EEF and NCEE are not the only funders who commission rigorous large-scale RCTs (for example, the National Center for Education Research (NCER), another US-based funder, also commissions similar trials). However, they are the only funders we know of who explicitly require all their trials to be published in a standard format that prevents publication bias. This is vital, as publication bias can substantially inflate effects in published results (Rosenthal, 1979).

The EEF and NCEE share many principles, but their trials are not identical. Both funders claim to evaluate promising interventions, but the way these are selected differs. For the EEF, the trials are initiated by investigators (e.g., universities, schools) through competitive grant programs. The applicant provides evidence for the principles behind the intervention and evidence of effectiveness, which is then evaluated via a review process. In contrast, the NCEE tests promising interventions that are initiated by the U.S. government. The two funders also differ in the type of trial they conduct. The EEF commissions both efficacy trials (trials meant to test the intervention in ideal conditions) and effectiveness trials (typically larger trials tested in more representative conditions with less oversight from the developers). In contrast, the NCEE only commissions effectiveness trials.

What Should We Expect From Rigorous Large-Scale RCTs?

The goal of all empirical research is to produce new information, and the same is true for rigorous large-scale RCTs in education. Unsurprisingly then, both the EEF and NCEE state that they aim to produce *informative* RCTs (EEF 2015a; NCEE, 2013). While there may be more direct classroom implications when an RCT finds that an intervention works (at least in comparison to the activity undertaken by the control group), RCTs which convincingly demonstrate that a given intervention does not work are equally valuable. Given this, in our terms a trial is *informative* if it allows us to determine with confidence that an educational intervention is either effective or ineffective. A trial is *uninformative* if its findings are

consistent with the associated intervention being either effective or ineffective. Whether or not an RCT is informative in these terms therefore depends upon both its effect size and the precision with which that effect size is estimated.

Effect Sizes. The typical effect of educational interventions is usually said to fall between 0.25 and 0.50 standard deviations (SDs) (e.g., Hattie, 2009; Hill, Bloom, Black, & Lipsey, 2008; Lipsey & Wilson, 1993). For example, Hattie's (2009) synthesis of more than 800 educational meta-analyses found an average effect size of 0.40 SDs. However, we might expect rigorous large-scale RCTs to produce smaller effect sizes than those present in the wider literature. One reason concerns the distinctive methodological features of these studies. For example, studies with randomized designs typically produce smaller effects than non-randomized studies: Cheung and Slavin (2016) found that the effect sizes from randomized educational experiments was 0.16 compared to 0.23 for non-randomized quasi-experimental studies. Likewise, studies using independent outcome measures, such as standardised tests, tend to produce smaller effects than studies using researcher-made measures. For instance, when comparing the performance of 5th and 6th graders on a standardized test of reading, the impact of an additional year of instruction and maturation is only around 0.23 SDs (Bloom, Hill, Black, & Lipsey, 2008). Similarly, studies comparing the intervention to an active control group, studies using conservative data-analyses (e.g., intention to treat), and studies sampling from large and heterogeneous populations also tend to produce smaller effect sizes (e.g., Cheung & Slavin, 2016; Karlsson & Bergmark, 2015). All these characteristics, which are present simultaneously in rigorous large-scale RCTs, are likely to reduce estimates of effect size.

Rigorous large-scale RCTs might also produce smaller effect sizes than those found in the wider literature because parts of this literature are biased. Unfavorable findings from traditional research are less likely to be published (Rosenthal, 1979), and many researchers selectively report analyses and conduct unplanned analyses (John, Lowenstein, & Prelec, 2012). Both phenomena – which are prevented by the EEF's and NCEE's state-of-the-art methodological requirements – increase the proportion of false positives and cause inflated effects in traditional research. Illustrative of this point are recent relatively unsuccessful attempts to replicate published psychology findings (Open Science Collaboration, 2015).

All of these factors suggest that the effect sizes which we should expect from rigorous large-scale RCTs will be lower than those found in the wider educational literature. Specifically, we would certainly expect effect sizes lower than the 0.4 reported by Hattie (2009), and probably lower than those associated with a year of maturation and instruction

(e.g., 0.23 SDs from 5th to 6th grade; Bloom et al., 2008). However, it is unclear how much lower. Addressing this question is one aim of the current study.

Precision. A second component of an RCT's informativeness is the precision with which the effect size is estimated (i.e., the width of the confidence interval around this estimate). Precision is largely determined by the number of participants in the trial: the more participants, the more precise the estimate. Precision is crucial to the interpretation of a trial's outcome. When the effects are small, low precision may mean that a trial cannot determine whether an intervention is effective or ineffective; i.e., that the trial is uninformative (for instance, an RCT which yielded an effect size estimate of 0 within a confidence interval of -0.25 to 0.25 would be consistent with three different possibilities: that the intervention is ineffective, that it has a positive effect of practical significance, and that it has a negative effect of practical significance). Consequently, measuring effect sizes with appropriate precision – with appropriate power – is critical. Unfortunately, appropriately powering a trial can be challenging, because of the large number of participants required and the clustered nature of educational data.

Bayes factors. An alternative way of evaluating a study's informativeness is to calculate a Bayes factor, which quantifies the relative evidence that the data provide for one hypothesis compared to another (Jeffreys, 1961). For example, a Bayes factor of 5 in favour of the alternative hypothesis against the null hypothesis implies that the observed data are 5 times more likely under the alternative than under the null. The Bayesian approach has the advantage over traditional null hypothesis significance testing in that it allows one to determine which of three possibilities the data support: the null hypothesis of no effect, an alternative hypothesis that models the effect expected if the intervention were effective, or neither of these (i.e. the data are uninformative) (Dienes, 2011). Jeffreys (1961, appendix B) offered guidelines by which Bayes factors can be interpreted, suggesting that figures between 3 and 1/3 are "hardly worth mentioning". In other words, if the observed data are less than 3 times as likely to occur under the alternative as the null (or vice versa) then the trial is uninformative. Jeffrey's further suggested that Bayes factors between 3 and 10 (or 1/3 and 1/10) indicate moderate evidence; those between 10 and 30 (1/10 and 1/30) indicate strong evidence; those between 30 and 100 (1/30 and 1/100) indicate very strong evidence; and those over 100 (below 1/100) indicate decisive evidence.

In sum, our goal was to assess the extent to which rigorous large-scale RCTs in education are informative. Addressing this goal is important. In view of the recent increased focus on educational RCTs, and the relatively high cost of conducting them, it is important that the field reflects on the extent to which they provide useful information. To address this

we: first, assessed the size of the effects produced by rigorous large-scale RCTs; second, considered how precisely these effects were estimated (by calculating associated confidence intervals); and third, directly determined whether or not these trials were informative by calculating Bayes factors.

Method

Identification

For the EEF trials, we retrieved all the evaluation reports available in the projects and evaluation section of the EEF website (98 reports). For the NCEE trials, we first retrieved the abstracts of all the reports with a NCEE number on the publications and products search database of the Institute of Education Sciences (IES) and on the ERIC database (302 abstracts). Both authors then read all the abstracts independently to determine their suitability for the study. Most NCEE reports were not describing trials, were summarizing trials described in other reports, or were describing trials that were not yet completed (interim reports). In total, only 56 reports were considered relevant. All 154 reports (98 EEF, 56 NCEE) were then read. Some of the reports included two or more trials testing different interventions with different participants. These trials were considered to be independent. In the end, 190 independent trials (119 EEF; 71 NCEE) were matched against our eligibility criteria. The search was finalized on June 1st, 2018.

Eligibility

For a trial to be eligible: (a) allocation to the intervention and control groups had to be random, (b) students had to be in grades K-12 (Key Stages 1 to 4 in the UK), and (c) the outcome(s) had to be of an academic nature. Pilot trials (i.e., small-scale trials evaluated mainly through qualitative measures) were excluded. Eligibility was determined by the two authors and discussion was used to resolve discrepancies.

The Sample

Of the 190 trials considered, 141 matched our eligibility criteria and were included in the analysis: 82 trials from the EEF (140 distinct effect sizes, 790,279 students) and 59 trials from the NCEE (131 distinct effect sizes, 431,745 students). A full list of trials included in our sample is given in the Supplemental Material available online.

Extraction and Coding

All the trials reported their outcomes in terms of standardized mean differences (which, for simplicity, we refer to as ‘effect sizes’). These were directly extracted from the reports. We recorded only effect sizes associated with primary academic outcomes (i.e., the

main outcomes that the trial was designed to address). When the report did not identify which outcome was primary, we used the effect sizes reported in the summary of the evaluation report. When a trial reported multiple primary outcomes, we only considered a single, randomly selected outcome to avoid violating statistical independence (Lipsey & Wilson, 2001). To compare, we also conducted additional analysis: (i) using the first outcome reported, (ii) using the outcome associated with the largest effect size, and (iii) using every outcome from every trial as if they were independent (as shown in the Supplemental Material, all these approaches gave broadly similar findings). Effect sizes were coded as positive when the intervention group performed better than the control group and negative when it performed worse.

To measure how precisely effect sizes were estimated, we coded the standard error of each effect (SE_d) which was retrieved from the report, or estimated from the 95% confidence interval or the p-value when not available. In 8 trials (14 distinct outcomes) there was not enough information to compute the SE_d. In these cases the value was estimated from the sample size and effect size (see Borenstein et al., 2009, p. 27), a procedure which ignores clusters, and thus can overstate the accuracy of the estimated effect. Excluding these 8 trials from our analysis does not materially affect our conclusions.

We also coded the topic of the outcome measures (e.g., reading, mathematics), the age of participants, the sample size, and the report's year of publication. For the EEF trials, we also coded the type of trial (efficacy or effectiveness), the total cost of the trial, the cost of the intervention per pupil (a number from 1 [low-cost] to 5 [high-cost]), and the quality of trial (a number from 0 [low-quality] to 5 [high-quality]). These latter two variables were determined by EEF-commissioned reviewers (EEF, 2015b; 2016).

To ensure the accuracy of the data entry, all the characteristics (e.g., type, cost…) of 43 randomly selected trials (30% of all trials) were recoded independently by a second rater. The match was 99%. Discussion was used to resolve the discrepancies. The raw data are available in the Supplemental Material available online.

Results

The included interventions targeted students in elementary school (59%), secondary school (22%), kindergarten (6%) or a combination of these levels (14%). Most outcome measures were related to language (63%) or mathematics (27%), but some were related to sciences (3%), economics (1%), or encompassed more than one topic (6%).

Figure 1 shows the distribution of observed effect sizes, which was unimodal. Figure 2 shows a funnel plot of the sample sizes (represented by the inverse of the variance) against the effect size of each trial, and indicates that more extreme effects (positive and negative)

were typically found in smaller, less precise, trials. Table 1 summarizes the findings of EEF trials, NCEE trials, and of both funders combined.

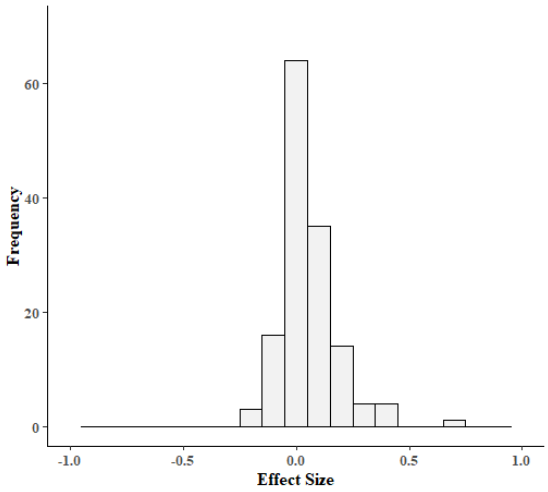


Figure 1. *The distribution of effect sizes from the 141 trials commissioned by the EEF and NCEE.*

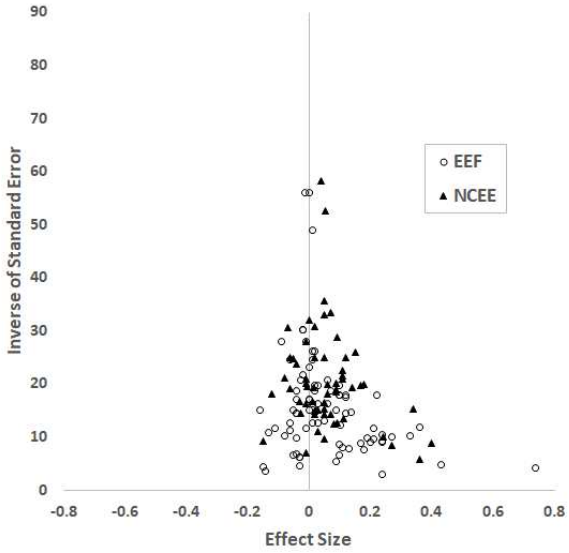


Figure 2. *A funnel plot of effect sizes from the trials commissioned by the EEF (82 trials) and NCEE (59 trials).*

Table 1. *Description of the trials commissioned by the EEF and NCEE.*

	EEF	NCEE	Overall
N. trials	82	59	141
Total N. participants	790,279	431,745	1,222,024
Median N. per trial	2,222	2,594	2,386
Effect size			
Min	-0.16	-0.15	-0.16
Max	0.74	0.40	0.74
Median	0.01	0.05	0.03
% positive	60%	71%	65%
Unweighted Mean	0.06	0.06	0.06
95% CI	(0.03; 0.09)	(0.03; 0.09)	(0.04; 0.08)
Weighted Mean	0.03	0.05	0.04
95% CI	(0.01; 0.05)	(0.03; 0.07)	(0.03; 0.05)
Q	159.69	147.03	325.03
I^2	66%	64%	68%
Precision			
Mean CI Width	0.34	0.23	0.30
Median CI Width	0.27	0.20	0.24
% effect size sig. > 0	18%	29%	23%
Mean MDES	0.24	0.17	0.21
Median MDES	0.19	0.15	0.17
Average Power	22%	25%	23%
Median Power	14%	21%	17%
Informativeness			
Bayes Factor			
% Uninformative	40%	39%	40%
% Supporting H0	40%	24%	38%
% Supporting Ha	20%	27%	23%
Median	0.50	0.67	0.56

Notes: Power was calculated assuming an effect size of 0.06. Bayes Factors were calculated by modelling the alternative hypothesis with a half normal distribution with mean 0 and SD 0.2.

