The misdirection of public policy: comparing and combining standardised effect sizes

Adrian Simpson
School of Education
Durham University

Abstract

Increased attention on 'what works' in education has led to an emphasis on developing policy from evidence based on comparing and combining a particular statistical summary of intervention studies: the standardised effect size. It is assumed that this statistical summary provides an estimate of the educational impact of interventions and combining these through meta-analyses and meta-meta-analyses results in more precise estimates of this impact which can then be ranked. From these, it is claimed, educational policy decisions can be driven. This paper will demonstrate that these assumptions are false: standardised effect size is open to researcher manipulations which violate the assumptions required for legitimately comparing and combining studies in all but the most restricted circumstances. League tables of types of intervention, which governments point to as an evidence base for effective practice may, instead, be hierarchies of openness to research design manipulations. The paper concludes that public policy and resources are in danger of being misdirected.

Keywords: effect size, experimental design, meta-analysis

To appear in: *Journal of Education Policy*

1 Introduction

Policy makers, schools and teachers make decisions about how to teach and organise education. We expect that they do so to improve outcomes. Different people may have different views of what constitutes better or worse outcomes and how to evaluate the impact of interventions. However, for most it would be uncontroversial to say that decisions should be made on the basis of evidence, even where people contest what counts as evidence.

Within the 'what works' movement there are a number of influential attempts to rank high level quantitative summaries of existing educational research under broad themes (e.g. Marzano 1998; Hattie 2009; Higgins et al. 2013). These work in similar ways: individual studies report quantitative measures of the outcomes of particular interventions; meta-analysts collect studies in a given area, convert outcome measures to a common metric and combine those to report an estimate which they claim represents the impact or influence of interventions in that area. Meta-meta-analysis then takes the results of meta-analyses, collected in broader fields, and combine those estimates to provide a rank ordering of those fields which make the most difference.

For example, Oladunni (1998) studied the teaching of problem solving in mathematics; Chang and Barufaldi (1999) studied a problem-solving based instructional model in earth sciences. From each of these studies, with seventeen others, Higgins et al. (2005) extracted a numerical value which they claim estimates the effectiveness of the study's intervention and combined all nineteen values to obtain a single number as an indication of the impact of 'thinking skills approaches'. This first level of the aggregation of studies is a meta-analysis: 'the best way to pool the results of a range of studies quantitatively … to compare the impact of thinking skills approaches with other researched educational interventions' (Higgins et al. 2005, 15).

Higgins et al. (2005) is combined with five other meta-analyses in the broad theme of 'metacognition' in a meta-meta-analysis (Higgins et al. 2013). The resulting numerical summary for this theme is then compared with the numerical summaries for many other educational areas (such as 'learning styles', 'digital technologies' and 'extending school time') to give a 'toolkit' of different areas which 'can help schools get the maximum "educational bang for their buck" … in terms of making an initial choice between possible strategies' (Higgins et al. 2013, 3).

This paper argues that this view is wrong.

The numerical summaries used to develop the toolkit (or the alternative 'barometer of influences': Hattie 2009) are not a measure of educational impact because larger numbers produced from this process are not indicative of larger educational impact. Instead, areas which rank highly in Marzano (1998), Hattie (2009) and Higgins et al. (2013) are those in which researchers can design more sensitive experiments.

As such, using these ranked meta-meta-analyses to drive educational policy is misguided.

2 'What Works' and evidence based education

There have long been calls for 'evidence-based education' often on the basis of analogies between education and medicine (Davies 1999). Such analogies are not always accepted: treatments in medicine are often standard and well specified, with agreed outcomes which are relatively easy to measure. This is not generally true in education and critics argue that there is a stark difference between the nature of evidence which underpins explanations of physical events and that which underpins explanations of human behaviour (Pring 2004). While many meta-analyses provide theoretical context (e.g. Kluger and DeNisi 1996), when reduced still further in meta-meta-analyses, the theory which might support practitioners connecting 'what worked' (in the study context) with 'what works here' (in a particular teacher's classroom) becomes obscured (Cartwright and Hardie 2012).

Nevertheless, governments promote the numerical summaries from meta-meta-analyses as a key evidence base for policy. In the UK, a recent government white paper urges teachers to use 'evidence which sets out what works and what doesn't' (Department for Education 2016, 37). It cites the meta-meta-analysis from the Education Endowment Foundation (EEF) as an example of such high quality evidence:

> The EEF's Teaching and Learning Toolkit is helping teachers to find and use evidence about the most effective teaching methods to improve standards for all children, including the most disadvantaged. A recent National Audit Office survey found that nearly two thirds of school leaders use the Toolkit showing that high quality evidence is now more accessible than ever before. (38)

The EEF is the UK government's designated 'what works' centre for education and they claim the toolkit provides a 'way of schools assessing the effectiveness of interventions …which identifies high-impact techniques such as improving the quality of feedback to pupils' (House of Commons Education Committee 2014, 43). It is promoted to help schools direct the £2.5bn aimed at closing the perceived attainment gap for disadvantaged children.

There is currently little evidence about how teachers and other policy makers use these meta-meta-analyses, but what there is suggests that the numerical summaries are taken uncritically with no doubts expressed that higher ranked interventions in these analyses are educationally more important (Cowan et al. 2015). It suggests teachers and policy makers are prone to 'metricophilia': the 'expectation that quantitative data — virtually on their own — will give us the answers on which to base policy in education' (Smith 2011, 633).

The argument here is not simply that the quantitative summaries are not a good representation of what works, but that the way in which they are derived (from comparing and combining standardised effect sizes) means that their rank order does not represent better and worse examples of 'what works' at all: rather they represent easier and harder contexts in which researchers can design sensitive studies.

3 Standardised effect size

Meta-analyses and meta-meta-analyses are based on a key statistic:

> One of the great virtues of meta-analysis is that it rests essentially on a measurement, 'Effect Size', which can be understood in the first week of a statistics course or by anyone who can remember what a *z*-score is.' (FitzGibbon 1984, 134).

This paper will argue that, while calculating an effect size may be simple enough for a first course in statistics, there are considerable subtleties in understanding it sufficiently well to ensure that the processes of combining effect sizes in meta-analyses allows valid conclusions to be drawn.

Standardised effect size came from Cohen's (1962) concern about whether psychology studies had sufficient *power*: that is, whether each study had a large enough chance to be able to reject a (false) null hypothesis. Cohen argued that too few researchers were calculating power and adjusting their designs to increase it when the chance of detecting an effect was too low.

In explaining power, Cohen noted the key question: 'How large an effect (a difference, a correlation coefficient, etc.) in the population do I expect actually exists' (Cohen 1962, 146). One of the most important of the measures he proposed is standardised mean difference, which came to be called 'Cohen's *d*': 'the difference in means expressed in units of standard deviations' (146). Variants of this are used as the common metric into which many meta-analysts (and meta-meta-analysts) convert original research results to compare them or combine them.

