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

Hugues Lortie-Forgues[1] and Matthew Inglis[2]

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 UK-based Education Endowment Foundation and the U.S.-based National Center for Educational Evaluation and Regional Assistance, 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. These sat within relatively large confidence intervals (mean width = 0.30 *SD*s), 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.

Keywords: educational policy; evaluation; meta-analysis; program evaluation

Large-scale randomized controlled trials (RCTs) are now regularly used to evaluate educational interventions. For example, the U.S.-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 article, 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.

Educational Researcher, Vol. XX No. X, pp. 1–9
DOI: 10.3102/0013189X19832850

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 that 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, the allocation of the participants is compromised, the outcomes are selectively reported, attrition is ignored, or the outcome measure provides an unfair advantage to the intervention group (e.g., by including material that is taught to the intervention group but not the control group).

[1]University of York, UK
[2]Loughborough University, UK

In this article, 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 analyze 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 (e.g., the National Center for Education Research [NCER], another U.S.-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 that 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 on 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 (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 *SD*s. 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 nonrandomized studies: Cheung and Slavin (2016) found that the effect sizes from randomized educational experiments was 0.16 compared to 0.23 for nonrandomized quasi-experimental studies. Likewise, studies using independent outcome measures, such as standardized tests, tend to produce smaller effects than studies using researcher-made measures. For instance, when comparing the performance of fifth and sixth graders on a standardized test of reading, the impact of an additional year of instruction and maturation is only around 0.23 *SD*s (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 that 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 *SD*s from fifth to sixth 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, namely, that the trial is uninformative (i.e., an RCT that 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 favor 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 one-third 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. Jeffreys 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 1, 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 on the journal website.

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: (a) using the first outcome reported, (b) using the outcome associated with the largest effect size, and (c) using every outcome from every trial as if they were independent (as shown in the Supplemental Material available on the journal website, 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 eight 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, Hedges, Higgins, & Rothstein, p. 27), a procedure that ignores clusters and thus can overstate the accuracy of the estimated effect. Excluding these eight trials from our analysis does not materially affect our conclusions.

We also coded the topic of the outcome measures (e.g., reading, mathematics), age of participants, sample size, and report's year of publication. For the EEF trials, we also coded the type of trial (efficacy or effectiveness), total cost of the trial, cost of the

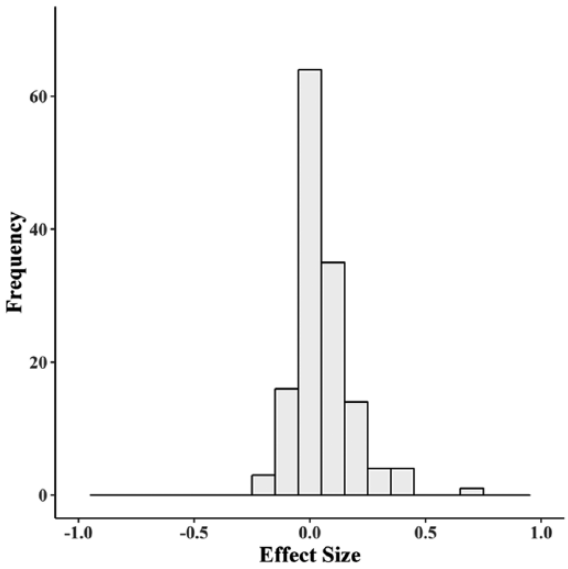

FIGURE 1. *The distribution of effect sizes from the 141 trials commissioned by the Education Endowment Foundation and National Center for Educational Evaluation and Regional Assistance.*