Effect Sizes

There were 141 distinct trials. The total number of participants was 1,222,024 and the median number of participants per trials was 2386. Of these trials, 91 (65%) reported effect sizes above zero. Effect size estimates ranged from -0.16 to 0.74, with a median of 0.03. The unweighted mean of the effect size estimates was 0.06, 95% CI [0.04, 0.08]. This mean was the same for the EEF and the NCEE trials and was minimally sensitive to the way effect sizes were selected in trials with multiple outcomes (see Supplemental Material available online).

Heterogeneity was moderate but statistically significant ($Q = 325.03$, $df = 140$, $p < .001$; $I^2 = 68\%$), suggesting that the effect sizes varied in magnitude beyond that expected by chance. Considering that the trials were substantially different to one another (e.g., different topics, participants, outcome measures), this was to be expected. Based on a random-effects model, the mean of the weighted effect size was 0.04, 95% CI [0.03, 0.05].

Subgroup analyses. We measured how stable effect sizes were across age groups, topics of outcome measure, cost of the trial, year of publication, type of trial, and reported quality of the trial. We analysed EEF and NCEE trials independently because not all the variables were comparable between the two funders. Moreover, because some of the trials involved multiple age groups and/or topic of outcome measures, we conducted the analysis at the effect size level (i.e., effect sizes of trials with multiple outcomes were treated as independent). Subgroups including less than five effect sizes were excluded from the analysis. As seen in Tables 2 and 3, none of the moderators tested were significant, except type of trial in the EEF sample. Efficacy trials were associated with slightly larger effect sizes than effectiveness trials.

Table 2. *Analysis of the subgroups identified in the trials commissioned by the EEF.*

Subgroup	k	Mean	95% CI	Q	df(Q)	p-value
Topic						
Language: Reading	63	0.04	(0.01, 0.04)			
Mathematics	35	0.04	(0.02, 0.07)			
Language: General	20	0.03	(-0.01, 0.07)	8.89	4	0.064
Combination	10	0.00	(-0.02, 0.02)			
Language: Writing	8	0.13	(0.06, 0.21)			
Level						
Kindergarten	5	0.08	(0.00, 0.17)			
Elementary	86	0.04	(0.02, 0.05)	4.45	3	0.216
Secondary	36	0.03	(-0.01, 0.06)			
Elem & Sec	13	0.09	(0.00, 0.18)			
Type of trial						
Efficacy trial	117	0.05	(0.03, 0.07)	4.23	1	0.040
Effectiveness trial	23	0.01	(0.00, 0.02)			

Subgroup	k	Coefficient	Z	p-value
Year of publication				
2014 to 2018	140	-0.01	-1.70	0.090
Quality trial				
0 (low) to 5 (high)	139	0.00	-0.70	0.486
Cost intervention per pupil				
1 (low) to 5 (high)	140	0.00	0.55	0.584
Cost trial				
70K to 1.4M	140	0.00	-1.76	0.078

Table 3. *Analysis of the subgroups identified in the trials commissioned by the NCEE.*

Subgroup	k	Mean	95% CI	Q	df(Q)	p-value
Topic						
Language: Reading	61	0.04	(0.02, 0.06)			
Mathematics	39	0.04	(0.01, 0.06)	7.66	3	0.054
Language: General	17	0.01	(-0.03, 0.04)			
Combination	6	0.15	(0.07; 0.23)			
Level						
Kindergarten	10	0.01	(-0.06, 0.08)			
Elementary	73	0.06	(0.04, 0.08)	7.74	3	0.052
Secondary	24	0.03	(0.00, 0.06)			
Elem & Sec	22	0.03	(0.02, 0.04)			

Subgroup	k	Coefficient	Z	p-value
Year of publication				
2008 to 2018	131	0.00	-0.05	0.96

Precision of Effect Sizes

Using the standard error (SE_d), we computed the 95% confidence interval surrounding each observed effect. Descriptive statistics are shown in Table 1. On average, the width of the confidence intervals was 0.30 (median: 0.24). The average width was larger in EEF trials (0.34) than in NCEE trials (0.23). Again, these values were not substantially influenced by the way effect sizes were selected in trials with multiple outcomes.

Statistical significance and power. Given the size of the effects observed and the relatively low precision at which they were measured, few effects reached statistical significance. In total, 32 effect sizes (23%) were significantly greater than zero and 4 (3%) were significantly lower than zero. Using the standard error associated with each effect size (SE_d), we computed the smallest effect size that each trial could reliably detect – the Minimal Detectable Effect Size (MDES) – by multiplying each trial's SE_d by 2.80. This gave the effect size that the trial had an 80% chance of detecting, given an alpha of .05 (Alasuutari, Bickman, & Brannen, 2008). The average MDES was 0.21 SDs. As shown in Figure 3, for more than 93% of the trials the MDES was greater than the effect size observed.

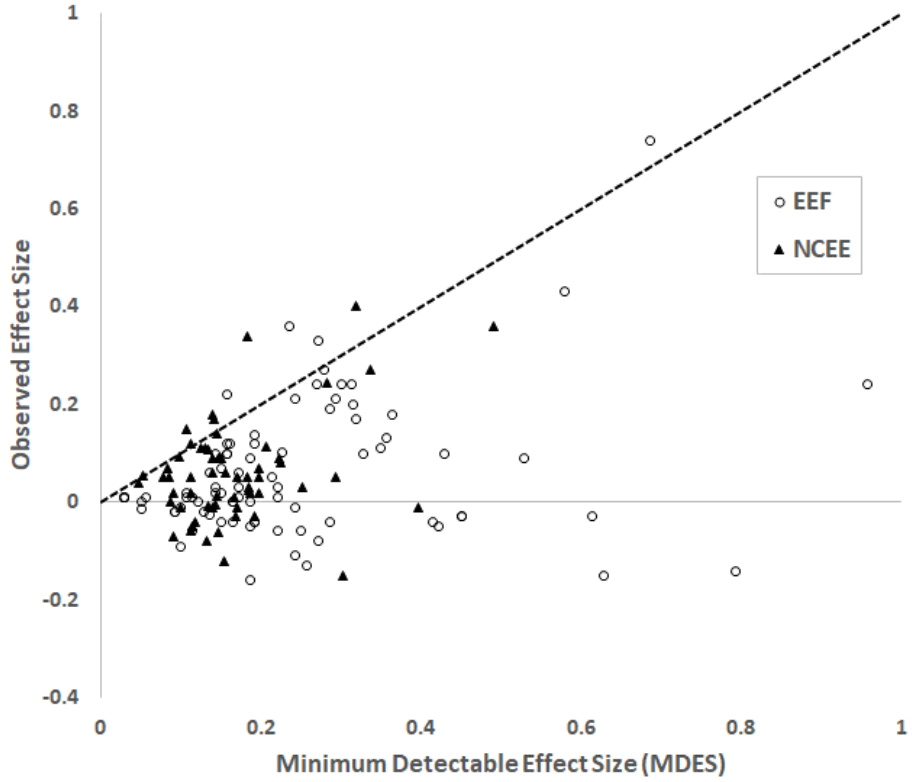


Figure 3. *A scatterplot showing the relation between the minimum detectable effect sizes (MDES) and observed effect sizes in the trials commissioned by the EEF and NCEE. The diagonal line represents obtained effect sizes equal to the MDES of the trial. Points below the diagonal represent obtained effect sizes below the MDES of the trial, points above the diagonal represent obtained effect sizes above the MDES of the trial.*

We also computed the statistical power that each trial had to detect an effect size of 0.06 – the mean effect size observed in our sample of trials (e.g., Cohen, 1988). On this method, the average power of the trials was 23% (median 17%), much lower than the commonly recommended 80%. Only 9 trials (6%) had at least 80% chance of detecting such an effect.

Bayes Factor

For each trial we calculated a Bayes factor, following the method suggested by Dienes, Coulton and Heather (2017). This quantified how likely the data were under the null hypothesis compared to the alternative hypothesis, which was defined to be an effect size taken from a half normal distribution with mean 0 and SD 0.2 (i.e., a distribution where effect sizes range from 0 to roughly 0.4, and where smaller effects are more likely than larger ones; our results were not highly sensitive to this choice, or to our choice of distribution; the Bayes factors associated with various different alternative hypotheses are given in the Supplemental Material available online). We interpreted the resulting Bayes factors, summarised in Table 1, following Jeffrey's (1961) guidelines. Many, 40%, fell between 3 and 1/3, indicating that the trial was uninformative; 38% were less than 1/3, indicating support for the null hypothesis (30% moderate, 7% strong, 0% very strong, and 0% decisive) and 23% were greater than 3, indicating support for the alternative hypothesis that the intervention is effective (13% moderate, 4% strong, 1% very strong, and 4% decisive). The overall median Bayes factor was 0.56.

Discussion

On average, the effect size of the rigorous large-scale RCTs commissioned by the EEF and NCEE was 0.06 SDs, much smaller than what is typically observed in the wider educational literature. The averaged effect size was even smaller when weighted by the precision of the estimates (0.04 SDs). By contrast, the confidence intervals of these effect sizes were comparatively large, on average 0.30 SDs wide. Consequently, many trials were uninformative: 40% of trials yielded Bayes factors between 3 and 1/3. These trials produced findings consistent with both the null hypothesis of no effect, and also with an effect comparable to that associated with one year of maturation and instruction (Bloom et al., 2008). Such trials neither allow us to conclude that an intervention should be implemented at scale, nor that this should be avoided to prevent the waste of public money.

For each of the trials in our sample the funding body felt that the intervention had promise. Why did so many of these trials fail to find unambiguous evidence of positive effects? In particular, why were the effect sizes found so much lower than the researchers expected, and the typical effect sizes found in the education literature? Our discussion centers around three broad, perhaps complementary, possibilities: (i) that many of the interventions studied are ineffective because the literature upon which they are based is unreliable, (ii) that many of the interventions studied are ineffective because they have been poorly designed or implemented; and (iii) that many of the interventions studied *are* effective, but that these

trials were not designed so that their effects could be reliably detected. We discuss each in turn.

One possibility is that the literature upon which educational interventions are based is unreliable. Recent developments, collectively referred to as the 'replication crisis', suggest that the psychological literature is not as robust as previously imagined (e.g., Open Science Collaboration, 2015). This is an issue that should particularly concern education researchers. Ioannidis (2005) has shown that if a scientific field ignores the importance of replication, a situation can arise where "most published research findings are false". This is worrying, as only 0.13% of articles in leading education journals report replication studies (Makel & Plucker, 2014). Equally, issues of *p*-hacking and other questionable research practices (e.g., John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011) seem to apply as much to education as to other areas of the psychological sciences. Interventions that are based on insights gained from unreliable basic research are unlikely to be effective, even if they are well designed, successfully implemented and appropriately trialed.

A second possibility is that the insights from basic research upon which the trials are based were not adequately translated into an effective intervention and/or successfully implemented. In education, basic research is generally developed in small, controlled settings, and often requires translation before being implemented in schools. This problem is compounded when trials are conducted at scale because an intervention implemented in many schools is less likely to be done so consistently. Unfortunately, as Burkhardt and Schoenfeld (2003) pointed out, the kind of translational work required to address this issue is undervalued by the research community, and therefore receives comparatively little attention or reward. Perhaps the reason that many EEF and NCEE trials failed to find unambiguously positive results is that the skills required to successfully translate insights from lab-based research into effective interventions that are possible to implement successfully are relatively rare, or perhaps not sufficient time or focus is devoted to this work.

A third possibility concerns the design of trials themselves. Educational RCTs are typically designed to have high external validity. Researchers achieve this by, among other things, conducting their trials in genuine educational settings and using real-world outcome measures that are often far removed from the intervention. For instance, the EEF's "increasing pupil motivation" trial evaluated whether providing financial incentives would improve motivation. The primary outcome measure was scores in a national examination, rather than a validated measure of motivation (Sibieta, Greaves, & Sianesi, 2014). This decision increased the external validity of the trial, but also increased the level of noise in the research design and reduced the range of plausible effect sizes (e.g., Baguley, 2009; Cheung

& Slavin, 2016). One plausible account for the relative lack of significant findings in many of these trials is that the interventions being studied do have positive effects, but the researchers underestimated the level of noise in their research designs, and therefore chose unrealistically high MDESs (cf. Norman, 2003). If this account is correct, many EEF and NCEE trials are inappropriately powered.