However, it is important to note that the aim of researchers in many studies is normally to determine whether there is a statistically significant difference between groups which they can attribute to the intervention, and understandably they make methodological decisions to maximise the chance of doing so (provided such a difference exists).

For example, Lumbelli et al. (1999) randomly assigned 28 children to a control group or to an experimental group which was given three sessions which involved examining texts for lapses in clarity or gaps and listening to a tape of an adult reading and commenting on the same passages. After the intervention, the researchers gave both groups a similar text to check and revise for lapses in clarity or gaps. The mean number of gaps/lapses for the experimental group was 3.64 (with a standard deviation of 1.39) and the mean for the control group was 2.36 (standard deviation 1.55). That is, the experimental group found or corrected on average 1.28 gaps/lapses more than the control group. Lumbelli et al. did not directly calculate an effect size: that was not their aim. However, to combine it with other studies, meta-analysts converted these raw mean differences into standardised mean differences (Graham, Hebert & Harris 2015).

To find this, the difference in means is divided by the standard deviation. Of course, there are two standard deviations (one for the control group and one for the experimental group), and different analysts choose different combinations of these to create the standardised mean difference. Some use the control group value (giving *d*=0.83) and some 'pool' the two values (giving *d*=0.89 for one way of pooling). Borenstein et al. (2009) describe this process, including formulae for many variants of standardised effect sizes and the mechanisms by which these are then combined in a meta-analysis. They note other possible adjustments to the calculation (for example, to compensate for a bias with small samples), but the detail of these does not affect the argument in this paper.

The key issue is that meta-analyses and meta-meta-analyses in education are generally based on some variant of *d*: the difference in mean scores of the experimental and comparison groups, divided by some measure of the extent to which those scores vary within the groups (some variant of standard deviation).

There are notions of 'raw effect size' (that is, working with the difference between group means, without dividing by any measure of spread) and one can even undertake meta-analyses using raw effect sizes (Bond et al. 2003) which addresses some but not all of the issues discussed below. However, all of the key educational meta-meta-analyses use standardised effect size, almost always in the form of standardised mean difference. So this paper will follow the practice of these analysts in using 'effect size' as shorthand for standardised mean difference.

4 The assumptions of educational meta-analysis

The use of effect size in educational research (amongst other areas) came in response to concerns with null hypothesis significance testing (NHST). This is the process of setting up a statistical model (for example, that the two groups of scores are random samples from the same population) and then using a test (such as the t-test) to indicate how incompatible the data is with that model, normally by choosing some significance level (α=0.05) below which the probability of such a test outcome resulting from the model is low enough to draw the conclusion that the data is incompatable with the model (e.g. the two groups of scores should be considered as having not come from the same population and thus that the difference is 'statistically significant').

The NHST process is controversial (Carver 1978; Nickerson 2000) and the American Statistical Association has recently released guidelines to try to ensure that NHST and the associated probability (p-value) are treated appropriately by researchers (Wasserstein and Lazar 2016). But it is not this controversy as such which led to the introduction of effect size. Instead, it was the realisation that statistical significance did not align with educational importance. A large enough sample size can result in a study drawing the conclusion that there is a difference between two groups when, from an educational perspective, that difference is trivial. FitzGibbon (1984) argued 'to interpret the educational significance of a difference between two groups one must interpret the difference in terms of the metric in which the outcomes were measured' (136) and suggests the use of effect size. While Cohen (1962) developed effect size as a way of thinking about the power of a study, for educational meta-analysts and meta-meta-analysts it is assumed that standardising "allows studies using different outcome measures to be compared using the same metric" (Higgins et al. 2005, 14). In particular, it is assumed that effect size "is the most important tool in reporting and interpreting effectiveness, particularly when drawing comparisons about *relative* effectiveness of different approaches." (Higgins et al. 2013, 6).

There have been many critiques of the meta-analytic methods, notably Eysenck's (1984) argument that they are "adding apples and oranges" (57): that is, that meta-analyses combine outcome measures which are conceptually incomparable. While there is much merit in that argument and it often applies to educational meta-analyses, the argument here is focussed on the metric in which those outcome measures are expressed: the effect size.

The assumptions of educational meta-analysis and meta-meta-analysis then are that interventions with larger effect sizes are generally associated with greater educational significance (that is, effect sizes can be compared across different studies) and that two or more different studies (potentially with different interventions in the same category, on different samples and with different outcome measures) can have their effect sizes combined to give a meaningful estimate of the educational significance of interventions within that category.

5 Violating the assumptions of meta-analysis

The remainder of this paper shows that these assumptions are not met: a larger effect size does not indicate a larger educational impact for the intervention, so one cannot generally compare across studies, and one cannot reasonably combine studies to obtain an estimate of the impact of a class of interventions. Moreover, this section shows that these violations do not cause random fluctuations in reported effect size: they are not just noise which gets factored out when large numbers of studies are combined. Instead, there is systematic and unadjusted bias in the violation of these assumptions: good experimenters legitimately manipulate *d* (as they might manipulate sample size) to enhance the sensitivity of their experiments to the impact of the interventions, but their freedom to do so varies between educational contexts. So league tables which purport to rank more or less effective categories of educational interventions are, if anything, ranking more or less easily manipulated experimental scenarios.

As such, using these league tables to drive educational policy is misguided.

Three areas in which experimental design leads to the violation of the assumptions of meta-analyses are detailed in the sections below: the role of comparison groups; the range of the population from which the sample is taken and the design of the measures. In each section, there is an illustration of the underlying issue. This is placed in a simple, but deliberately non-educational context in which (unlike many educational contexts) the raw outcomes are easy to see and compare to one another. These illustrations are intended to show that, even in experimental situations which are much simpler than the majority of educational interventions, reported effect sizes can vary widely on the basis of experimental design decisions. Even when one should argue that outcomes are identical, effect sizes are very different.

Each section also provides references to just a small sample of the huge number of studies and meta-analyses that suffer from each of the issues discussed.

Recall that the issues are not problems inherent in the original studies: providing an estimate of effect size when reporting the outcomes of an intervention study may be good practice to support future power analyses for replication studies. Problems arise when those estimates are taken to be indications of the relative importance or impact of different interventions across studies with different samples and different measures. That is, they become problems only when meta-analysts and meta-meta-analysts compare and combine effect sizes inappropriately.

5.1 Comparison groups

Fundamental to intervention studies is the contrast between an ‘experimental’ and a ‘comparison’ (or control) group.

A simple thought experiment highlights a serious problem for comparing or combining effect sizes when comparison groups differ between studies: Consider a farmer who wishes to find out if a fertilizer increases the mean length of beans on her plants. She plants two rows of beans: one (the experimental row) she treats with her new fertilizer and on the other (the comparison row) she uses no fertilizer at all. Otherwise, she treats the two rows equally. On her comparison row the beans grow to a mean length of 10cm with a standard deviation of 1cm; on her experimental row the beans have grown to a mean length of 11cm with a standard deviation of 1cm. So she reports an effect size of d=1.