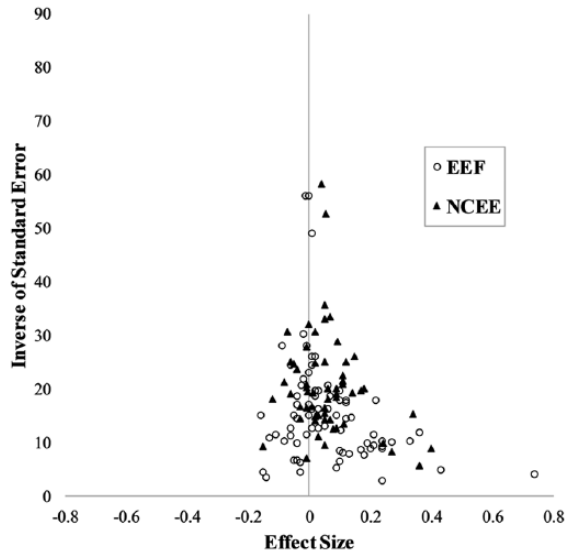

FIGURE 2. *A funnel plot of effect sizes from the trials commissioned by the Education Endowment Foundation (82 trials) and National Center for Educational Evaluation and Regional Assistance (59 trials).*

intervention per pupil (a number from 1 = *low cost* to 5 = *high cost*), and 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, etc.) 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 on the journal website.

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 both funders combined.

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 2,386. 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 on the journal website).

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 analyzed 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.

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.

Table 1
Description of the Trials Commissioned by the Education Endowment Foundation (EEF) and National Center for Educational Evaluation and Regional Assistance (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			
Minimum	–0.16	–0.15	–0.16
Maximum	0.74	0.40	0.74
Median	0.01	0.05	0.03
Percentage 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
Percentage effect size significant > 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			
Percentage uninformative	40	39	40
Percentage supporting H0	40	34	38
Percentage supporting Ha	20	27	23
Median	0.50	0.67	0.56

Note. Power was calculated assuming an effect size of 0.06. Bayes factors were calculated by modeling the alternative hypothesis (Ha) with a half normal distribution with mean 0 and *SD* 0.2. CI = confidence interval; H0 = null hypothesis; MDES = minimal detectable effect size.

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 *SD*s. As shown in Figure 3, for more than 93% of the trials, the MDES was greater than the effect size observed.

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 nine 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 (2018). 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 on the journal website). We interpreted the resulting Bayes factors, summarized in Table 1, following Jeffrey's (1961) guidelines. Many, 40%, fell between 3 and one-third, indicating that the trial was uninformative; 38% were less than one-third, 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%

Table 2
Analysis of the Subgroups Identified in the Trials Commissioned by the Education Endowment Foundation

Subgroup	*k*	Mean	95% CI	*Q*	*df*(*Q*)	*p* Value
Topic						
Language: Reading	63	0.04	[0.01, 0.04]	8.89	4	.064
Mathematics	35	0.04	[0.02, 0.07]			
Language: General	20	0.03	[–0.01, 0.07]			
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]	4.45	3	.216
Elementary	86	0.04	[0.02, 0.05]			
Secondary	36	0.03	[–0.01, 0.06]			
Elementary and secondary	13	0.09	[0.00, 0.18]			
Type of trial						
Efficacy trial	117	0.05	[0.03, 0.07]	4.23	1	.040
Effectiveness trial	23	0.01	[0.00, 0.02]			

Subgroup	*k*	Coefficient	*Z*	*p* Value
Year of publication				
2014–2018	140	–0.01	–1.70	.090
Quality trial				
0 (low) to 5 (high)	139	0.00	–0.70	.486
Cost intervention per pupil				
1 (low) to 5 (high)	140	0.00	0.55	.584
Cost trial, £				
70,000 to 1.4 million	140	0.00	–1.76	.078

Note. CI = confidence interval.

Table 3
Analysis of the Subgroups Identified in the Trials Commissioned by the National Center for Educational Evaluation and Regional Assistance

Subgroup	*k*	Mean	95% CI	*Q*	*df*(*Q*)	*p* Value
Topic						
Language: Reading	61	0.04	[0.02, 0.06]	7.66	3	.054
Mathematics	39	0.04	[0.01, 0.06]			
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]	7.74	3	.052
Elementary	73	0.06	[0.04, 0.08]			
Secondary	24	0.03	[0.00, 0.06]			
Elem and secondary	22	0.03	[0.02, 0.04]			

Subgroup	*k*	Coefficient	*Z*	*p* Value
Year of publication				
2008–2018	131	0.00	–0.05	.96

Note. CI = confidence interval.