Implications

Determining which of these three accounts is correct (or, if each plays a role, which is the primary factor), is vitally important. Each account demands a different change to current practice.

The first account is simply that the basic research upon which educational interventions are based is unreliable. Two reforms could improve this situation. First, methodological improvements such as a greater emphasis on preregistration and data sharing would likely lead to a more reliable literature (e.g., Open Science, 2015; Simmons et al., 2011; SREE n.d.). Second, more care could be taken when assessing the reliability of existing insights. For instance, a direct replication of basic research could be required prior to an RCT being commissioned (the "goal" structure used by the NCER is an example of this approach; NCER, 2012). Alternatively, critical reviews of the wider literature might lead to some interventions to be questioned in advance of an RCT.

If our results can be explained by poor translation from basic research into effective practice, then the research community needs to devote more effort to the kind of engineering research advocated by Burkhardt and Schoenfeld (2003) by encouraging, for example, greater collaboration between researchers, educational designers and professional development providers.

Finally, if the interventions being trialed have positive effects, but for various reasons the ways that trials are currently designed are not capable of reliably detecting them, then methodological reform is necessary. Trials would need to be powered to much lower MDESs, perhaps even to lower than 0.05. Given existing resource constraints, it seems impractical to achieve this with larger samples (nearly 20,000 participants would be required for an independent-samples *t*-test to detect an effect size of 0.04 with 80% power); and larger samples do not in any event guarantee higher power (Weisburd, Petrosino, & Mason, 1993). Alternatively, the power of trials could be increased through other means, perhaps by focusing on more targeted subgroups of the population, using more targeted outcome measures, or having greater oversight from the developers (indeed, in line with this latter

point, we found that EEF efficacy trials produced slightly greater effect sizes than EEF effectiveness trials). These modifications would increase the power of trials, but might limit the external validity of their findings. However, this need not limit the usefulness of such research (Mook, 1983). To take the earlier example, the EEF's "increasing pupil motivation" trial could have used a validated measure of motivation as its primary outcome variable, rather than a national examination. Arguably, using a more targeted outcome measure in this fashion, coupled with a reliance on the theoretically well-established causal link between self-motivation and attainment (e.g., Zimmerman, Bandura, & Martinez-Pons, 1992), would have increased the power of the trial without necessarily affecting its cost or usefulness. Such an approach would, however, have the unfortunate consequence of making it difficult to legitimately compare effect sizes between trials that use different outcome measures (Baguley, 2009).

It has only been possible to conduct the analysis reported in this paper because of the extremely high methodological standards adopted by the EEF and NCEE. Specifically, both funding bodies require analysis plans to be preregistered, and all results to be published. This gives us confidence that EEF and NCEE trials are not affected by either the selective reporting of analyses or by publication bias. This is not true for large-scale educational RCTs in general. Had we conducted our analysis on the wider literature, we may have found that a larger proportion of (published) RCTs are informative. However, such a finding would likely be misleading due to the so-called 'winner's curse', the observation that those papers which make it through the review process typically overestimate effect sizes (Young, Ioannidis & Al-Ubaydli, 2008). Without being able to study an unbiased sample of trials – including those which did not find significant effects – it would not be possible to accurately estimate the proportion that are informative. This observation reinforces the need for the level of rigor insisted upon by the EFF and NCEE.

Given the significant level of educational research funding currently being spent on rigorous large-scale RCTs, it is clearly unsatisfactory that so many trials are uninformative. Understanding why educational RCTs often yield small and uninformative effects should be seen as a priority for our field.

References

Alasuutari, P., Bickman, L., & Brannen, J. (Eds.). (2008). *The SAGE handbook of social research methods*. London, UK: Sage. doi: 10.4135/9781446212165

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*, 603-617.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological), 57*, 289-300.

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

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Burkhardt, H., & Schoenfeld, A. H. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational researcher*, *32* (9), 3-14.

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hilsdale. *NJ: Lawrence Earlbaum Associates*.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on?. *Perspectives on Psychological Science, 6(3),* 274-290.

Dienes, Z., Coulton, S., & Heather, N. (2018). Using Bayes factors to evaluate evidence for no effect: examples from the SIPS project. *Addiction, 113(2),* 240-246.

Education Endowment Foundation (EEF) (2015a). *Annual Report 2014/15*. London: Education Endowment Foundation.

Education Endowment Foundation (EEF) (2015b). *EEF Guidance on Cost Evaluation*. London: Education Endowment Foundation.

Education Endowment Foundation (EEF) (2016). *Classification of the security of findings from EEF evaluations*. London: Education Endowment Foundation.

Education Endowment Foundation (EEF) (2017). *EEF standards for independent evaluation panel members.* London: Education Endowment Foundation.

Hattie, J.A.C. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement.* Oxford, UK: Routledge.

Hill, C. J., Bloom, H. S., Black, A. R., and Lipsey, M. W., (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives, 2* (3), 172-177.

Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA, Cochrane Bias Methods Group., (2011). Cochrane Statistical Methods Group: The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ: British Medical Journal, 343*

Ioannidis J. P. A. (2005). Why most published research findings are false. *PLoS Med.* (8):e124.

Jeffreys, H. (1961). *Theory of probability (3rd ed.).* Oxford, UK: Oxford University Press

John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524-532.

Karlsson, P., & Bergmark, A. (2015). Compared with what? An analysis of control-group types in Cochrane and Campbell reviews of psychosocial treatment efficacy with substance use disorders. *Addiction*, *110* (3), 420-428.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. Thousand Oaks, CA: Sage.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*, 304-316.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*, 379-387.

National Center for Education Evaluation and Regional Assistance (NCEE) (2013). *NCEE Guidance for REL Study Proposals, Reports, and Other Products.* Retrieved from https://ies.ed.gov/ncee/edlabs/relresources/pdf/NCEE_Guidance_for_REL_Products_042013.pdf

National Center for Education Evaluation and Regional Assistance (NCEE) (2017). *Evaluation Principles and Practices.* Retrieved from https://ies.ed.gov/ncee/projects/pdf/IESEvaluationPrinciplesandPractices_011117.pdf

National Center for Education Research (NCER) (2012). *2012 National Board for Education Sciences Annual Report Briefing Material for Board Members*. Retrieved from https://ies.ed.gov/director/board/briefing/ncer_structure.asp

Norman, G. (2003). RCT=results confounded and trivial: the perils of grand educational experiments. *Medical Education, 37,* 582-584.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. Science, 349, aac4716.

Pocock S. J. (1983) *Clinical trials: A practical approach*. Chichester, England: Wiley.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Sibbald, B., & Roland, M. (1998). Understanding controlled trials. Why are randomised controlled trials important? *BMJ: British Medical Journal*, *316* (7126), 201.

Sibieta, L., Greaves, E., & Sianesi, B. (2014). *Increasing Pupil Motivation: 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,* 1359-1366.

Society for Research on Educational Effectiveness (SREE). (n.d.). *Registry of efficacy and effectiveness studies.* Retrieved from https://www.sree.org/pages/registry.php

Weisburd, D., Petrosino, A., & Mason, G. (1993). Design sensitivity in criminal justice experiments. *Crime and justice, 17*, 337-379.

Young, N. S., Ioannidis, J. P., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS medicine, 5(10),* e201.

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal, 29*, 663-676.

Figure S.1. *Effect sizes and confidence intervals of EEF trials.*

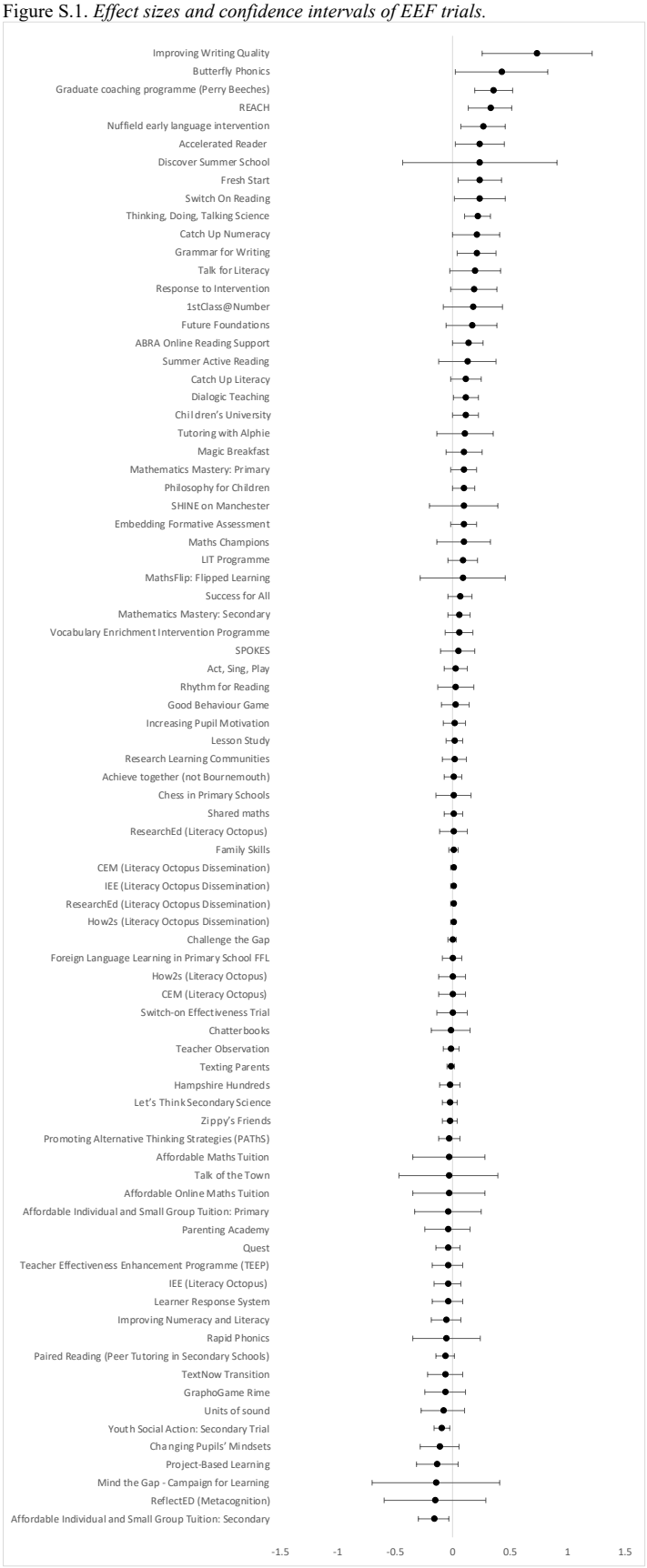
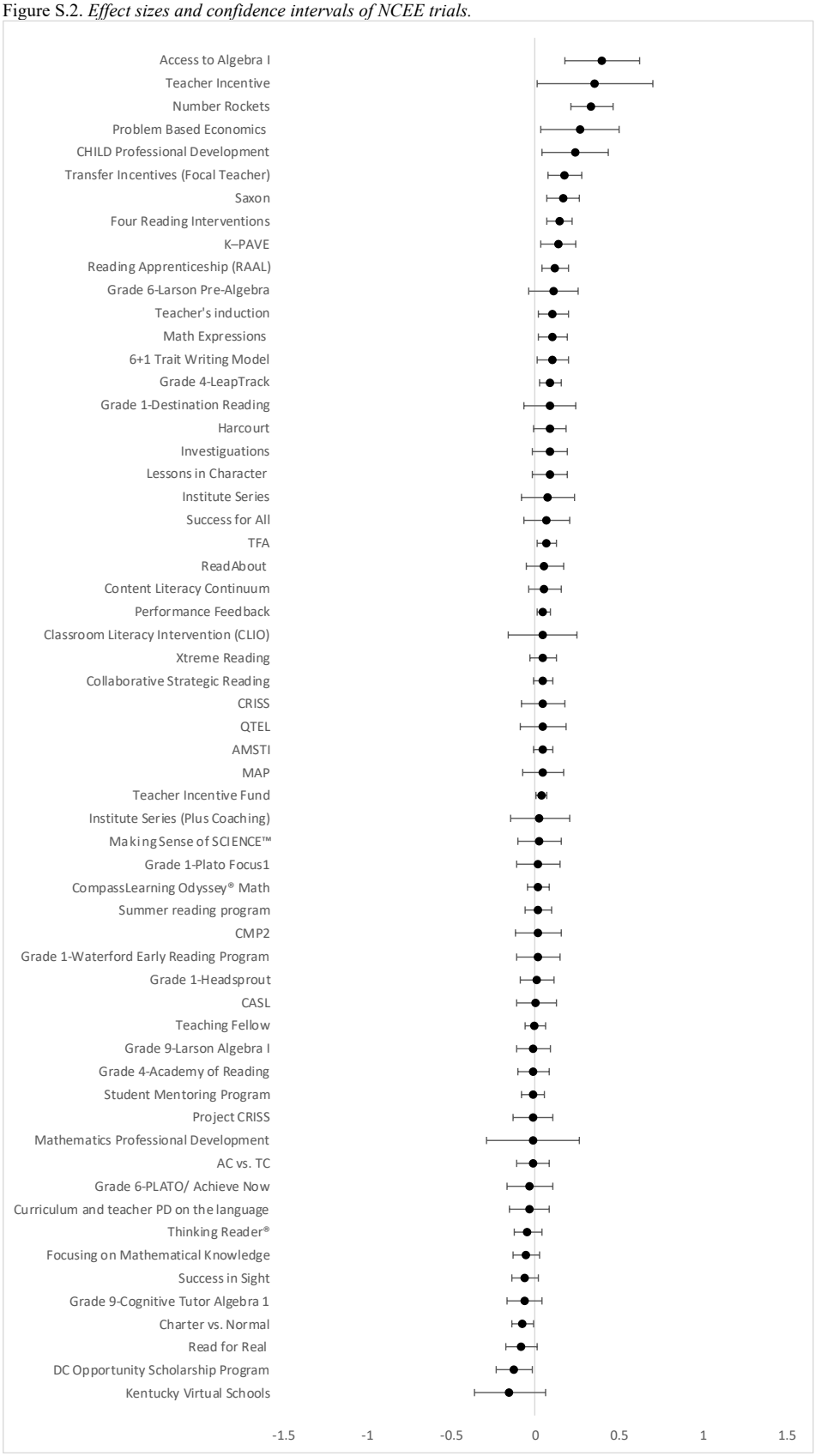


Figure S.2. *Effect sizes and confidence intervals of NCEE trials.*



Supplementary material for:

Lortie-Forgues, H., & Inglis, M. (Accepted/In press). Rigorous Large-Scale Educational RCTs are Often Uninformative: Should We Be Concerned? *Educational Researcher*.