A second farmer believes he has a fertilizer which is better than manure at increasing bean length. He plants two rows of beans and treats the experimental row with his new fertilizer. The comparison row he treats with manure. He otherwise conducts the trial using the same protocols and measures as the first farmer. He finds his comparison beans grow to an average of 10.5cm with a standard deviation of 1cm and his experimental beans grow to an average of 11cm with a standard deviation of 1cm. That is, he reports an effect size of d=0.5, half that of the first farmer.

We cannot, however, conclude that the first farmer’s fertilizer has a larger impact on bean length than the second farmer’s. Nor can we combine the two *d* values to provide a meaningful estimate of the effectiveness of fertilizer. The farmers were comparing to different controls (the first to no fertilizer and the second to manure). Both were legitimate studies, with slightly different underlying questions.

Unequal comparison groups are commonplace in the studies included in educational meta-analyses and meta-meta-analyses. This is understandable: where it is possible to make an active choice about a comparison, it is not always obvious what it should be. For example, if a researcher wishes to study the effect of providing feedback to pupils on their essays by text message, the comparison group could be given no feedback, identical length feedback written at the end of each essay, or feedback using their teacher's usual system. Each may be legitimate, depending on the research aim, though studies may also be constrained by practical and ethical considerations: away from a laboratory setting, it may be unacceptable to provide no feedback at all to a class.

If two studies report the effect size of text message feedback, we cannot reasonably compare them or combine them if one study uses no feedback and the other uses short written feedback.

Yet combining studies with very different choices of comparison group is common in many meta-analyses. For example, in one meta-analysis (Paschal et al. 1984) studies are combined comparing homework with no-homework (Gray and Allison 1971), enriched homework with usual homework (Singh 1970) and even homework in which every question was graded with homework where a random half of the questions were graded (Austin and Austin 1974). This meta-analysis is then combined in the EEF 'homework' category with another (Cooper et al. 2006) in which only studies with 'no homework' comparison conditions were accepted.

The same issue occurs across many different areas in the EEF 'toolkit'. For example, Graham et al. (2015) produced a meta-analysis of studies in formative assessment supporting learners' writing performance in grades 1-8. This included many studies with different comparison groups including:

- the experimental group's summaries of texts were given graphical feedback on latent semantic analysis while the comparison group were given feedback on length alone (Wade-Stein and Kintsch 2004),
- the experimental group were given peer feedback on draft writing while the comparison group were given feedback from a teacher on spelling and mechanical errors as well as a general statement about clarity (Prater and Bermudez 1993),
- the teachers of the experimental group were taught to use curriculum based approaches to recording and using assessment and the comparison group's teachers used standard forms to set goals on teacher-made spelling tests (Fuchs et al. 1991)

This meta-analysis was combined in a meta-meta-analysis with other meta-analyses, including one which, unusually, has taken care to ensure each study had equivalent comparison groups (Bangert-Drowns et al. 1991).

However, the problem of unequal comparisons can become more subtle than this: 'business as usual' comparisons. This is where the only information given about the comparison group is simply that the normal regime was followed.

In Graham et al.'s (2015) meta-analysis, many studies did not specify the comparison or control groups, either giving no indication or suggesting a business as usual comparison. For example, Ross et al. (1999) used an intervention in which students were taught some self-evaluation methods while the comparison group teachers ‘continued teaching language as they usually did, including self-evaluation if that was part of their practice, but not emphasizing it’ (117). Meyer et al. (2010) compared students using an electronic portfolio tool to a comparison group where the only information given in the published report is that their teachers had been incentivised with a small stipend.

Using unequal comparisons or using unspecified ones makes it impossible to compare or combine effect sizes meaningfully. However, one might note a further issue in the examples given above: while Bangert-Drowns et al. (1991) includes only ‘no feedback’ comparisons, the interventions are also highly constrained. Two thirds of the studies used feedback consisting simply of giving participants the correct answer. However, in other meta-analyses, most studies had comparison groups with something more than ‘no feedback’ including being given correct answers.

That is, the experimental condition in some studies and meta-analyses is the comparison condition in others.

In some cases, meta-analysts undertake ‘moderator analyses’ (analysing subsets of studies for differences in effect size). For example, Camilli et al. (2010) analysed separately those studies which compared early years interventions to comparison groups ‘who received no intervention or an unsystematic intervention’ (592) from those which compared intervention groups to comparison groups who had an alternative treatment. Unsurprisingly, the calculated average effect size for the former was higher: over three times as large. Few other meta-analyses examine this issue and it is not addressed in meta-meta-analyses.

The choice of comparison group is part of the research design process: the researcher has a particular question in mind, often if one strategy is better than another, whether that ‘other’ is well specified or simply business as usual. In some contexts it may be possible (and ethical) to have a comparison group who receives no element of the strategy, but in others it may not. For example, it is possible and in some circumstances may be ethical to provide a comparison group with no feedback (for example, in laboratory based experiments), but most behavior intervention studies take place in schools and it would be unethical to use comparison groups with no behaviour intervention at all. Studies in behaviour interventions compare more intensive interventions with existing (business as usual) ones, while many studies in feedback compare different feedback strategies to no feedback. So, ceteris paribus, we might expect that effect sizes across a collection of feedback studies to be higher than the effect sizes across a collection of behaviour intervention studies simply because of the available research design choices for comparison groups. One ought not conclude from a comparison of combined effect sizes that feedback is more effective at improving educational outcomes than behaviour interventions.

5.2 Range restriction

The unequal comparison problem affects comparing or combining any form of effect size, whether raw or standardised. However, some issues arise particularly from the process of standardising: recalibrating mean differences in terms of some form of standard deviation. Researchers can make legitimate design decisions which alter the standard deviation and thus report very different effect sizes for identical interventions. One such design decision is range restriction.

In another variant of the thought experiment, the first farmer goes to the nursery to select plants for her trial and she excludes all those plants which have particularly long or short young beans, selecting from those where the mean length of the young beans on the plant are all about the same. The second farmer undertakes exactly the same trial (with the same fertilizer and the same experimental and comparison regimes) with the sole exception that he chooses entirely at random from all the plants in the nursery.

At the end of their trials the first farmer will report an effect size very much larger than the second. This is because the first farmer chose to restrict the range of her sample. While both farmers may find similar mean differences in average bean length, the first farmer will have a smaller variance. So she divides by a smaller number in calculating Cohen's *d*.

Roughly speaking, in this thought experiment, there are two sources of variance at work: the random variance by which bean lengths vary anyway and the variance which might be attributed to the improved fertilizer regime. By substantially reducing the former, the standard deviation of the bean lengths is reduced and the effect size is much larger. However, the effect of the fertilizer on the beans is identical in both cases.