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 *SD*s, 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 *SD*s). By contrast, the confidence intervals of these effect sizes were comparatively large, on average 0.30 *SD*s wide. Consequently, many trials were uninformative: 40% of trials yielded Bayes factors between 3 and

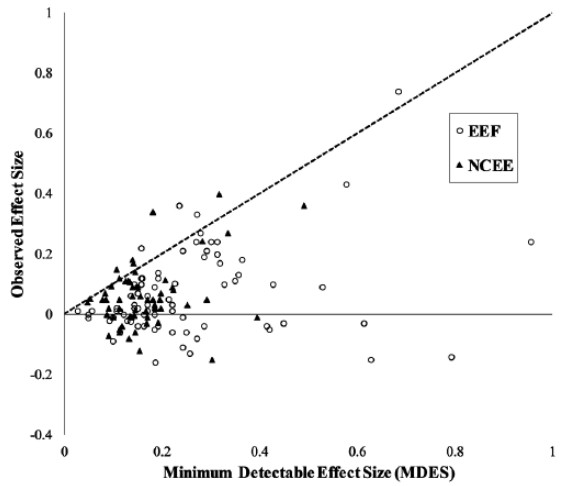

FIGURE 3. *A scatterplot showing the relation between the minimum detectable effect sizes (MDES) and observed effect sizes in the trials commissioned by the Education Endowment Foundation and National Center for Educational Evaluation and Regional Assistance.*
Note. The diagonal line represents obtained effect sizes equal to the MDES of the trial. Points below the diagonal represent obtained effect sizes below the MDES of the trial, and points above the diagonal represent obtained effect sizes above the MDES of the trial.

one-third. These trials produced findings consistent with both the null hypothesis of no effect and also an effect comparable to that associated with one year of maturation and instruction (Bloom et al., 2008). Such trials allow us to conclude neither 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: (a) that many of the interventions studied are ineffective because the literature on which they are based is unreliable, (b) that many of the interventions studied are ineffective because they have been poorly designed or implemented, and (c) 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 on 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 et al., 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 on 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, insufficient 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 on 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 Collaboration, 2015; Simmons et al., 2011; Society for Research on Educational Effectiveness, 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 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 that 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 that 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 on 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.

NOTE

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 also thank David W. Braithwaite, Steve Higgins, Robert M. Klassen, Michael Schneider, and ZhiMin Xiao for their comments; Erin Pollard (Institute of Education Sciences) for assistance with National Center for Educational Evaluation and Regional Assistance trials; and Anh Nguyen Van Pham for assistance with coding.

REFERENCES

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

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

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*, 289–328.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: 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. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 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*, 240–246.

Education Endowment Foundation. (2015a). *Annual report 2014/15*. London: Author.

Education Endowment Foundation. (2015b). *EEF guidance on cost evaluation*. London: Author.

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

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

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

Higgins, J. P., Altman, D. G., Gotzsche, P. C., Juni, P., Moher, D., Oxman, A. D., . . . 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*, d5928.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *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*, 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. (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. (2017). *Evaluation principles and practices*. Retrieved from https://ies.ed.gov/ncee/projects/pdf/IESEvaluationPrinciplesandPractices_011117.pdf

National Center for Education Research. (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, UK: 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. (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.

AUTHORS

HUGUES LORTIE-FORGUES, PhD, is an assistant professor in the Department of Education at the University of York, York, YO10 5DD, United Kingdom; *hugues.lortie-forgues@york.ac.uk*. His research focuses on mathematics education and the evaluation of educational interventions.

MATTHEW INGLIS, PhD, is a professor of mathematical cognition at Loughborough University, Leicestershire, LE11 3TU, United Kingdom; *m.j.inglis@lboro.ac.uk*. His research focuses on understanding the cognitive processes involved in mathematical thinking and reasoning.

Manuscript received April 11, 2018
Revisions received August 28, 2018, and January 10, 2019
Accepted January 22, 2019