Note:

Raw data and R source code can be downloaded at: https://doi.org/10.6084/m9.figshare.c.4421087

Table S.1
Variables and effect sizes of EEF trials.

Intervention	Year	Total N	Total Cost (£)	Cost per pupil	Topic	Level	Subgroup	ES	SEd	Bayes Factor	Strength
1stClass@Number	2018	532	287500.00	1	Mathematics	Elementary		0.18	0.13	1.88	4
ABRA Online Reading Support	2016	2241	643,467.00	1	Language: Reading	Elementary	Non-ICT	0.23	0.07	80.92	4
					Language: Reading	Elementary	ICT	0.14	0.07	3.81	4
Accelerated Reader	2015	349	147,000.00	1	Language: Reading	Secondary		0.24	0.11	5.83	3
Achieve together (not Bournemouth)	2017	8,581	437,831.00	1	Combination	Secondary		0.01	0.04	0.24	3
Act, Sing, Play	2015	909	415,000.00	2	Language: Reading	Elementary		0.03	0.05	0.41	4
					Mathematics	Elementary		0.00	0.05	0.24	4
Affordable Individual and Small Group Tuition: Primary	2015	2,519	78,600.00	3	Mathematics	Elem & Sec	Maths	-0.04	0.15	0.51	0
					Language: General	Elem & Sec	English	-0.08	0.13	0.38	0
Affordable Individual and Small Group Tuition: Secondary	2015	125,968	185,000.00	3	Language: General	Secondary	English	-0.16	0.07	0.10	0
					Mathematics	Secondary	Maths	0.05	0.03	1.10	0
Affordable Maths Tuition	2016	600	196,499.00	3	Mathematics	Elementary		-0.03	0.16	0.56	3
Affordable Online Maths Tuition	2016	600	196499.00	3	Mathematics	Elementary		-0.03	0.16	0.56	3
Butterfly Phonics	2015	370	457,980.00	2	Language: Reading	Secondary		0.43	0.21	3.62	0
Catch Up Literacy	2015	631	430,000.00	4	Language: Reading	Elem & Sec		0.12	0.07	2.32	4
Catch Up Numeracy	2014	224	71,000.00	1	Mathematics	Elementary		0.21	0.10	5.06	3
Challenge the Gap	2017	40,025	961,778.00	1	Combination	Elementary	KS2	0.00	0.02	0.10	2
					Combination	Secondary	KS4	-0.01	0.01	0.03	1
Changing Pupils' Mindsets	2015	1791	368,460.00	3	Language: General	Elementary	Pupil workshops	0.18	0.10	3.09	2
	2015		368,460.00	3	Mathematics	Elementary	Pupil workshops	0.10	0.09	1.16	2
	2015		368,460.00	1	Mathematics	Elementary	Teacher training	0.01	0.10	0.48	3
	2015		368,460.00	1	Language: General	Elementary	Teacher training	-0.11	0.09	0.20	3
Chatterbooks	2014	577	397,314.00	1	Language: Reading	Secondary	Chatterbooks Plus	-0.01	0.09	0.38	3
					Language: Reading	Secondary	Chatterbooks	-0.14	0.09	0.17	3
Chess in Primary Schools	2016	4009	689,150.00	1	Mathematics	Elementary		0.01	0.08	0.41	5
Children's University	2017	2603	559862.00	1	Language: Reading	Elementary		0.12	0.06	3.50	2
					Mathematics	Elementary		0.15	0.06	10.02	3
Dialogic Teaching	2017	4958	499,485.00	1	Mathematics	Elementary		0.09	0.06	1.49	3
					Language: General	Elementary		0.15	0.08	3.24	3
					Sciences	Elementary		0.12	0.06	3.50	3
Discover Summer School	2014	124	240,652.00	5	Language: Writing	Elem & Sec		0.24	0.34	1.18	0
					Language: Reading	Elem & Sec		0.21	0.32	1.15	0
Embedding Formative Assessment	2018	25000	489602.00	1	Combination	Secondary		0.10	0.06	1.94	5
Family Skills	2018	1,985	556940.00	1	Language: General	Kindergarten		0.01	0.02	0.16	5
Foreign Language Learning in Primary School FFL	2017	4967	200,000.00	1	Language: General	Elementary		0.00	0.04	0.20	1
Fresh Start	2015	433	350,000.00	2	Language: Reading	Elem & Sec		0.24	0.10	8.81	3
Future Foundations	2014	435	543,000.00	5	Language: General	Elementary	English	0.17	0.11	2.20	2
					Mathematics	Elementary	Maths	0.00	0.11	0.48	2
Good Behaviour Game	2018	3084	820000.00	1	Language: Reading	Elementary		0.03	0.06	0.44	4
Graduate coaching programme (Perry Beeches)	2015	373	306,000.00	5	Language: General	Elem & Sec		0.36	0.08	4587.45	3

Grammar for Writing	2014	2,394	339,000.00	1	Language: Writing	Elementary	Class level vs. control	0.10	0.10	1.09	3
					Language: Writing	Elementary	Class level vs. control no small group	0.06	0.11	0.74	3
					Language: Writing	Elementary	Small group vs. control	0.24	0.13	3.39	3
					Language: Writing	Elementary	Small group vs. whole group	0.21	0.09	7.76	3
GraphoGame Rime	2018	398	364527.00	1	Language: Reading	Elementary		-0.06	0.09	0.27	5
Hampshire Hundreds	2014	1440	121,000.00	2	Combination	Elementary	Disadvantaged pupils	0.03	0.06	0.44	2
					Combination	Elementary	Other pupils	-0.02	0.05	0.18	2
Improving Numeracy and Literacy	2015	3189	489,471.00	1	Language: Reading	Elementary		-0.05	0.07	0.21	5
					Mathematics	Elementary		0.20	0.09	6.26	5
Improving Writing Quality	2014	845	395,850.00	1	Language: Writing	Elem & Sec		0.74	0.24	10.52	2
Increasing Pupil Motivation	2014	15,710	1,100,000.00	3	Mathematics	Secondary	Event incentives	0.08	0.05	1.52	2
					Mathematics	Secondary	Financial incentives	0.04	0.05	0.51	2
					Language: General	Secondary	Event incentives	0.04	0.06	0.52	2
					Language: General	Secondary	Financial incentives	0.02	0.05	0.34	2
Learner Response System	2017	6500	1013991.00	1	Mathematics	Elementary	Maths (cohort A)	0.00	0.07	0.33	3
					Language: Reading	Elementary	Reading (cohort A)	0.00	0.06	0.29	3
					Mathematics	Elementary	Maths (cohort B)	-0.08	0.07	0.17	5
					Language: Reading	Elementary	Reading (cohort B)	-0.04	0.07	0.23	5
Lesson Study	2017	12,700	543425.00	1	Combination	Elementary	1 Year of Lesson	0.02	0.04	0.30	5
					Combination	Elementary	2 Year of Lesson	0.03	0.05	0.41	5
Let's Think Secondary Science	2016	8000	639,485.00	1	Sciences	Secondary		-0.02	0.03	0.09	3
LIT Programme	2014	5,565	310,000.00	1	Language: Reading	Secondary		0.09	0.07	1.22	1
Literacy Octopus Dissemination: CEM	2017	88,088		1	Language: General	Elementary	Passive Arm vs control	0.01	0.01	0.14	5
Literacy Octopus Dissemination: How2s	2017	87,701		1	Language: General	Elementary	Passive Arm vs control	0.01	0.01	0.14	5
Literacy Octopus Dissemination: IEE	2017	86,742		1	Language: General	Elementary	Passive Arm vs control	0.01	0.01	0.14	5
Literacy Octopus Dissemination: ResearchEd	2017	86,155		1	Language: General	Elementary	Passive Arm vs control	0.01	0.01	0.14	5
Literacy Octopus: CEM	2017	2,174		1	Language: Reading	Elementary	Passive Arm vs control	0.00	0.06	0.29	5
Literacy Octopus: CEM	2017	2,080		1	Language: Reading	Elementary	Active Light Arm vs control	0.03	0.06	0.44	5
Literacy Octopus: CEM	2017	2,386		1	Language: Reading	Elementary	Active Arm vs control	0.03	0.06	0.44	5
Literacy Octopus: How2s	2017	2,337		1	Language: Reading	Elementary	Passive Arm vs control	0.00	0.06	0.29	5
Literacy Octopus: How2s	2017	2,448		1	Language: Reading	Elementary	Active Arm vs control	-0.03	0.06	0.20	5
Literacy Octopus: IEE	2017	2,291		1	Language: Reading	Elementary	Passive Arm vs control	-0.02	0.06	0.23	5
Literacy Octopus: IEE	2017	2,203		1	Language: Reading	Elementary	Active Arm vs control	-0.04	0.06	0.18	5
Literacy Octopus: ResearchEd	2017	2,474		1	Language: Reading	Elementary	Passive Arm vs control	0.00	0.06	0.29	5
Literacy Octopus: ResearchEd	2017	2,122		1	Language: Reading	Elementary	Active Arm vs control	0.01	0.06	0.33	5
Magic Breakfast	2016	8,841	425,967.00	1	Language: Reading	Elementary		0.10	0.08	1.28	4
					Mathematics	Elementary		0.08	0.07	1.02	4
					Mathematics	Elementary		0.15	0.05	33.45	4
					Language: Writing	Elementary		0.14	0.05	19.35	4
					Language: Reading	Elementary		0.10	0.05	3.10	4
Mathematics Mastery: Primary	2015	5,108	600,000.00	2	Mathematics	Elementary		0.10	0.06	1.94	3
Mathematics Mastery: Secondary	2015	7,712	174,000.00	1	Mathematics	Secondary		0.06	0.05	0.84	4
Maths Champions	2018	628	380000.00	1	Mathematics	Kindergarten		0.10	0.12	1.01	2