We cannot argue that the fertilizer was more effective in the first case than the second – it was the same fertilizer – even though the effect size reported from the first study was much higher. It makes no sense to combine the effect size values to give a meaningful estimate of the effect of this fertilizer.

One can correct for range restriction (Thorndike 1949), though it requires good estimates of parameters relating the sample to the population, which is rarely available. Moreover, the original research studies would not be expected to correct for it, since it is irrelevant to their aim. However without adjustment one study may appear to show a larger effect size than another because the first was conducted on a restricted range, not because the intervention had a bigger educational impact. Indeed, whole families of studies noted below are conducted, deliberately and for good reason, on restricted ranges, but are then compared and combined with families of studies which are either unrestricted or restricted in different ways, with no attempts to adjust.

Any study where a sample is chosen using some criteria (whether explicit or implicit) which correlates with the outcome measures will result in an effect size larger than that from an identical study conducted with a random sample from the whole population. This particularly affects education because it is common to group pupils in classes for many subjects according to some achievement measure and researchers often use these classes for reasons of practicality and convenience.

On the basis of previous mathematics tests, say, pupils may be placed in classes of similar mathematical achievement. If a subsequent study uses any form of mathematics test as the outcome measure and is conducted with any one of the sets, the range of scores they will achieve on the test will be more restricted than the same study conducted across the year group, and the effect size reported would be much larger.

Figure **Error! Reference source not found.** illustrates this. It simulates a population where a particular study with broad, fully representative sampling would result in d=0.5. The simulation considers a year group split into four ability sets (classes 1-4) on the basis of a previous test which correlates well, but not precisely, with the study measures (r=0.7).

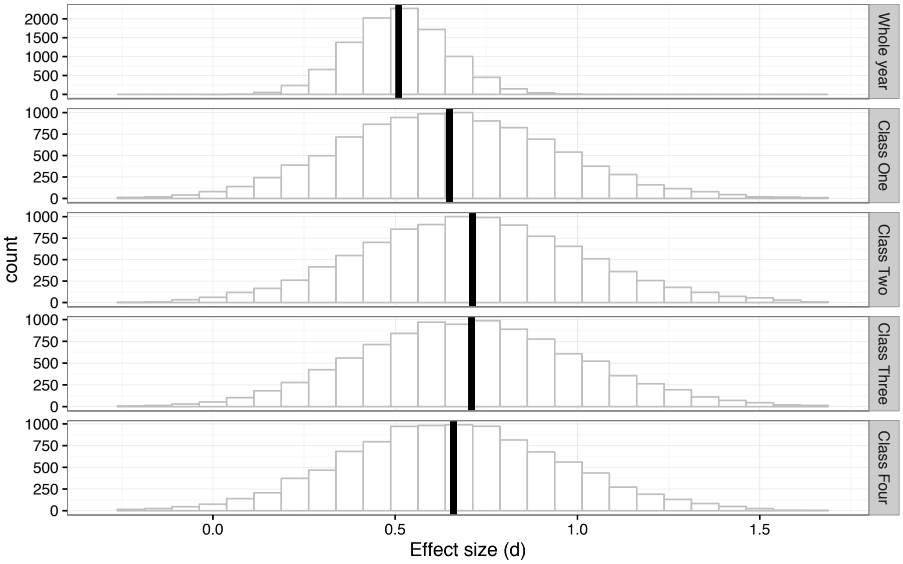

Figure 1: A simulation of range restriction

Consider two scenarios: In one, we conduct four separate studies, one for each set. In the second we conduct the same study with the whole year group. We randomly assign pupils to the experimental or control conditions. In both cases, the intervention is the same and the impact on every pupil's score is identical. The only difference is how the data is analysed.

Figure 1 shows the distributions of the resulting effect sizes for 10000 such simulated studies. As expected the studies across the whole year group have mean effect sizes clustered around 0.5. However, the studies of the highest and lowest ability groups have mean effect sizes about 30% higher (around 0.65) and the two middle ability groups around 40% higher (around 0.7).

A number of factors can affect how restricted range impacts effect size. Selecting disproportionately from the middle of a normal distribution results in a even higher effect size than selecting from the tails (as shown in figure **Error! Reference source not found.** where the study conducted on the middle sets give a higher mean effect size than the top and bottom sets). The closer the correlation between the selection measure and the outcome measure, the higher the inflation of restricted range studies over whole population studies. Figure **Error! Reference source not found.** shows the graph of the inflation factor for a study using the top or bottom set (of four) or the middle two sets (of four) compared to a study using the whole year group, across different correlations between the selection criterion measure and the study measure.

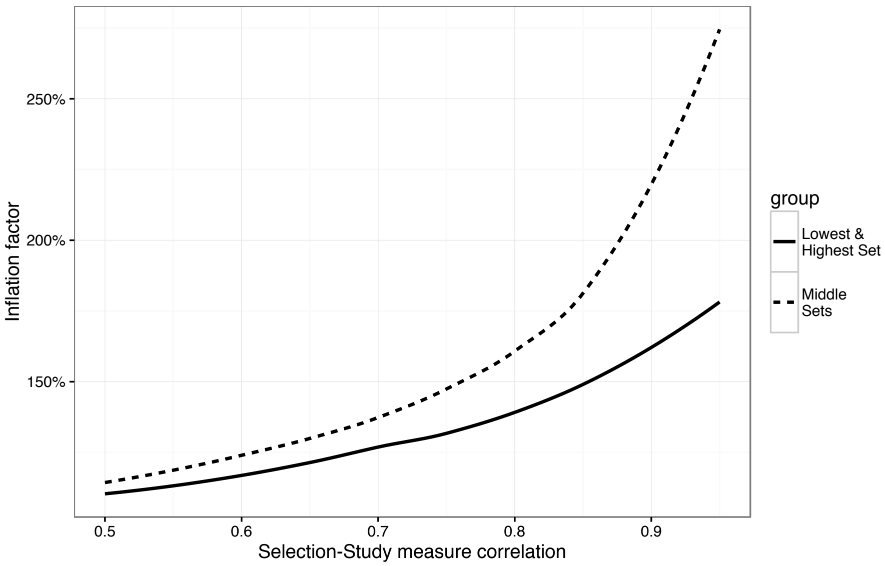

Figure 2: The impact of selection-outcome measure correlation on effect size

For example, the correlation between tests of mathematics at age 14 and public mathematics examination scores at age 16 in the UK is reported as around 0.85 (Sammons et al. 2014). So if a study was conducted on the top or bottom mathematics set of four, based on their mathematics achievement at age 14 and using the public mathematics examination grades as an outcome measure, the effect size would be 50% higher than the identical study (with the identical intervention and identical changes in each individual pupil's score) conducted across the year group. If the study used one of the middle sets of four, the effect size would be around 75% higher. Even if the selection was apparently less well related (say classes selected by their English performance but studied on their mathematics performance) we would still expect effect sizes inflated by 30-50% for restricted ranges compared to the whole population.

