This is an Accepted Manuscript of a paper to appear in *British Educational Research Paper*

Princesses are bigger than Elephants: effect size as a category error in evidence based education.

Adrian Simpson
School of Education,
Durham University

Abstract

Much of the evidential basis for recent policy decisions is grounded in effect size: the standardized mean difference in outcome scores between a study's intervention and comparison groups. This is interpreted as measuring educational influence, importance or effectiveness of the intervention. This paper shows this is a category error at two levels. At the individual study level, the intervention plays only a partial role in effect size, so treating effect size as a measure of the intervention is a mistake. At the meta-analytic level, the assumptions needed for a valid comparison of relative effectiveness of interventions on the basis of relative effect size are absurd. While effect size continues to have a role in research design, as a measure of the *clarity* of a study, policy makers should recognize the lack of a valid role for it in practical decision making.

The Size of Elephants

Adam photographs an elephant. The elephant's image covers 0.2 of the area of the photograph.

Belinda photographs another elephant. Her elephant's image covers 0.3 of the area of her photograph.

Simon compares these numbers, concluding the second elephant must be the larger.

Catherine collects many photographs of elephants. For each she works out the 'photo-size' of each elephant (the proportion of the photograph filled by the elephant's image) and, averaged across the collection, finds a photo-size of 0.18.

Douglas collects photographs of princesses. The average photo-size of princesses in his collection is 0.24.

Tabitha compares these numbers, concluding princesses are bigger than elephants, cautioning this does not mean a particular princess is bigger than any particular elephant: this is about averages.

Uri draws together Catherine's work (with other collections of elephants), Douglas's work (with other collections of princesses) with averaged photo-sizes from multiple collections of microbes, politicians, sea-creatures, white rhinos, ants and many other categories. He produces a league table: politicians and microbes towards the top, ants and white rhinos

towards the bottom. This league table is promoted as the 'best bet' for indicating which creatures are really bigger or smaller, with caveats about assumptions needed to interpret it.

This story is clearly designed to expose the argument's absurdity. An object's physical size plays only a partial role in photo-size, so it is a category error to treat relative photo-size as a proxy for relative physical size. Only if Simon had reason to believe Adam and Belinda used the same camera and lens and had stood the same distance away, might he legitimately argue that Belinda's elephant is the larger.

Similarly Tabitha's conclusion that relative averaged photo-size can act as proxy for relative averaged actual size relies on strong assumptions: that other elements affecting photo-size are distributed equally for photographs of princesses and for photographs of elephants.

The same assumptions are needed for Uri in comparing actual average sizes of classes of creatures on the basis of relative averaged photo-sizes. Not only are these heroically strong assumptions (which Uri leaves unchecked), it is clear they cannot be met. They require that the design decisions of photographers are distributed equally across areas, but photographers do not use the same cameras for microbes and politicians; nor stand at the same distance when photographing white rhinos and ants. Design decisions vary systematically between areas, so the argument for using relative photo-size as relative actual size is invalid.

This paper will show identical, fundamentally flawed arguments underpin much of the 'evidence based education' movement. First, effect size does not measure the effectiveness of an intervention (nor its educational importance or influence) since the intervention plays only a partial role in the calculation of effect size. Second, when comparing studies, relative effect size can be a proxy for relative effectiveness of interventions only in the highly restricted circumstances that all other factors impacting on effect size are equal. Third, when comparing groups of studies, relative averaged effect size can be a proxy for relative average effectiveness for types of intervention only in the highly restricted circumstances that all other factors impacting on effect size are distributed identically across those groups of studies. While meta-meta-analysts may assume those circumstances hold, they do not: instead, these factors vary systematically between types of intervention.

Evidence Based Education

There is a growing movement in educational research and policy described as 'evidence based education' or 'evidence informed education'. The UK government promotes an evidential hierarchy which sees randomised controlled trials (RCTs) and meta-analyses of trials as the highest standards of evidence for policy makers (Campbell and Harper, 2012).

A key measure dominates this approach to policy: 'Effect size' [1]. These are reported in (or can be calculated from) individual studies; studies are collated according to some criteria

[1] The paper uses 'effect size' to mean standardised mean difference (in keeping with most educational literature), retaining 'raw effect size' for the difference between the mean scores of the groups (unscaled by a measure of spread). Working with raw effect size (or other forms of effect size, such as odds ratios and correlation coefficients) addresses some but not all of the issues raised.

and their effect sizes averaged in meta-analyses. Meta-analyses are further aggregated on broad educational areas, in meta-meta-analyses, from which rankings are constructed. For example, the Educational Endowment Foundation (EEF) produces a 'teaching and learning toolkit' purporting to give "information about the relative effects of different approaches to improving learning" (Higgins and Katsipataki, 2016, p.237). Hattie (2009) used a similar meta-meta-analysis to indicate the 'relative efficacy of different influences that teachers use'[2] (p.6). 'Relative' in both statements means relative averaged effect size.