MathsFlip: Flipped Learning	2017	1100	890080.00	2	Mathematics	Elementary		0.09	0.19	0.93	3
Mind the Gap - Campaign for Learning	2014	1177	550,000.00	2	Combination	Elementary		-0.14	0.28	0.65	1
Nuffield early language intervention	2016	394	736,546.00	1	Language: General	Kindergarten	20 weeks	0.16	0.09	2.89	4
					Language: General	Kindergarten	30 weeks	0.27	0.10	16.39	4
Paired Reading (Peer Tutoring in Secondary Schools)	2015	2736	520,064.00	1	Language: Reading	Secondary		-0.02	0.07	0.27	4
					Language: Reading	Secondary		-0.06	0.04	0.08	4
Parenting Academy	2016	2153	991,400.00	3	Language: Reading	Elem & Sec	Unincentivised	0.02	0.09	0.49	4
					Mathematics	Elementary	Incentivised	0.01	0.11	0.51	4
					Language: Reading	Elem & Sec	Incentivised	0.00	0.09	0.41	4
					Mathematics	Elem & Sec	Unincentivised	-0.04	0.10	0.34	4
Philosophy for Children	2015	3159	272,000.00	1	Language: Writing	Elem & Sec		0.03	0.05	0.41	3
					Language: Reading	Elementary		0.12	0.05	7.22	3
					Mathematics	Elementary		0.10	0.05	3.10	3
Project-Based Learning	2016	4,074	906,000.00	1	Language: Reading	Secondary		-0.13	0.09	0.18	1
Promoting Alternative Thinking Strategies (PAThS)	2015	3336	90,000.00	1	Mathematics	Elementary	Year 5	0.03	0.05	0.41	4
					Mathematics	Elementary	Year 6	-0.03	0.05	0.16	4
					Language: Reading	Elementary	Year 5	-0.03	0.05	0.16	4
					Language: Reading	Elementary	Year 6	-0.11	0.05	0.08	4
Quest	2015	3,641	572,832.00	2	Language: Reading	Secondary		-0.04	0.05	0.14	1
Rapid Phonics	2015	201	148,000.00	3	Language: Reading	Secondary		-0.05	0.15	0.49	3
REACH	2016	287	525,000.00	1	Language: Reading	Secondary	Reading + Comprehension	0.51	0.09	516200.92	2
					Language: Reading	Secondary	Reading	0.33	0.10	69.61	2
ReflectED (Metacognition)	2016	1858	253,000.00	1	Mathematics	Elementary		0.30	0.17	2.91	4
					Language: Reading	Elementary		-0.15	0.22	0.53	4
Research Learning Communities	2017	5462	237000.00	1	Language: Reading	Elementary		0.02	0.05	0.34	5
Response to Intervention	2014	517	496,000.00	2	Language: Reading	Elementary		0.19	0.10	3.62	1
					Language: Reading	Elementary		-0.09	0.13	0.36	
Rhythm for Reading	2014	419	78,755.00	1	Language: Reading	Secondary		0.03	0.08	0.50	3
Shared maths	2015	6,472	766,945.00	1	Mathematics	Elementary		0.02	0.04	0.30	4
					Mathematics	Elementary		0.01	0.04	0.24	4
SHINE on Manchester	2016	1376	510,175.00	4	Language: Reading	Elementary	SHINE vs control (Year 3)	0.10	0.15	0.97	3
					Language: Reading	Elementary	SHINE vs control (Year 1)	0.03	0.16	0.70	3
					Language: Reading	Elementary	SHINE vs control (Year 2)	-0.10	0.08	0.18	3
SPOKES	2016	808	1,000,000.00	4	Language: Reading	Elementary		0.08	0.08	0.94	3
					Language: Reading	Elementary		0.05	0.08	0.63	3
					Language: Reading	Elementary		0.03	0.08	0.50	3
Success for All	2017	1767	1,410,000.00	1	Language: Reading	Elementary	End of Year 1	0.07	0.05	1.11	3
					Language: Reading	Kindergarten	End of Reception	0.04	0.05	0.51	3
Summer Active Reading	2014	205	218,414.00	2	Language: Reading	Elementary		0.13	0.13	1.24	3
Switch On Reading	2014	314	70,575.00	3	Language: Reading	Secondary		0.24	0.11	5.83	3
Switch-on Effectiveness Trial	2017	999	670000.00	2	Language: Reading	Elementary		0.00	0.07	0.33	4
Talk for Literacy	2015	236	148,110.00	1	Language: Reading	Secondary		0.20	0.11	3.24	4
Talk of the Town	2016	3,299	967,780.00	1	Language: Reading	Elementary	Treatment vs. control, NGRT	-0.03	0.22	0.69	4
Teacher Effectiveness Enhancement Programme (TEEP)	2016	13,990	997,000.00	1	Mathematics	Secondary		-0.02	0.06	0.23	3

					Language: General	Secondary		-0.04	0.07	0.23	3
Teacher Observation	2017	14,100	1180000.00	1	Combination	Secondary		-0.01	0.04	0.16	5
Texting Parents	2016	19,298	532,620.00	1	Sciences	Secondary		-0.01	0.02	0.07	3
					Mathematics	Secondary		0.07	0.01	4103348604	3
					Language: General	Secondary		0.03	0.02	0.56	3
TextNow Transition	2014	501	480,953.00	2	Language: Reading	Elem & Sec		-0.06	0.08	0.23	3
Thinking, Doing, Talking Science	2015	1513	270,000.00	1	Sciences	Elementary		0.22	0.06	273.98	3
Tutoring with Alphie	2015	248	153,280.00	3	Language: Reading	Elementary		0.11	0.13	1.07	0
Units of sound	2015	786	390,206.00	3	Language: Reading	Secondary		-0.08	0.10	0.27	1
Vocabulary Enrichment Intervention Programme	2014	649	393,570.00	2	Language: Reading	Secondary		0.06	0.06	0.76	4
Youth Social Action: Secondary Trial	2016	7,781	676,142.00	2	Language: General	Secondary		-0.09	0.04	0.06	2
					Mathematics	Secondary		-0.09	0.04	0.06	2
Zippy's Friends	2018	3904	190000.00	1	Language: Reading	Elementary		-0.02	0.03	0.09	2

Notes: Bayes Factors were calculated by modelling the alternative hypothesis with a half normal distribution with mean 0 and SD 0.2.

Table S.2
Variables and effect sizes of NCEE trials.

NCEE Report ID	Intervention	Year	Total N	Topic	Level	Subgroup	ES	SEd	Bayes Factor
NCEE 2008-4013	Four Reading Interventions	2008	729	Language: Reading	Elementary	Word Attack (Grade 3)	0.36	0.07	86407.34
				Language: Reading	Elementary	TOWRE PDE (Grade 3)	0.26	0.06	3164.43
				Language: Reading	Elementary	Word identification (Grade 3)	0.15	0.04	338.59
				Language: Reading	Elementary	TOWRE SWE (Grade 3)	0.11	0.03	212.52
				Language: Reading	Elementary	AIMSweb (Grade 3)	0.14	0.07	3.81
				Language: Reading	Elementary	Passage Comprehension (Grade 3)	0.14	0.06	6.89
				Language: Reading	Elementary	GRADE (Grade 3)	0.06	0.11	0.74
				Language: Reading	Elementary	Word Attack (Grade 5)	0.18	0.04	6632.06
				Language: Reading	Elementary	TOWRE PDE (Grade 5)	0.11	0.06	2.58
				Language: Reading	Elementary	Word identification (Grade 5)	-0.04	0.05	0.14
				Language: Reading	Elementary	TOWRE SWE (Grade 5)	0.09	0.06	1.49
				Language: Reading	Elementary	AIMSweb (Grade 5)	-0.08	0.04	0.07
				Language: Reading	Elementary	Passage Comprehension (Grade 5)	-0.08	0.06	0.13
				Language: Reading	Elementary	GRADE (Grade 5)	0.05	0.07	0.62
NCEE 2008-4028	Classroom Literacy Intervention (CLIO)	2008	2790	Language: General	Kindergarten	Expressive language: English	-0.11	0.07	0.14
				Language: General	Kindergarten	Expressive language: Spanish	0.05	0.10	0.66
				Language: General	Kindergarten	Phonological awareness: Blending	-0.13	0.07	0.12
				Language: General	Kindergarten	Phonological awareness: Elision	0.00	0.07	0.33
				Language: General	Kindergarten	Print knowledge	0.05	0.07	0.62
				Language: General	Kindergarten	Receptive vocabulary	-0.09	0.06	0.12
				Language: General	Kindergarten	Syntax and grammar	-0.08	0.06	0.13
NCEE 2008-4030	Institute Series	2008	2620	Language: Reading	Elementary		0.08	0.08	0.94
NCEE 2008-4030	Institute Series (Plus Coaching)	2008	2436	Language: Reading	Elementary		0.03	0.09	0.53
NCEE 2009-4041	Grade 9-Cognitive Tutor Algebra 1	2009	755	Mathematics	Secondary		-0.06	0.05	0.12
NCEE 2009-4041	Grade 9-Larson Algebra I	2009	1204	Mathematics	Secondary		0.00	0.05	0.24
NCEE 2009-4041	Grade 1-Destination Reading	2009	742	Language: Reading	Elementary		0.09	0.08	1.09
NCEE 2009-4041	Grade 1-Headsprout	2009	1079	Language: Reading	Elementary		0.01	0.05	0.28
NCEE 2009-4041	Grade 1-Plato Focus1	2009	618	Language: Reading	Elementary		0.02	0.07	0.42
NCEE 2009-4041	Grade 1-Waterford Early Reading Program	2009	1155	Language: Reading	Elementary		0.02	0.07	0.42
NCEE 2009-4041	Grade 4-Academy of Reading	2009	899	Language: Reading	Elementary		-0.01	0.05	0.21
NCEE 2009-4041	Grade 4-LeapTrack	2009	1274	Language: Reading	Elementary		0.09	0.03	24.15

NCEE 2009-4041	Grade 6-Larson Pre-Algebra	2009	2588	Mathematics	Elementary		0.11	0.07	1.85
NCEE 2009-4041	Grade 6-PLATO/ Achieve Now	2009	1037	Mathematics	Elementary		-0.03	0.07	0.25
NCEE 2009-4043	AC vs. TC	2009	2600	Language: Reading	Kinder & Elem		-0.01	0.05	0.21
				Mathematics	Kinder & Elem		-0.05	0.03	0.06
NCEE 2009-4047	Student Mentoring Program	2009	2360	Mathematics	Elem & Sec	Math	-0.05	0.04	0.09
				Mathematics	Elem & Sec	Math—Percent Proficient	-0.03	0.04	0.12
				Language: Reading	Elem & Sec	Reading/ELA	-0.04	0.05	0.14
				Language: Reading	Elem & Sec	Reading/ELA—Percent Proficient	-0.03	0.04	0.12
				Sciences	Elem & Sec		-0.03	0.04	0.12
				Social Studies	Elem & Sec		-0.01	0.04	0.16
NCEE 2009-4068	CompassLearning Odyssey® Math	2009	2446	Mathematics	Elementary		0.02	0.03	0.27
NCEE 2009-4077	Harcourt	2009	1936	Mathematics	Elementary	Harcourt: Cohort 1	0.09	0.04	4.41
				Mathematics	Elementary	Harcourt: Cohort 2	0.09	0.05	2.14
NCEE 2009-4077	Success for All	2009	1531	Language: Reading	Elementary	SAT 10 reading (cohort 1)	-0.08	0.04	0.07
				Language: Reading	Elementary	DIBEL Fluency (cohort 1)	0.05	0.06	0.62
				Language: Reading	Elementary	DIBEL Nonsense Fluency (cohort 1)	0.07	0.07	0.85
				Language: Reading	Elementary	SAT 10 reading (cohort 2)	0.01	0.07	0.37
				Language: Reading	Elementary	DIBEL Fluency (cohort 2)	0.09	0.07	1.22
				Language: Reading	Elementary	DIBEL Nonsense Fluency (cohort 2)	0.14	0.09	2.07
NCEE 2010-4014	K–PAVE	2010	1296	Combination	Kindergarten	Academic knowledge	0.14	0.06	6.89
				Language: General	Kindergarten	Listening comprehension	0.11	0.07	1.85
				Language: General	Kindergarten	Expressive vocabulary	0.14	0.05	19.35
NCEE 2010-4015	Project CRISS	2010	3372	Language: Reading	Elementary	Cohort 1	-0.01	0.06	0.25
				Language: Reading	Elementary	Cohort 2	0.00	0.04	0.20
NCEE 2010-4015	Read for Real	2010	3188	Language: Reading	Elementary	Cohort 1	-0.08	0.05	0.10
				Language: Reading	Elementary	Cohort 2	-0.02	0.07	0.27
NCEE 2010-4015	ReadAbout	2010	3298	Language: Reading	Elementary	Cohort 1	-0.04	0.05	0.14
				Language: Reading	Elementary	Cohort 2	0.06	0.06	0.76
NCEE 2010-4021	Reading Apprenticeship (RAAL)	2010	2255	Language: Reading	Secondary		0.12	0.04	29.65
NCEE 2010-4021	Xtreme Reading	2010	2329	Language: Reading	Secondary		0.05	0.04	0.74
NCEE 2010-4022rev	Problem Based Economics	2010	3752	Economics	Secondary	Economic Content	0.32	0.13	8.99
				Economics	Secondary	Economic Problem Solving	0.27	0.12	6.44
NCEE 2010-4027	Teacher's induction	2010	1690	Language: Reading	Elementary	Two-year districts	0.11	0.05	4.65