There appears to be no evidence that either meta-analyses or meta-meta-analyses in education adjust for range restriction (even though the issue is noted by Higgins et al. 2013). That is, for each original study which restricts the range of their sample, the effect size used in the meta-analysis and meta-meta-analysis is larger than would be expected for the identical study on the population as a whole and, without adjustment, should not be compared to or combined with another study with a sample with a different range.

This can result directly in meta-analysts drawing inappropriate conclusions by failing to consider the issue of range restriction. For example, Li and Ma (2010) looked at the effect of computer technology on mathematics achievement, noting that the effect size for studies restricted to pupils with special needs was higher than those undertaken with general education pupils. Rather than considering whether this is a result of range restriction, they argued 'technology was strongly more effective in promoting mathematics achievement when used to help special need students than to help general education students' (226).

Of course, if a given meta-analysis included only studies which sampled entirely randomly from a consistent, agreed target population, then they would not need to adjust for range restriction, but education meta-analyses do not appear to do this. Moreover, even if a particular meta-analysis did this, one could only combine or compare effect sizes with other meta-analyses with exactly the random sampling from the same target population (and then only if they also address the problems of identical comparison groups and, as discussed below, test focus and precision).

In fact the problem of range restriction is seen in a large number of the studies which are ultimately summarised in meta-meta-analyses. In a number of cases, studies are conducted where the selection criteria directly correlate with study measures. Some examples taken within the EEF 'feedback' collection include:

- A study of whether verbal feedback improved performance on area and volume conservation tasks using a sample of those who had failed the tasks in an earlier study (Hornblum and Overton 1976).
- A study of the influence of feedback on a maze completion task using a sample of those who had performed above a given level on the task (Neenan and Routh 1986).
- A study of the effect of trial by trial feedback on recall. The study used a group selected for their failure to attend some trials and then restricted the range still further by splitting the groups into those with low pretest scores and those with high pretest scores (Lacher 1983).
- A study which examined the effect of feedback on children's toothbrushing performance using children where 'a large number of caries present in the children at annual check-ups suggested that appropriate toothbrushing skills were not being performed by the children' (Murray and Epstein 1981, 362).

That is, researchers were designing their studies to look at the impact of an intervention on samples chosen specifically because they scored in a restricted range on the measure the feedback condition was hypothesised to improve. That is very direct range restriction and likely to lead to effect sizes very much larger than identical studies on samples chosen from the wider population.

In a very large number of studies, while the selection criterion did not match the study measure this closely, they would still be expected to be well correlated. In some cases this comes from selecting or excluding sections of the population on correlated measures or choosing a sample from a population who are not fully representative of the whole population. E.g.

- A study of the effect of failure feedback on task performance, taking students in the top and bottom third on a test of locus of control (Lewis-Beck 1978).
- A study of the improvement in legal hit rates following feedback interventions in an ice hockey team with a notable losing record, and, where within this group, the best and most experienced players were excluded (Anderson et al. 1988).
- A study of the impact of feedback on measures of intrinsic motivation. While a wide range of students took part in the study, the analysis used only from those 'whose average grade in language and mathematics on their most recent report card was in the top or bottom 25% for their class' (Butler 1987, 476).

Again, each of these studies reflects a research design choices but the effect sizes reported will be larger than those expected from identical studies with unrestricted samples. Without adjustment, these studies cannot be meaningfully compared or combined, yet all of the above studies are combined without adjustment in just one meta-analysis (Kluger and DeNisi 1996) in the EEF ‘feedback’ collection, which is then compared to other intervention areas.In some cases, meta-analyses collect studies which are deliberately range restricted, because the analysts have a particular focus. For example, Elbaum et al. (2000) undertook a meta-analysis focussed on the effectiveness of one-to-one tutoring on reading for pupils at risk of reading failure. That is, they chose studies whose ‘participants were elementary students identified as at risk for reading failure, scoring in the lowest 20-30 percentile on grade level reading assessments, or possessing learning disabilities’ (606). However, in the EEF toolkit, the results of this meta-analysis are combined in the ‘one-to-one tutoring’ theme, with those from other meta-analyses which do not have a restricted focus and without adjusting for the issue.

The choice of sample is part of the research design process: the researcher may have a reason to choose a particular range of participants or they may simply want to maximise their chance of finding a significant difference between groups by reducing within-group variance. In some contexts it may be easier to restrict the range of a study than in others. For example, it may be easy to choose a very restricted range when one is studying feedback, but if one is studying the effect of extending the school day or having a school uniform, there is no obvious way in which the study can be conducted on just a restricted sample of pupils. So, ceteris paribus, we might expect that effect sizes across a collection of feedback studies to be higher than the effect sizes across a collection of school uniform or extended school day studies simply because of the available research design choices for the sample. One ought not conclude from a comparison of combined effect sizes that feedback is more effective at improving educational outcomes than, say, extending the school day.

5.3 Measure design

Range restriction impacts on effect size by reducing the within-group variance and thus reduces the standard deviation in the calculation of d. That is, researchers can impact effect size through the choice of sample. However, researchers can also directly impact effect size through the choice of measure in at least two ways: its focus and its precision. In simple terms, a researcher can increase their chances of finding a significant difference between groups if they use a test tied closely to the nature of the intervention or if they increase the number of test items.

The extent to which the measure is focussed on the proposed impact of the intervention can alter the effect size (whether standardised or raw). Again, the researcher can have legitimate reason to choose measures more or less tightly bound to the intervention, depending on their research question, but this makes it impossible to meaningfully compare and combine studies on the basis of effect sizes.

Consider another variant of the thought experiment. Assume that, unknown to the farmers, the fertilizer is only effective on beans which are exposed to direct sunlight, not those shaded under the leaves of the plant. Assume further that both farmers undertake identical trials with the same fertilizer, comparisons and samples. However, the first farmer selects the beans to measure from those which are easy to reach - that is those that tend to be exposed to sunlight - while the second farmer selects from across the plant including those hidden under leaves. The first farmer will report a larger mean difference (and thus a larger effect size) compared to the second, since all of the beans in her sample on each plant have been affected by the improved fertilization while the second farmer will have included many beans which have not.

Again, the higher effect size reflects only that the measure was more tightly focussed on the beans which the fertilizer affects, not that the fertilizer was more effective.

The same issue of test focus can occur in educational studies. For example, an intervention may (successfully) target pupils' understanding of algebra, with little impact on other areas of mathematics. One research team may measure the participants' marks on a standard, public mathematics examination (including both algebra and non-algebra questions). Another research group may use the marks on their own test (focussed entirely on algebra). Both research teams may do so for good reason. The first may be interested specifically on the impact of a classroom intervention on a publicly accepted measure, the second may be interested specifically on the impact on algebra. Alternatively, the two research groups may simply have theorised the impact of the intervention differently, one believing that the intervention affects many areas of mathematics and one believing that it mainly targets understanding of algebra.