These meta-meta-analyses are heavily promoted to policy makers. For example, a recent UK government paper encourages teachers to use the toolkit, arguing it "sets out what works and what doesn't" (Department for Education, 2016, p. 37), with the toolkit being consulted by nearly two thirds of school leaders (NAO, 2015).

Concerns about meta-analyses and meta-meta-analyses have been raised from the beginning. Many questioned the coherence of the collection of studies (the so-called 'apples and oranges' issue, Eysenck, 1984). Certainly there are categorisations that make little sense: the EEF area 'mastery learning' collects studies on Bloom's learning for mastery and Keller's personalised systems of instruction with mathematics mastery approaches from Singapore and Shanghai. The last concept has little in common with the first two: they simply share a name. Linda Wang (personal communication) likens this to combining a nutritional measure of boiled leaves with a nutritional measure of the British evening meal to obtain a measure for 'tea'.

Similarly, it is not clear that "digital technology" combining computer-based frog dissection, the impact of colour animations of 3D vectors and using video in vocabulary development is a coherent category. At times, meta-analysts seem less like they are combining apples and oranges, than combining aphorisms and orangutans.

While acknowledging the category coherence as a crucial issue alongside many others (e.g. publication bias and study quality), this paper focuses on the *measure* used to combine and compare those interventions.

Effect Size

The roots of effect size do not lie in measuring effectiveness of interventions, but in evaluating the research design process. Cohen (1962) introduced it to see if psychology studies were designed with good 'power': a good chance of finding a difference between groups (provided there was, in fact, a real difference to be found). That is, it was not developed to quantify the *results* of a study, but to help *plan* higher quality studies.

At its simplest, a study might have a planned intervention treatment, a comparison treatment against which it is contrasted, a sample of participants assigned (preferably randomly) to the treatments and some test taken afterwards. The researcher's interest has traditionally been whether there is a difference in test scores (which might be attributed to the difference in treatments).

[2] Albeit 'sickle cell anemia', 'gender', 'self-assessment' and other areas in Hattie (2009) could not be described as things 'teachers use': the work often confuses 'influences', 'correlations' and 'interventions'.

To explore the notion of statistical power, Cohen introduced standardized effect size (often called Cohen's *d*), commonly taken to be the mean test score of the comparison group, subtracted from the mean score for the intervention group, divided by some measure of spread of scores[3].

The chance of detecting a difference between the groups does not only depend on *d*, but also on the size of sample and the significance level (often 0.05). A researcher might estimate how large this effect size might be in the population from which they draw their sample. Power analysis allows them to adjust the size of their sample to improve the chance that the difference between the groups would be deemed 'statistically significant': large *d* may be detectable with smaller samples, but if *d* is expected to be small, researchers may need a larger sample.

The focus on detecting a difference has been questioned: researchers became concerned that very small effect sizes were detectable with very large samples and so worried that unimportant between-group differences were reported as 'statistically significant' without differentiating them from important differences. It was argued we needed some way of measuring 'practical significance' (Kirk, 1996).

Effect size came to be used to compare individual studies and aggregations of studies, with relative effect sizes standing for relative effectiveness of interventions. This switched the role of effect size from before the experiment (supporting design) to after the experiment (interpreting results). Interventions in studies with larger effect sizes are now promoted as more important or influential than interventions in studies with smaller effect sizes.

As an example, take Gray and Alison's (1971) study of an intervention treatment involving three twenty-minute homework tasks per week for four weeks; a comparison treatment having similar classroom teaching (on fraction arithmetic), but no homework; with grade 6 pupils in a suburban Canadian school being randomly assigned to the two conditions; with a test on the material prepared by the researchers. They found the intervention group averaged 22.29 (with standard deviation 1.82) and the comparison group averaged 21.21 (with standard deviation 2.94). Depending on the particular definition chosen, the standardized mean difference is about 0.45.

While some compare effect sizes between individual studies as a basis for policy recommendations (e.g. Wasik & Slavin, 1993; Gorard, Siddiqui & See 2017), effect size is most commonly encountered by policy makers in meta-analyses and meta-meta-analyses. These averaged effect sizes are promoted as more accurate measures of the effectiveness of interventions which can be compared to identify more or less effective interventions.

When Gray and Allison's study is aggregated with other studies of primary school homework in the EEF toolkit, an average effect size of 0.10 is obtained. This is small compared to other

[3] Different research designs have different definitions of effect size (Lakens, 2013), though analysts often neglect to convert between them. Definitions differ even for independent means designs assumed here, particularly in regard to measuring spread. Some use standard deviation of the comparison group and others the pooled standard deviation of both groups.

areas (e.g. ‘feedback’ or ‘meta-cognition’) so primary school homework interventions are promoted as less effective than interventions from these other areas.

This paper shows this argument is flawed in two fundamental ways. Just as the identification of relative photo-size with relative actual size is a category error, so is the identification of relative effect size with relative effectiveness of interventions. Second, just as using relative *average* photo-size as proxy for relative *average* actual size requires assumptions which will not hold, the paper shows that using relative averaged effect sizes to provide an ordering of the effectiveness of classes of interventions requires assumptions which obviously do not hold.