				Mathematics	Elementary	Two-year districts	0.20	0.05	903.16
				Language: Reading	Elementary	One-year districts	0.01	0.04	0.24
				Mathematics	Elementary	One-year districts	-0.10	0.06	0.11
NCEE 2010-4029	Charter vs. Normal	2010	2150	Language: Reading	Secondary		-0.07	0.03	0.04
				Mathematics	Secondary		-0.06	0.04	0.08
NCEE 2010-4035	Thinking Reader®	2010	2147	Language: Reading	Elementary	Vocabulary	-0.04	0.04	0.10
				Language: Reading	Elementary	Comprehension	0.03	0.06	0.44
NCEE 2011-4001 (maths)	Investiguations	2011	2634	Mathematics	Elementary	Grade 1	0.00	0.04	0.20
				Mathematics	Elementary	Grade 2	0.09	0.05	2.14
NCEE 2011-4001 (maths)	Math Expressions	2011	2729	Mathematics	Elementary	Grade 1	0.11	0.04	14.83
				Mathematics	Elementary	Grade 2	0.12	0.05	7.22
NCEE 2011-4001 (maths)	Saxon	2011	2698	Mathematics	Elementary	Grade 1	0.07	0.05	1.11
				Mathematics	Elementary	Grade 2	0.17	0.05	111.73
NCEE 2011-4001 (reading)	Collaborative Strategic Reading	2011	1355	Language: Reading	Elementary		0.05	0.03	1.10
NCEE 2011-4005	CASL	2011	9596	Mathematics	Elementary		0.01	0.06	0.33
NCEE 2011-4007	CRISS	2011	4959	Language: Reading	Secondary		0.05	0.07	0.62
NCEE 2011-4024	Mathematics Professional Development	2011	2132	Mathematics	Secondary		-0.01	0.14	0.55
NCEE 2012-4002	Making Sense of SCIENCE™	2012	5130	Sciences	Secondary	ATLAST	0.11	0.05	4.65
				Sciences	Secondary	California Standards Test	0.03	0.07	0.47
NCEE 2012-4004	Lessons in Character	2012	4683	Combination	Elementary		0.08	0.08	0.94
				Language: General	Elementary		0.09	0.05	2.14
NCEE 2012-4005	QTEL	2012	36017	Language: General	Secondary	Grade 7: CELDT	0.05	0.07	0.62
				Language: General	Secondary	Grade 8: CELDT	0.03	0.08	0.50
				Language: General	Secondary	Grade 7: CST-ELA	-0.01	0.03	0.12
				Language: General	Secondary	Grade 8: CST-ELA	0.01	0.05	0.28
				Language: General	Secondary	Grade 7 (Low): CST-ELA	0.03	0.06	0.44
				Language: General	Secondary	Grade 8 (Low): CST-ELA	0.01	0.04	0.24
NCEE 2012-4006	Summer reading program	2012	1571	Language: Reading	Elementary		0.02	0.04	0.30
NCEE 2012-4007	Number Rockets	2012	994	Mathematics	Elementary		0.34	0.07	24201.18
NCEE 2012-4008	AMSTI	2012	18713	Mathematics	Elem & Sec		0.05	0.02	4.36
				Sciences	Elem & Sec		0.05	0.03	1.10
NCEE 2012-4010	6+1 Trait Writing Model	2012	4134	Language: Writing	Elementary		0.11	0.05	4.65
NCEE 2012-4013	Curriculum and teacher PD on the language	2012	2612	Language: General	Elementary		-0.03	0.06	0.20

NCEE 2012-4014	Success in Sight	2012	8213	Language: Reading	Elementary		-0.01	0.03	0.12
				Mathematics	Elementary		-0.06	0.04	0.08
NCEE 2012-4017	CMP2	2012	5677	Mathematics	Elementary		0.02	0.07	0.42
NCEE 2012-4020	Kentucky Virtual Schools	2012	5864	Mathematics	Secondary		-0.15	0.11	0.23
NCEE 2012-4021	Access to Algebra I	2012	440	Mathematics	Secondary		0.40	0.11	154.24
NCEE 2013-4000	MAP	2013	1914	Language: Reading	Elementary	ISAT reading scale score	0.05	0.06	0.62
				Language: Reading	Elementary	MAP composite score	0.07	0.06	0.93
NCEE 2013-4001	Content Literacy Continuum	2013	9557	Language: Reading	Secondary	Grade 10	0.10	0.07	1.49
				Language: Reading	Secondary	Grade 9	0.06	0.05	0.84
NCEE 2013-4002	CHILD Professional Development	2013	3078	Language: Reading	Elementary		0.24	0.10	8.81
NCEE 2013-4015	TFA	2013	4573	Mathematics	Elem & Sec		0.07	0.03	4.21
NCEE 2013-4015	Teaching Fellow	2013	4116	Mathematics	Elem & Sec		0	0.03	0.15
NCEE 2014-4003	Transfer Incentives (Focal Teacher)	2014	17052	Language: Reading	Elementary		0.10	0.05	3.10
				Mathematics	Elementary		0.18	0.05	215.97
				Language: Reading	Secondary		0.01	0.05	0.28
				Mathematics	Secondary		0.04	0.09	0.59
NCEE 2016-4004	Teacher Incentive Fund	2016	121592	Language: Reading	Elem & Sec	Cohort 1	0.03	0.02	0.56
				Mathematics	Elem & Sec	Cohort 1	0.02	0.02	0.27
				Language: Reading	Elem & Sec	Cohort 2	0.03	0.01	8.88
				Mathematics	Elem & Sec	Cohort 2	0.04	0.02	1.41
				Language: Reading	Elem & Sec	Cohort 3	0.04	0.02	1.41
				Mathematics	Elem & Sec	Cohort 3	0.05	0.02	4.36
NCEE 2016-4010	Focusing on Mathematical Knowledge	2016	3677	Mathematics	Elementary	NWEA Test	-0.05	0.04	0.09
				Mathematics	Elementary	State Assessment	-0.06	0.03	0.05
NCEE 2017-4022	DC Opportunity Scholarship Program	2017	1771	Language: Reading	Elem & Sec		-0.09	0.058	0.11
				Mathematics	Elem & Sec		-0.12	0.055	0.09
NCEE 2018-4001	Performance Feedback	2018	30093	Language: Reading	Elem & Sec	Year 1	0.01	0.018	0.15
				Mathematics	Elem & Sec	Year 1	0.053	0.019	8.92
				Language: Reading	Elem & Sec	Year 2	0.024	0.026	0.32
				Mathematics	Elem & Sec	Year 2	0.058	0.03	1.79
NCEE 2018-4004	Teacher Incentive	2018	57897	Combination	Elementary	Year 1	0.36	0.18	3.76
				Combination	Elementary	Year 2	0.27	0.15	3.13
				Combination	Elementary	Year 3	0.04	0.13	0.68

				Combination	Elementary	Year 4	0.39	0.18	4.63

Notes: Bayes Factors were calculated by modelling the alternative hypothesis with a half normal distribution with mean 0 and SD 0.2.

Table S.3
Description of the trials commissioned by the EEF and NCEE, using the first outcome reported.

	EEF	NCEE	Overall
N. trials	82	59	141
Total N. participants	790,279	431,745	1,222,024
Median N. per trial	2,222	2,594	2,386
Effect size			
Min	-0.14	-0.15	-0.15
Max	0.74	0.40	0.74
Median	0.03	0.02	0.02
% positive	68%	66%	67%
Unweighted Mean	0.07	0.05	0.06
95% CI	(0.04; 0.10)	(0.02; 0.08)	(0.04; 0.08)
Weighted Mean	0.04	0.04	0.04
95% CI	(0.02; 0.06)	(0.02; 0.06)	(0.03; 0.05)
Q	153.36	166.58	322.79
I^2	59%	71%	69%
Precision			
Mean CI Width	0.34	0.23	0.29
Median CI Width	0.27	0.20	0.23
% effect size sig. > 0	21%	27%	23%
Mean MDES	0.24	0.16	0.21
Median MDES	0.19	0.14	0.17
Average Power	22%	26%	24%
Median Power	14%	22%	17%
Informativeness			
Bayes Factor			
% Uninformative	45%	25%	37%
% Supporting H0	35%	47%	40%
% Supporting Ha	20%	27%	23%
Median	0.54	0.42	0.51

Notes: Power was calculated assuming an effect size of 0.06. Bayes Factors were calculated by modelling the alternative hypothesis with a half normal distribution with mean 0 and SD 0.2.

Table S.4
Description of the trials commissioned by the EEF and NCEE, using the outcome associated with the largest effect size.

	EEF	NCEE	Overall
N. trials	82	59	141
Total N. participants	790,279	431,745	1,222,024
Median N. per trial	2,222	2,594	2,386
Effect size			
Min	-0.14	-0.15	-0.15
Max	0.74	0.40	0.74
Median	0.03	0.05	0.05
% positive	71%	75%	72%
Unweighted Mean	0.09	0.07	0.08
95% CI	(0.05; 0.12)	(0.04; 0.10)	(0.06; 0.10)
Weighted Mean	0.06	0.06	0.06
95% CI	(0.04; 0.08)	(0.04; 0.08)	(0.04; 0.07)
Q	201.35	147.54	362.22
I^2	77%	66%	73%
Precision			
Mean CI Width	0.34	0.24	0.30
Median CI Width	0.27	0.21	0.24
% effect size sig. > 0	23%	32%	27%
Mean MDES	0.24	0.17	0.21
Median MDES	0.19	0.15	0.17
Average Power	22%	24%	23%
Median Power	14%	20%	16%
Informativeness			
Bayes Factor			
% Uninformative	44%	36%	40%
% Supporting H0	30%	32%	31%
% Supporting Ha	26%	32%	28%
Median	0.67	0.78	0.74

Notes: Power was calculated assuming an effect size of 0.06. Bayes Factors were calculated by modelling the alternative hypothesis with a half normal distribution with mean 0 and SD 0.2.

Table S.5
Description of the trials commissioned by the EEF and NCEE, using every outcome from every trial as if they were independent

	EEF	NCEE	Overall
N. trials	82	59	141
Total N. participants	790,279	431,745	1,222,024
Median N. per trial	2,222	2,594	2,386
Effect size			
Min	-0.16	-0.15	-0.16
Max	0.74	0.40	0.74
Median	0.03	0.04	0.03
% positive	64%	69%	66%
Unweighted Mean	0.06	0.05	0.05
95% CI	(0.04; 0.08)	(0.03; 0.07)	(0.04; 0.07)
Weighted Mean	0.04	0.04	0.04
95% CI	(0.02; 0.05)	(0.02; 0.05)	(0.03; 0.05)
Q	301.85	349.15	656.74
I^2	68%	70%	69%
Precision			
Mean CI Width	0.32	0.23	0.28
Median CI Width	0.26	0.20	0.23
% effect size sig. > 0	20%	25%	23%
Mean MDES	0.23	0.16	0.20
Median MDES	0.19	0.15	0.16
Average Power	21%	27%	24%
Median Power	15%	21%	17%
Informativeness			
Bayes Factor			
% Uninformative	45%	36%	41%
% Supporting H0	35%	39%	37%
% Supporting Ha	20%	25%	23%
Median	0.50	0.62	0.52

Notes: Power was calculated assuming an effect size of 0.06. Bayes Factors were calculated by modelling the alternative hypothesis with a half normal distribution with mean 0 and SD 0.2.

Table S.6.1
Bayes factors associated with various different alternative hypotheses (Model of Ha: half normal distribution)

				% Supporting H0				% Supporting Ha			
Model of Ha	k	Median	% Uninformative (1/3 < BF < 3)	% Moderate (1/10 < BF < 1/3)	% Strong (1/30 < BF < 1/10)	% Very Strong (1/100 < BF < 1/30)	% Decisive (BF < 1/100)	% Moderate (3 < BF < 10)	% Strong (10 < BF < 30)	% Very Strong (30 < BF < 100)	% Decisive (BF > 100)
Half Normal (Mean 0; SD 0.10)	141	0.80	58%	17%	1%	0%	0%	15%	4%	1%	4%
Half Normal (Mean 0; SD 0.15)	141	0.66	47%	27%	2%	0%	0%	15%	3%	3%	4%
Half Normal (Mean 0; SD 0.20)	141	0.56	40%	30%	7%	0%	0%	13%	4%	1%	4%
Half Normal (Mean 0; SD 0.25)	141	0.48	36%	32%	10%	0%	0%	13%	4%	1%	5%
Half Normal (Mean 0; SD 0.30)	141	0.42	35%	31%	13%	1%	0%	11%	4%	1%	5%

Table S.6.2
Bayes factors associated with various different alternative hypotheses (Model of Ha: normal distribution)

				% Supporting H0				% Supporting Ha			
Model of Ha	k	Median	% Uninformative (1/3 < BF < 3)	% Moderate (1/10 < BF < 1/3)	% Strong (1/30 < BF < 1/10)	% Very Strong (1/100 < BF < 1/30)	% Decisive (BF < 1/100)	% Moderate (3 < BF < 10)	% Strong (10 < BF < 30)	% Very Strong (30 < BF < 100)	% Decisive (BF > 100)
Normal (Mean 0; SD 0.10)	141	0.85	79%	6%	0%	0%	0%	9%	2%	1%	3%
Normal (Mean 0; SD 0.15)	141	0.72	69%	16%	0%	0%	0%	9%	1%	2%	4%
Normal (Mean 0; SD 0.20)	141	0.57	57%	23%	4%	0%	0%	10%	1%	2%	4%
Normal (Mean 0; SD 0.25)	141	0.48	49%	30%	5%	0%	0%	9%	1%	3%	3%
Normal (Mean 0; SD 0.30)	141	0.41	43%	38%	5%	0%	0%	8%	1%	3%	3%

Figure S.1
Effect sizes and confidence intervals of EEF trials.