Across the studies included in meta-analyses and meta-meta-analyses in education, some measures are based on researcher designed tests while others use standard tests (albeit different standard tests in different studies and areas). One would expect the researcher designed tests to be more targeted on the hypothesised impact of the intervention. Across just one meta-analysis on meta-cognition (Higgins et al. 2005) there is a wide range of more or less tightly focussed tests:

- In a study of the effectiveness of problem solving techniques, researchers used a test with three 'creative' mathematics problems of a type identical to those in the intervention (Oladunni 1998).
- In a study of the effects of a problem based instructional model on earth science teaching, researchers used questions selected from a database of widely available earth science tests (Chang and Barufaldi 1999).
- In a study of the impact of teaching meta-cognitive skills in a mathematics context, researchers used a standard test: the Iowa Test of Basic Skills (Cardelle-Elawar 1992).

The extent of the test focus issue can be seen in the small number of meta-analyses with 'moderator analyses' for the nature of the measure. Normally they contrast studies which use standard measures against those which use non-standard (often researcher designed) tests.

For example, Abrami et al. (2008) undertook a meta-analysis of critical thinking instruction combining 161 different effects. Of these, 91 came from studies using standard measures, the calculated average effect size was 0.24; the other 70 had a calculated average effect size of 0.53. Of the 14 meta-analyses in the EEF collection which clearly report a moderator analysis for measure design, only two identified higher effect sizes for standard tests and on average, the effect size using non-standard testing was just under 40% higher than the effect size using standardised tests.

In addition to the focus of the test, its precision can also affect the reported effect size for identical interventions.

Consider a further variant of the thought experiment. Suppose the two farmers undertake identical trials of the same fertilizer against identical comparison regimes on similar samples (suitably representing the same range of the same population). The difference this time is in the measure they choose: the first farmer measures the mean length of 5 beans chosen at random on each plant and the second farmer measures the mean length of 10 beans chosen at random on each plant. The second farmer could report an effect size as much as 40% larger than the first.

There are two sources of variance of interest here: *within* the same plant beans will vary in length and *between* different plants the beans will vary in length. By choosing a larger number of beans on a given plant to measure, the second farmer gets a more precise estimate of the mean length of beans on that plant (he reduces the contribution of the within-plant variance). Again this means that while the mean difference between the experimental and comparison plants will be about the same, to calculate the effect size we divide by the standard deviation which will be smaller for the second farmer. That is, while the intervention is identical, the reported effect size for the second farmer is much larger.

The same happens in research studies in education with different numbers of items in the outcome measure. Adding questions to a test tends to increase effect size by increasing the precision with which an individual's achievement is measured. In the case where every question is independent of every other question, the inflation effect will be large: multiplying the length of the test by four will double the effect size (as the mean difference increases fourfold, but the standard deviation only doubles).

In reality, responses to test questions are not independent, they tend to have a relationship to one another (that is, tests tend to have some internal consistency). Figure **Error! Reference source not found.** shows the results of a simulation of four different situations: tests consisting of 2, 4, 10 or 20 questions. In the comparison group, every student gets a mean mark of 10 on each question; in the experimental group, every student gets a mean mark of 11 on each question (and the pooled standard deviation for each question is 2). That is, for a test consisting of a single question, the effect size would be expected to be 0.5. The mark given to each student for the test as a whole is the mean of the marks on each question. The proportion by which the effect size is inflated compared to the one-question test is plotted against the level of internal consistency (given as Cronbach's alpha, the most commonly reported test of internal consistency).

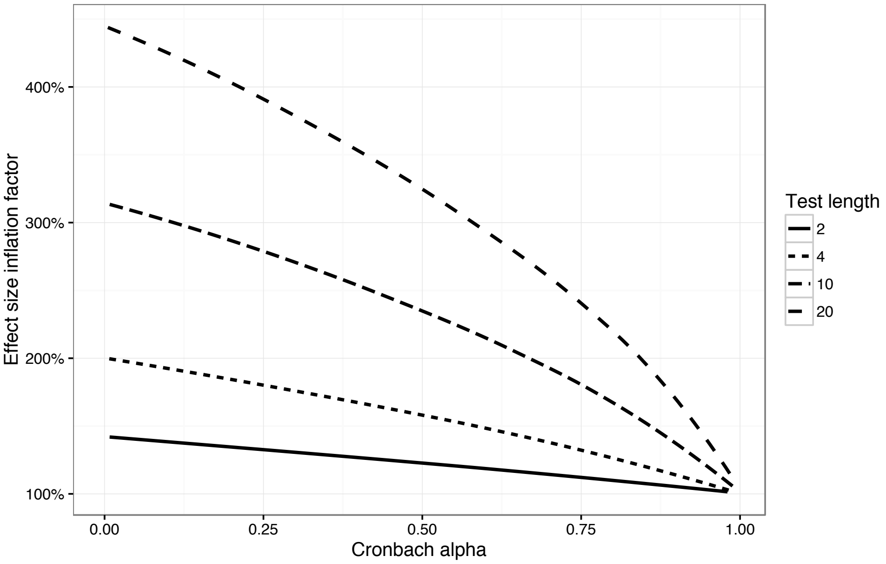

Figure 3: The impact of test length on effect size

That is, for a relatively well designed test (say with α=0.75) we can expect that having two questions instead of one inflates the effect size by 20% and having twenty questions more than doubles the effect size.

The number of trials or test items can vary widely. For example, from the studies combined by Kluger and DeNisi (1996) there are:

- Six tasks each scored from 0-5 (Hornblum and Overton 1976).
- Ten trials of picture recall (scored 0 or 1) (Lacher 1983).
- Fifteen multiple choice questions (0 or 1) (Arkin and Schumann 1984).
- Twenty multiple choice questions (0 or 1) (Lewis-Beck 1978)
- Three hundred arithmetic tasks (0 or 1) (Dossett et al. 1979).

Figure 3 shows that, theoretically, adjusting for test length is possible, but it requires knowledge of the test reliability. In fact, few studies in the EEF toolkit report a suitable statistic for internal consistency. The study by Dossett et al. (1979) is an exception, with a reported α=0.73, but it would be reasonable to assume that many studies which do not report such statistics will have lower consistency and thus be more affected by test length.

There is no evidence of any adjustment made for number of trials or test length at either the meta-analysis or meta-meta-analysis level. Dossett et al. (1979) used a correct/incorrect measure 20 times the length of Arkin and Schumann (1984) and thus, had the tests otherwise been comparable and with similar internal consistency, conducted on comparable samples from the population against comparable control groups, the former would have reported an effect size well over twice as large.

The choice of test is clearly part of the research design process: either because of their research question or practicality, a researcher may choose a particular length of test and may either choose a standard test or design their own. In some contexts it may be easier to design more focussed tests than in others. For example, it may be easy to design a test which is highly focussed on the particular academic discipline in which a feedback intervention study is conducted, but if one is studying early years interventions, one would expect to use standardised measures. Indeed, of the 68 effect sizes which can be clearly identified in the feedback meta-analyses of Bangert-Drowns et al. (1991) and Kingston and Nash (2011) only 9 used standard tests; while of the 39 effect sizes which can be clearly identified in the early years intervention meta-analyses of Nelson et al. (2003) and Anderson et al. (2003) 37 used standard tests. So, ceteris paribus, we might expect that effect sizes across a collection of feedback studies to be higher than the effect sizes across a collection of early years intervention studies simply because of the ease with which researchers can design precise and focussed tests. So one ought not to conclude from a comparison of combined effect sizes that feedback is more effective at improving educational outcomes than, say, early years interventions.

6 Conclusions

Recall that the notion of (standardised) effect size was introduced to help researchers make reasonable decisions about their chance of detecting a difference between experimental and comparison groups (should one, in fact, exist) - that is, the power of a study. It still has that important role, but researchers can also legitimately directly manipulate effect size when they are looking to increase their chance of detecting a difference.

For example, it is good experimental design to reduce variance other than that induced by the intervention: Lipsey and Hurley (2009) recommend minimising within-group variation to make a treatment difference stand out more clearly. If the aim is to detect a difference which might otherwise be small, this is entirely sensible advice, but as shown above it can dramatically increase the standardised mean difference.

Indeed, it might be argued that 'effect size' is badly named. It is not simply a measure of the size of an effect at all and it might have been better named 'effect clarity': a large *d* indicates that, for that particular intervention, between the two groups used and on the measure selected, the difference is very clear. But that does not mean the difference is large or important or educationally significant.

In each case, the arguments above shows that identical interventions can lead to dramatically different standardised mean differences. By contrasting them with different comparison groups, by measuring them on samples selected with a measure which correlates with the outcome measure, by increasing test length or tightening the focus of the test on the intervention, the difference becomes clearer, but it does not mean that the intervention gives more 'educational bang for their buck' (Higgins et al. 2013, 3).

If one wishes to make judgements about more or less effective educational interventions, then studies must use the same comparisons, measures and range of participants. Indeed, in such cases, one can use raw (unstandardised) effect sizes anyway (Bond et al. 2003). One cannot compare standardised mean differences between sets of studies which tend to use restricted ranges of participants with researcher designed, tightly focussed measures and sets of studies which tend to use a wide range of participants and use standardised tests as measures.

These issues are not noise randomly affecting the categories of studies and which might be factored out by combining large numbers of studies. They vary systematically with the educational areas which are more or less susceptible to sensitive research design: it is easier to design a study on a restricted range, with a tightly focussed measure in studies of feedback or meta-cognition than in studies of extending school time. So one should not conclude that feedback or metacognition are more effective at enhancing educational outcomes than extending school time on the basis of combining and comparing *d* values from these areas. One should say, instead, that differences tend to be clearer or easier to spot in feedback than extending school time, irrespective of their educational importance.

The argument here is not simply that educational policy has fallen prey to metricophilia: the unjustified faith in numerical quantities as having particularly special status as 'evidence' (Smith 2011), though that is a real concern. The main issue is that whatever the quality of evidence provided by the aggregation of effect sizes in the manner seen in Marzano (1998), Hattie (2009) and the EEF toolkit (Higgins et al. 2013), it is not evidence of more and less effective educational interventions, it does not indicate where there is more bang for our educational buck. It is evidence of research areas which are more or less susceptible to research design manipulation: areas where it is easier to make what may be educationally unimportant differences stand out through methodological choices. That is, standardised effect size is a research tool for individual studies, not a policy tool for directing whole educational areas.

These meta-meta-analyses which order areas on the basis of effect size are thus poor selection mechanisms for driving educational policy and should not be used for directing large portions of a country's education budget.

Acknowledgements

I am grateful for the insightful comments on earlier drafts from a number of colleagues, including Nancy Cartwright, Matthew Inglis and Sean McCusker, as well as those from the anonymous reviewers.

References

Abrami, P. C., R. M. Bernard, E. Borokhovski, A. Wade, M. A. Surkes, R. Tamim, and D. Zhang. 2008. "Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis." *Review of Educational Research 78*(4): 1102–1134.

Anderson, D. C., C. R. Crowell, M. Doman, and G. S. Howard. 1988. "Performance posting, goal setting, and activity-contingent praise as applied to a university hockey team." *Journal of Applied Psychology 73*(1): 87–95.

Anderson, L. M., C. Shinn, M. T. Fullilove, S. C. Scrimshaw, J. E. Fielding, J. Normand, and V. G. Carande-Kulis. 2003. "The effectiveness of early childhood development programs: A systematic review." *American Journal of Preventive Medicine 24*(3): 32–46.

Arkin, R. M. and D. W. Schumann. 1984. "Effects of corrective testing: An extension." *Journal of Educational Psychology 76*(5): 835–843.

Austin, J. D. and K. A. Austin. 1974. "Homework grading procedures in junior high mathematics classes." *School Science and Mathematics 74*(4): 269–72.

Bangert-Drowns, R. L., C.-L. C. Kulik, J. A. Kulik, and M. Morgan. 1991. "The instructional effect of feedback in test-like events." *Review of Educational Research 61*(2): 213–238.

Bond, C. F., W. L. Wiitala, and F. D. Richard. 2003. "Meta-analysis of raw mean differences." *Psychological Methods 8*(4): 406–418.

Borenstein, M., L.V. Hedges, J.P.T. Higgins, and H.R. Rothstein, *Introduction to Meta-analysis,* Chichester: Wiley.

Butler, R. 1987. "Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance." *Journal of Educational Psychology 79*(4): 474–482.

Camilli, G., S. Vargas, S. Ryan, and W. Barnett. 2010. "Meta-analysis of the effects of early education interventions on cognitive and social development." *Teachers College Record 112*(3): 579–620.

Cardelle-Elawar, M. 1992. "Effects of teaching metacognitive skills to students with low mathematics ability." *Teaching and Teacher Education 8*(2): 109–121.

Cartwright, N. and J. Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford: Oxford University Press.

Carver, R. P. 1978. "The case against statistical significance testing." *Harvard Educational Review 48*(3): 378–399.

Chang, C.-Y. and J. P. Barufaldi. 1999. "The use of a problem-solving-based instructional model in initiating change in students' achievement and alternative frameworks." *International Journal of Science Education 21*(4): 373–388.

Cohen, J. 1962. "The statistical power of abnormal-social psychological research: a review." *Journal of Abnormal and Social Psychology 65*(3): 145–53.