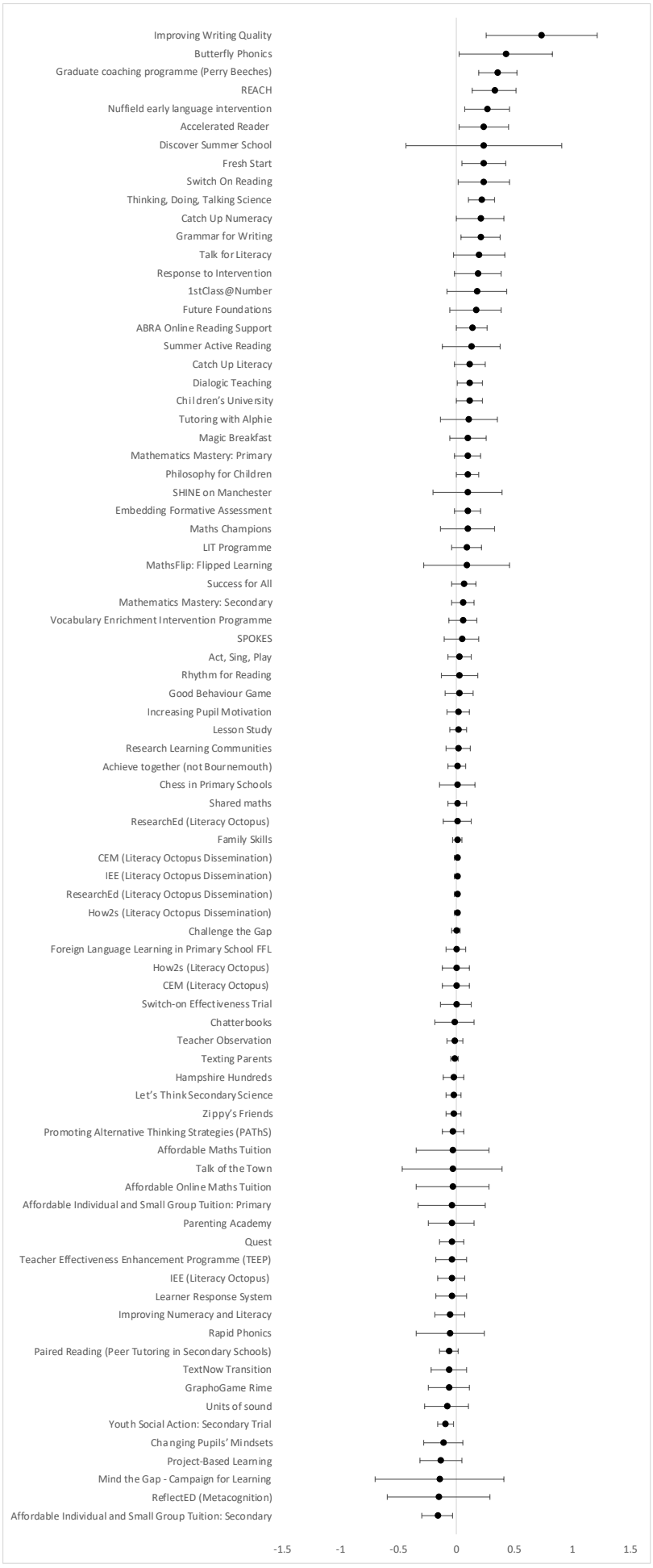


Figure S.2
Effect sizes and confidence intervals of NCEE trials.

References EEF trials

Buchanan, E., Morrison, J., Walker, M., Aston, H., & Cook, R. (2015). *Tutor Trust Secondary Evaluation report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Publications/Campaigns/Evaluation_Reports/EEF_Project_Report_AffordableIndividualAndSmallGroupTuition_Secondary.pdf

Buchanan, E., Worth, J., & Aston, H. (2015). *Tutor Trust Primary Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Affordable_Individual_and_Small_Group_Tuition_Primary.pdf

Centre for Effective Education, & Institute for Effective Education. (2015). *Quest Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Quest.pdf

Centre for Effective Education Queen's University Belfast. (2015). *Tutoring with Alphie Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Tutoring_with_Alphie.pdf

Crawford, C., & Skipp, A. (2014). *LIT Programme Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_LITProgramme.pdf

Crawford, C., Edwards, A., Farquharson, C., Greaves, E., Trevelyan, G., Wallace, E., & White, C. (2016). *Magic Breakfast Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Magic_Breakfast.pdf

Dorsett, R., Rienzo, C., Rolfe, H., Burns, H., Robertson, B-A., Thorpe, B., & Wall, K. (2014). *Mind the Gap Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_MindTheGap.pdf

Gorard, S., See, B. H., & Siddiqui, N. (2014). *Switch-on Reading Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Publications/Campaigns/Evaluation_Reports/EEF_Project_Report_SwitchOnReading.pdf

Gorard, S., See, B. H., Siddiqui, N., Smith, E., & White, P. (2016). *Youth Social Action Trials: Youth United Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Publications/Campaigns/Evaluation_Reports/EEF_Project_Report_Youth_Social_Action_Trials.pdf

Gorard, S., Siddiqui, N., & See, B. H. (2015). *Accelerated Reader Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_AcceleratedReader.pdf

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

https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_FreshStart.pdf

Gorard, S., Siddiqui, N., & See, B. H. (2015). *Philosophy for Children Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_PhilosophyForChildren.pdf

Gorard, S., Siddiqui, N., See, B. H. (2014). *Response to Intervention Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_ResponseToIntervention.pdf

Gorard, S., Siddiqui, N., See, B. H., Smith, E., & White, P. (2017). *Children's University Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Childrens_University.pdf

Greaves, E., Sianesi, B., Sibieta, L., Amin-Smith, N., Callanan, M., & Hudson, R. (2017). *Achieve Together Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Achieve_Together_Evaluation_Report.pdf

Hanley, P., Böhnke, J. R., Slavin, B., Elliott, L., & Croudace, T. (2016). *Let's Think Secondary Science Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Lets_Think_Secondary_Science.pdf

Hanley, P., Slavin, R., & Elliott, L. (2015). *Thinking, Doing, Talking Science Evaluation report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Oxford_Science.pdf

Haywood, S., Griggs, J., Lloyd, C., Morris, S., Kiss, Z., & Skipp, A. (2015). *Creative Futures: Act, Sing, Play Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_ActSingPlay.pdf

Humphrey, N., Hennessey, A., Ashworth, E., Frearson, K., Black, L., Petersen, K., Wo, L., Panayiotou, M., Lendrum, A., Wigelsworth, M., Birchinall, L., Squires, G., & Pampaka, M. (2018). *Good Behaviour Game Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/the-good-behaviour-game/

Husain, F., Jabin, N., Haywood, S., Kasim, A., & Paylor, J. (2016). *Parent Academy Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Parent_Academy.pdf

Husain, F., Wishart, R., Marshall, L., Frankenberg, S., Bussard, L., Chidley, S., Hudson, R., Votjkova, M., & Morris, S. (2018). *Family Skills Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Family_Skills.pdf

Institute for Effective Education. (2016). *Teacher Effectiveness Enhancement Programme Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/TEEP.pdf

Jay, T., Willis, B., Thomas, P., Taylor, R., Moore, N., Burnett, C., Merchant, G., & Stevens, A. (2017). *Dialogic Teaching Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Dialogic_Teaching_Evaluation_Report.pdf

Jerrim, J., Austerberry, H., Crisan, C., Ingold, A., Morgan, C., Pratt, D., Smith, C., & Wiggins, M. (2015). *Mathematics Mastery Secondary Evaluation Report.* https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_Mathematics_Mastery_Secondary

Jerrim, J., Macmillan, L., Micklewright, J., Sawtell, M., & Wiggins, M. (2016). *Chess in Schools Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_Chess_in_Schools.pdf

King, B., & Kasim, A. (2015). *Rapid Phonics Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_RapidPhonics.pdf

Lloyd, C., Edovald, T., Kiss, Z., Morris, S., Skipp, A., & Ahmed, H. (2015). *Paired Reading Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Peer_Tutoring_in_Secondary_Schools.pdf

Lloyd, C., Edovald, T., Morris, S., Kiss, Z., Skipp, A., & Haywood, S. (2015). *Durham Shared Maths Project Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Shared_Maths.pdf

Lord, P., Bradshaw, S., Stevens, E., & Styles, B. (2015). *Perry Beeches Coaching Programme Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_GraduatesCoachingProgramme.pdf

Lord, P., Rabiasz, A., Roy, P., Harland, J., Styles, B., & Fowler, K. (2017). *Evidence-based Literacy Support: the 'Literacy Octopus' Trial Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Evidence-based_Literacy_Support_-_the_Literacy_Octopus_Trial.pdf

Lord, P., Rabiasz, A., & Styles, B. (2017). *'Literacy Octopus' Dissemination Trial.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Literacy_Octopus_Dissemination_Trial.pdf

Manchester Institute of Education, & PATHS to Success. (2015). *Promoting Alternative Thinking Strategies (PATHS) Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from

https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/PAThS.pdf

Maxwell, B., Connolly, P., Demack, S., O'Hare, L., Stevens, A., & Clague, L. (2014). *Summer Active Reading Programme Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_SummerActiveReading.pdf

Maxwell, B., Connolly, P., Demack, S., O'Hare, L., Stevens, A., & Clague, L. (2014). *TextNow Transition Programme Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_TextNow.pdf

McNally, S., Challen, A., Wyness, G., West, A., & Noden, P. (2014). *Hampshire Hundreds Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Hampshire_Hundreds.pdf

McNally, S., Ruiz-Valenzuela, J., & Rolfe, H. (2016). *ABRA: Online Reading Support Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_ABRA.pdf

Menzies, V., Hewitt, C., Kokotsaki, D., Collyer, C., & Wiggins, A. (2016). *Project Based Learning Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_Project_Based_Learning.pdf

Menzies, V., Kasim, A., Kokotsaki, D., Hewitt, C., Akhter, N., Collyer, C., Younger, K., Wiggins, A., & Torgerson, C. (2016). *Hallé SHINE on Manchester Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_SHINE.pdf

Merrell, C., & Kasim, A. (2015). *Butterfly Phonics.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_ButterflyPhonics.pdf

Miller, S., Biggart, A., Sloan, S., & O'Hare, L. (2017). *Success for All Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Success_for_All_Evaluation_Report.pdf

Miller, S., Davison, J., Yohanis, J., Sloan, S., Gildea, A., & Thurston, A. (2016). *Texting Parents Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Texting_Parents.pdf

Motteram, G., Choudry, S., Kalambouka, A., Hutcheson, G., & Barton, A. (2016). *ReflectED Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_ReflectED.pdf

Murphy, R., Weinhardt, F., Wyness, G., & Rolfe, H. (2017). *Lesson Study Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Lesson_Study.pdf

NFER. (2014). *Catch Up® Numeracy Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_CatchUpNumeracy.pdf

Nunes, T., Barros, R., Evangelou, M., Strand, S., Mathers, S., & Sanders-Ellis, D. (2018). *1stClass@Number Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/1stClass@Number_evaluation_report.pdf

Patel, R., Jabin, N., Bussard, L., Cartagena, J., Haywood, S., & Lumpkin, M. (2017). *Switch-on Effectiveness Trial Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_Switchon_Effectiveness.pdf

Rienzo, C., Rolfe, H., & Wilkinson, D. (2015). *Changing Mindsets Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Changing_Mindsets.pdf

Robinson-Smith, L., Fairhurst, C., Stone, G., Bell, K., Elliott, L., Gascoine, L., Hallett, S., Hewitt, C., Hugill, J., Torgerson, C., Torgerson, D., Menzies, V., & Ainsworth, H. (2018). *Maths Champions Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Maths_champions_evaluation_report.pdf

Rose, J., Thomas, S., Zhang, L., Edwards, A., Augero, A., & Roney, P. (2017). *Research Learning Communities Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Research_Learning_Communities.pdf

Rudd, R., Aguilera, A. B. V., Elliott, L., & Chambers, B. (2017). *MathsFlip: Flipped Learning Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Flipped_Learning.pdf

NFER (2015). *Catch Up® Literacy Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_CatchUpLiteracy.pdf

Sheard, M., Chambers, B., & Elliott, L. (2015). *Units of Sound Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Publications/Campaigns/Evaluation_Reports/EEF_Project_Report_UnitsOfSound.pdf

Sibieta, L. (2016). *REACH Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from

https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_REACH

Sibieta, L., Greaves, E., & Sianesi, B. (2014). *Increasing Pupil Motivation Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Pupil_Incentives.pdf

Sibieta, L., Kotecha, M., & Skipp, A. (2016). *Nuffield Early Language Intervention Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_Nuffield_Early_Language_Intervention.pdf

Siddiqui, N., & See. B. H. (2014). *Future Foundations Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_FutureFoundationsSummerSchool.pdf

Sloan, S., Gildea, A., Miller, S., & Thurston, A. (2018). *Zippy's Friends Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Zippys_Friends.pdf

Speckesser, S., Runge, J., Foliano, F., Bursnall, M., Hudson-Sharp, N., Rolfe, H., & Anders, J. (2018). *Embedding Formative Assessment Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/EFA_evaluation_report.pdf

Styles, B., & Bradshaw, S. (2015). *Talk for Literacy Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Publications/Campaigns/Evaluation_Reports/EEF_Project_Report_TalkForLiteracy.pdf

Styles, B., Clarkson, R., & Fowler, K. (2014). *Chatterbooks Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_Chatterbooks.pdf

Styles, B., Clarkson, R., & Fowler, K. (2014). *Rhythm for Reading Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Rhythm_for_Reading.pdf