Cooper, H., J. C. Robinson, and E. A. Patall. 2006. "Does homework improve academic achievement? A synthesis of research, 1987 – 2003." *Review of Educational Research 76*(1): 1–62.

Cowan, N., N. Cartwright, B. Virk, and S. Mascarenhas-Keyes. 2015. *Making the Most of the Evidence: Evidence-based policy in the classroom (CHESS working paper 2015-03)*. Durham: Durham University.

Davies, P. 1999. “What is evidence-based education? ” *British Journal of Educational Studies 47*(2): 108–121.

Department for Education. 2016. *Educational Excellence Everywhere*. London: HMSO.

Dossett, D. L., G. P. Latham, and T. R. Mitchell. 1979. “Effects of assigned versus participatively set goals, knowledge of results, and individual differences on employee behavior when goal difficulty is held constant.” *Journal of Applied Psychology 64*(3): 291–298.

Elbaum, B., S. Vaughn, M. T. Hughes, and S. W. Moody. 2000. “How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research.” *Journal of Educational Psychology 92*(4): 605–619.

Eysenck, H.J. (1984). “Meta-analysis: an abuse of research integration” *Journal of Special Education* 18, 41-59.

FitzGibbon, C. 1984. “Meta-analysis: an explication.” *British Educational Research Journal 10*(2): 135–144.

Fuchs, L. S., D. Fuchs, C. L. Hamlett, and R. M. Allinder. 1991. “The contribution of skills analysis to curriculum-based measurement in spelling.” *Exceptional Children 57*(5): 443–452.

Graham, S., M. Hebert, and K. R. Harris. 2015. “Formative assessment and writing: A meta-analysis.” *The Elementary School Journal 115*(4): 523–547.

Gray, R. F. and D. E. Allison. 1971. “An experimental study of the relationship of computation with fractions.” *School Science and Mathematics 71*(4): 339–346.

Hattie, J. 2009. *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Abingdon: Routledge.

Higgins, S., E. Hall, V. Baumfield, and D. Moseley. 2005. *A meta-analysis of the impact of the implementation of thinking skills approaches on pupils*. EPPI-Centre, Social Science Research Unit, Institute of Education, University of London, London.

Higgins, S., M. Katsipataki, D. Kokotsaki, R. Coleman, L. Major, and R. Coe. 2013. *The Sutton Trust - Education Endowment Foundation Teaching and Learning Toolkit: Technical Appendices*. Sutton Trust and the Education Endowment Foundation

Hornblum, J. N. and W. F. Overton. 1976. “Area and volume conservation among the elderly: Assessment and training.” *Developmental Psychology 12*(1): 68–74.

House of Commons Education Committee. 2014. *Underachievement in Education by White Working Class Children*. London: The Stationery Office.

Kingston, N. and B. Nash. 2011. “Formative Assessment : A Meta-Analysis and a Call for Research.” *Educational Measurement: Issues and Practice 30*(4): 28–37.

Kluger, A. N. and A. DeNisi. 1996. “The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory.” *Psychological Bulletin 119*(2): 254–284.

Lacher, M. B. 1983. “Effects of feedback, instruction, and initial performance upon training and persistence of verbal rehearsal.” *The Journal of General Psychology 108*(1): 43–54.

Lewis-Beck, J. A. 1978. “Locus of control, task expectancies, and children’s performance following failure.” *The Journal of Educational Research 71*(4): 207–210.

Li, Q. and X. Ma. 2010. “A meta-analysis of the effects of computer technology on school students’ mathematics learning.” *Educational Psychology Review 22*(3): 215–243.

Lipsey, M. W. and S. M. Hurley. 2009. “Design sensitivity.” In L. Bickman and D. J. Rog (Eds.), *The SAGE handbook of applied social research methods*, Chapter 2, pp. 44–76. London: Sage.

Lumbelli, L., G. Paoletti, and T. Frausin. 1999. “Improving the ability to detect comprehension problems: from revising to writing.” *Learning and Instruction 9*(2): 143–166.

Marzano, R. J. 1998. *A Theory-Based Meta-Analysis of Research on Instruction*. Aurora, Colarado: Mid-continent Regional Educational Laboratory.

Meyer, E., P. C. Abrami, C. A. Wade, O. Aslan, and L. Deault. 2010. “Improving literacy and metacognition with electronic portfolios: Teaching and learning with ePEARL.” *Computers and Education 55*(1): 84–91.

Murray, J. A. and L. H. Epstein. 1981. “Improving oral hygiene with videotape modeling.” *Behavior Modification 5*(3): 360----371.

Neenan, D. M. and D. K. Routh. 1986. “Response cost, reinforcement, and children’s Porteus Maze qualitative performance.” *Journal of Abnormal Child Psychology 14*(3): 469–480.

Nelson, G., A. Westhues, and J. MacLeod. 2003. “A meta-analysis of longitudinal research on preschool prevention programs for children.” *Prevention and Treatment 6*(1): 1–67.

Nickerson, R. S. 2000. “Null hypothesis significance testing: a review of an old and continuing controversy.” *Psychological Methods 5*(2): 241–301.

Oladunni, M. O. 1998. “An experimental study on the effectiveness of metacognitive and heuristic problem solving techniques on computational performance of students in mathematics.” *International Journal of Mathematical Education in Science and Technology 29*(6): 867–874.

Paschal, R. A., T. Weinstein, and H. J. Walberg. 1984. “The effects of homework on learning: A quantitative synthesis.” *The Journal of Educational Research 78*(2): 97–104.

Prater, D. and A. Bermudez. 1993. “Using peer response groups with limited English proficient writers.” *Bilingual Research Journal 17*(1-2), 99–116.

Pring, R. 2004. *The Philosophy of Education*. London: Continuum.

Ross, J., C. Rolheiser, and A. Hogaboam-Gray. 1999. “Effect of self-evaluation on narrative writing.” *Assessing Writing 6*(1): 107–132.

Sammons, P., K. Sylva, E. Melhuish, I. Siraj, B. Taggart, K. Toth, and R. Smees. 2014. *Influences on students’ GCSE attainment and progress at age 16*. London: Department for Education.

Singh, J. M. 1970. “Research in homework as the motivating factor in reading achievement.” *Journal of Reading Behavior 3*(3): 51–60.

Smith, R. 2011. “Beneath the skin: Statistics, trust, and status.” *Educational Theory 61*(6): 633–645.

Thorndike, R. 1949. *Personal Selection*. New York: Wiley.

Wade-Stein, D. and E. Kintsch. 2004. “Summary Street: Interactive computer support for writing.” *Cognition and Instruction 22*(3): 333–362.

Wasserstein, R. L. and N. A. Lazar. 2016. "The ASA's statement on p-values: Context, process, and purpose." *The American Statistician 70*(2): 129–133.