Styles, B., Stevens, E., Bradshaw, S., Clarkson, R. (2014). *Vocabulary Enrichment Intervention Programme Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_VocabularyEnrichment.pdf

Thurston, A. (2016). *Talk of the Town Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Talk_of_the_Town.pdf

Torgerson, C., Ainsworth, H., Buckley, H., Hampden-Thompson, G., Hewitt, C., Humphry, D., Jefferson, L., Mitchell, N., & Torgerson, D. (2016). *Affordable Online Maths Tuition.* London: Education Endowment Foundation. Retrieved from

https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Affordable_Maths.pdf

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. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Evaluation_Report_-_Improving_Writing_Quality.pdf

Torgerson, D., Torgerson, C., Jefferson, L., Buckley, H., Ainsworth, H., Heaps, C., & Mitchell, N. (2014). *Discover Summer School Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_DiscoverSummerSchool.pdf

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. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_GrammarForWriting.pdf

Tracey, L., Chambers, B., Bywater, T., & Elliott, L. (2016). *SPOKES Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/SPOKES.pdf

Vignoles, A., Jerrim, J., & Cowan, R. (2015). *Mathematics Mastery Primary Evaluation Report*. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_MathematicsMasteryPrimary.pdf

West, M., Ainscow, M., Wigelsworth, M., & Troncoso, P. (2017). *Challenge the Gap Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Challenge_the_Gap_Evaluation_Report.pdf

Wiggins, M., Parrao, C. G., Austerberry, H., & Ingold, A. (2017). *Foreign Language Learning in Primary School Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/EEF_Project_Report_FLL.pdf

Wiggins, M., Sawtell, M., & Jerrim, J. (2017). *Learner Response System Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Learner_Response_System.pdf

Worth, J., Sizmur, J., Ager, R., & Styles, B. (2015). *Improving Numeracy and Literacy Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_ImprovingNumeracyAndLiteracyInKeyStage1.pdf

Worth, J., Sizmur, J., Walker, M., Bradshaw, S., & Styles, B. (2017). *Teacher Observation Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Teacher_Observation.pdf

Worth, J., Nelson, J., Harland, J., Bernardinelli, D., & Styles, B. (2018). *GraphoGame Rime Evaluation Report and Executive Summary.* London: Education Endowment Foundation. Retrieved from

https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/GraphoGame_Rime.pdf

References NCEE trials

Abe, Y., Thomas, V., Sinicrope, C., & Gee, K. A. (2012). *Effects of the Pacific CHILD Professional Development Program: Final Report. (NCEE 2013–4002).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/pacific/pdf/REL_20134002.pdf

Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). *Achievement Effects of Four Early Elementary School Math Curricula: Findings for First and Second Graders. (NCEE 2011-4001).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20114001/pdf/20114001.pdf

Arens, S. A., Stoker, G., Barker, J., Shebby, S., Wang, X., Cicchinelli, L. F., & Williams, J. M. (2012). *Effects of Curriculum and Teacher Professional Development on the Language Proficiency of Elementary English Language Learner Students in the Central Region. (NCEE 2012-4013).* Denver, CO: Mid-continent Research for Education and Learning. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/central/pdf/REL_20124013.pdf

Bernstein, L., Rappaport, C. D., Olsho, L., Hunt, D., & Levin, M. (2009). *Impact Evaluation of the U.S. Department of Education's Student Mentoring Program Final Report. (NCEE 2009-4047)*. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20094047/pdf/20094047.pdf

Black, A. R., Somers, M-A., Doolittle, F., Unterman, R., & Grossman, J. B. (2009). *The Evaluation of Enhanced Academic Instruction in After-School Programs Final Report. (NCEE 2009-4077)*. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20094077/pdf/20094077.pdf

Bos, J. M., Sanchez, R. C., Tseng, F., Rayyes, N., Ortiz, L., and Sinicrope, C. (2012). *Evaluation of Quality Teaching for English Learners (QTEL) Professional Development: Final Report. (NCEE 2012-4005).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/west/pdf/REL_20124005.pdf

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). *Effectiveness of Reading and Mathematics Software Products - Findings From Two Student Cohorts. (NCEE 2009-4041).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20094041/pdf/20094041.pdf

Cavalluzzo, L., Lowther, D. L., Mokher, C., & Fan, X. (2012). *Effects of the Kentucky Virtual Schools' Hybrid Program for Algebra I on grade 9 Student Math Achievement: Final Report. (NCEE 2012-4020).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/appalachia/pdf/20124020.pdf

Chiang, H., Speroni, C., Herrmann, M., Hallgren, K., Burkander, P., & Wellington, A. (2017). *Evaluation of the Teacher Incentive Fund: Final Report on Implementation and Impacts of Pay-for-Performance Across Four Years. (NCEE 2017-4004).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20184004/pdf/20184004.pdf

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). *The Effectiveness of Secondary Math Teachers from Teach for America and the Teaching Fellows Programs. (NCEE 2013-4015).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20134015/pdf/20134015.pdf

Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (2011). *An Investigation of the Impact of the 6+1 Trait Writing Model on Grade 5 Student Writing Achievement: Final Report. (NCEE 2012–4010)*. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://files.eric.ed.gov/fulltext/ED527445.pdf

Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). *An Evaluation of Teachers Trained Through Different Routes to Certification, Final Report. (NCEE 2009-4043)*. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20094043/pdf/20094044.pdf

Cordray, D., Pion, G., Brandt, C., Molefe, A., & Toby, M. (2012). *The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement: Final Report. (NCEE 2013–4000)*. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://files.eric.ed.gov/fulltext/ED537982.pdf

Corrin, W., Lindsay, J. J., Somers, M-A., Myers, N. E., Meyers, C. V., Condon, C. A., & Smith, J. K. (2012). *Evaluation of the Content Literacy Continuum: Report on Program Impacts, Program Fidelity, and Contrast: Final Report. (NCEE 2013-4001).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/REL_20134001.pdf

Drumond, K., Chinen, M., Duncan, T. G., Miller, H. R., Fryer, L., Zmach, C., & Culp, K. (2011). *Impact of the Thinking Reader® Software Program on Grade 6 Reading Vocabulary, Comprehension, Strategies, and Motivation. (NCEE 2010–4035).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://files.eric.ed.gov/fulltext/ED517968.pdf

Dynarski, M., Rui, N., Webber, A., & Gutmann, B. (2017). *Evaluation of the DC Opportunity Scholarship Program: Impacts After One Year. (NCEE 2017-4022).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20174022/pdf/20174022.pdf

Finkelstein, N., Hanson, T., Huang, C-W., Hirschman, B., & Huang, M. (2011). *Effects of Problem Based Economics on High School Economics Instruction: Final Report. (NCEE 2010-4002rev).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/west/pdf/REL_20104022.pdf

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). *Reading First Impact Study: Final Report. (NCEE 2009-4038).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pdf/20094038.pdf

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., ... & Sztejnberg, L. (2008). *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement. (NCEE 2008-4030)*. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pdf/20084030.pdf

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., … & Borman, G. D. (2016). *Focusing on Mathematical Knowledge: The impact of Content-Intensive Teacher Professional Development. (NCEE 2016-4010).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20164010/pdf/20164010.pdf

Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M., & Manzeske, D. (2017). *The Impact of Providing Performance Feedback to Teachers and Principals. (NCEE 2018-4001).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20184001/pdf/20184001.pdf

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., … & Doolittle, F. (2011). *Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation. (NCEE 2011-4024)*. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://files.eric.ed.gov/fulltext/ED519922.pdf

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010). *Impacts of Comprehensive Teacher Induction: Final Results From a Randomized Controlled Study. (NCEE 2010-4027).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://files.eric.ed.gov/fulltext/ED565837.pdf

Glazerman, S., Protik, A., Teh, B-r., Bruch, J., & Max, J. (2013). *Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment. (NCEE 2014-4003).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20144003/pdf/20144003.pdf

Gleason, P., Clark, M., Tuttle, C. C., & Dwoyer, E. (2010). T*he Evaluation of Charter School Impacts: Final Report. (NCEE 2010-4029).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20104029/pdf/20104030.pdf

Goodson, B., Wolf, A., Bell, S., Turner, H., & Finney, P. B. (2010). *The Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB): Kindergarten Final Evaluation Report. (NCEE 2010- 4014).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://files.eric.ed.gov/fulltext/ED512900.pdf

Hanson, T., Dietsch, B., & Zheng, H. (2012). *Lessons in Character Impact Evaluation: Final Report. (NCEE 2012-4004).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/west/pdf/20124004.pdf

Heller, J. I. (2012). *Effects of Making Sense of SCIENCE™ Professional Development on the Achievement of Middle School Students, Including English Language Learners: Final Report. (NCEE 2012-4002).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/west/pdf/REL_20124002.pdf

Heppen, J.B., Walters, K., Clements, M., Faria, A-M., Tobey, C., Sorensen, N., & Culp, K. (2011). *Access to Algebra I: The Effects of Online Mathematics for Grade 8 Students: Final Report. (NCEE 2012–4021).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_20124021.pdf

Hitchcock, J., Dimino, J., Kurki, A., Wilkins, C., & Gersten, R. (2011). *The Impact of Collaborative Strategic Reading on the Reading Comprehension of Grade 5 Students in Linguistically Diverse Schools. (NCEE 2011–4001).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/REL_20114001.pdf

James-Burdumy, S., Deke, J., Lugo-Gil, J., Carey, N., Hershey, A., Gersten, R., … & Faddis, B. (2010). *Effectiveness of Selected Supplemental Reading Comprehension Interventions: Findings From Two Student Cohorts. (NCEE 2010-4015)*. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20104015/pdf/20104015.pdf

Judkins, D., St.Pierre, R., Gutmann, B., Goodson, B., Glatz, A. v., Hamilton,, J., … & Rimdzius, T. (2008). *A Study of Classroom Literacy Interventions and Outcomes in Even Start. (NCCE 2008-4028)*. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20084028/pdf/20084028.pdf

Kushman, J., Hanita, M., & Raphael, J. (2011). *An Experimental Study of the Project CRISS Reading Program on Grade 9 Reading Achievement in Rural High Schools. (NCEE 2010-4007)*. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/northwest/pdf/REL_20114007.pdf

Martin, T., Brasiel, S. J., Turner, H., & Wise, J. C. (2012). *Effects of the Connected Mathematics Project 2 (CMP2) on the Mathematics Achievement of Grade 6 Students in the Mid-Atlantic Region: Final Report. (NCEE 2012-4017).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_20124017.pdf

Newman, D., Finney, P. B., Bell, S., Turner, H., Jaciw, A. P., Zacamy, J. L., & Gould, L. F. (2012). *Evaluation of the Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI): Final Report. (NCEE 2012–4008).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/southeast/pdf/REL_20124008.pdf

Randel, B., Beesley, A. D., Apthorp, H., Clark, T. F., Wang, X., Cicchinelli, L. F., & Williams, J. M. (2011). *Classroom Assessment for Student Learning: Impact on Elementary School Mathematics in the Central Region. (NCEE 2011–4005).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/central/pdf/REL_20114005.pdf

Rolfhus, E., Gersten, R., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2012). *An evaluation of Number Rockets: a Tier-2 intervention for grade 1 students at risk for difficulties in mathematics: Final Report. (NCEE 2012-4007).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/REL_20124007.pdf

Somers, M-A., Corrin, W., Sepanik, S., Salinger, T., Levin, J., & Zmach, C. (2010). *The Enhanced Reading Opportunities Study Final Report: The Impact of Supplemental Literacy Courses for Struggling Ninth-Grade Readers. (NCEE 2010-4021)*. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20104021/pdf/20104021.pdf

Torgesen, J., Schirm, A., Castner, L., Vartivarian, S., Mansfield, W., Myers, D., Stancavage, F., Durno, D., Javorsky, R., & Haan, C. (2007). *National Assessment of Title I, Final Report: Volume II: Closing the Reading Gap, Findings from a Randomized Trial of Four Reading Interventions for Striving Readers. (NCEE 2008-4013).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pdf/20084013.pdf

Wellington, A., Chiang, H., Hallgren, K., Speroni, C., Herrmann, M., & Burkander, P. (2016). *Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance After Three Years. (NCEE 2016-4004).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20164004/pdf/20164004.pdf

Wijekumar, K., Hitchcock, J., Turner, H., Lei, P., & Peck, K. (2009). *A Multisite Cluster Randomized Trial of the Effects of CompassLearning Odyssey® Math on the Math Achievement of Selected Grade 4 Students in the Mid-Atlantic Region. (NCEE 2009–4068).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_20094068.pdf

Wilkerson, S. B., Shannon, L. C., Styers, M. K., & Grant, B-J. (2012). *A Study of the Effectiveness of a School Improvement Intervention (Success in Sight): Final Report. (NCEE 2012-4014).*

Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/central/pdf/REL_20124014.pdf

Wilkins, C., Gersten, R., Decker, L. E., Grunden, L., Brasiel, S., Brunnert, K., & Jayanthi, M. (2012). *Does a Summer Reading Program Based on Lexiles Affect Reading Comprehension? Final Report. (NCEE 2012-4006).* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/REL_20124006.pdf