The misidentification of effect size with the intervention

It is clear from its definition that the identification of effect size from a study with the effectiveness of the intervention is a category error. The definition has three explicit elements: the mean intervention group test score; the mean comparison group test score and the spread of those scores. Each element can be altered (affecting effect size) as a result of design decisions, without altering the intervention. Since the effect size does not depend solely on the intervention it cannot be a straightforward measure of the intervention’s effectiveness.

While one may argue that, all other things being equal, these other elements are factored out by averaging across studies; it should be clear that all other things are *not* equal: tests, samples and comparison activities vary systematically between educational areas.

The paper outlines simple thought experiments showing how each design factor results in studies with different effect sizes for identical interventions, along with illustrations from studies and meta-analyses.

Effect size at the study level

a) The comparison treatment

In the example above, researchers chose to compare three homework tasks to no homework: a reasonable decision in the context of their study. In different circumstances researchers may have chosen a comparison treatment with one homework task, or two; or given the same number of tasks in a different form. Doing so would not change the intervention treatment (or the sample or test), but each study would result in a different (presumably smaller) effect size.

The impact of choice of comparison treatment can sometimes be seen in individual studies. In evaluating the ‘catch up numeracy’ programme (NFER, 2014), the evaluators used two comparison groups (with the same intervention, test and sample). The intervention treatment was a particular numeracy curriculum delivered one-to-one. The first comparison treatment was ‘business as usual’: the normal teaching regime. The second comparison treatment was delivery of ‘time equivalent’ one-to-one numeracy support with content chosen by teaching assistants, provided it was not ‘catch up numeracy’. The comparison to the first group led to an effect size of d=0.21 while the effect size comparing the intervention and second groups appeared much smaller and negative (around d=-0.05).

Examination of meta-analyses also highlights the role of the comparison treatment. Bangert-Drowns et al. (1991) collected studies of frequent testing: those comparing frequent testing to no testing had average effect size 0.56; studies comparing frequent testing to two or more tests per semester had 0.07. Again, the more active comparison was associated with much smaller effect sizes.

Most extremely, meta-analyses sometimes include studies with starkly inactive comparisons: not teaching the topic at all. For example, Steenbergen-Hu and Cooper's (1994) meta-analysis for intelligent tutorial system interventions includes studies where the comparison activity was human tutoring (average d=-0.25) and studies where the comparison was reading computerised material (average d=0.25). Most interestingly a group of studies had 'no treatment' comparisons (average d=0.90) including a study with a sample with no previous economics teaching, using an intervention treatment teaching economics using an intelligent tutoring system, while the comparison group were not taught economics at all; the outcome was measured with an economics test – unsurprisingly the effect size was rather large, around d=1.5 (Shute & Glaser, 1990). While some may argue that researchers should know that effect sizes are relative to comparison treatments, meta-analysist show no qualms in combining very different comparison treatments in a single average (here, d=0.35), with meta-meta-analysts using that summary value to rank order interventions (e.g. Schneider & Preckel, 2017).

The choice of comparison activity is neither arbitrary nor random; researchers select it to meet their intentions, within restrictions laid down by convention and practicality. Where possible, having a less active comparison allows increased power (the chance of detecting a group difference) and therefore effect size without altering the intervention.

Thus it is a category error to read relative effect sizes as relative effectiveness of interventions, let alone to assume they signal educational importance, relevance or influence.

b) Sample

The choice of sample can impact on effect size for a study in two interacting ways, independent of the intervention, control treatment or test. The first follows from the definition of effect size: the more homogenous the sample on the outcome measure, the higher the effect size. The second relates to the mechanism through which the intervention treatment works.

The first issue has long been known (e.g. Fitzgibbon, 1984), albeit meta-analysts rarely address it[4]. Fixing intervention treatment, comparison treatment and test, the researcher can choose a sample with a wider or narrower range of ability. They may do so explicitly to increase power (and hence effect size) as some recommend (e.g. Lipsey, 1990), or implicitly

[4] Despite its prevalence throughout the EEF Toolkit, only one meta-analysis appears to mention range restriction: this is to acknowledge that they *introduced* range restriction problems by trimming outliers (Kluger & DiNisi, 1996).

because the focus is on a particular ability range or because a restricted ability range is convenient. Bobko, Roth and Bobko (2001) note that adjusting for this requires knowing a great deal about the samples or making assumptions which may be little more than guesswork. Few educational meta-analyses appear to attempt adjustment: most just ignore the issue.

The second issue with the sample is that an intervention may be expected to work differently with different people – in particular, effectiveness may vary with pre-existing ability. A computer based activity aimed at improving test taking techniques may be effective with pupils struggling with such techniques; so, a sample consisting of these pupils may show a larger (raw) difference in mean scores on a suitable test. The same activity may be ineffective with confident test takers; so, a sample of those may show little difference in mean scores (Martindale, Pearson, Curda and Pilcher, 2005).

The closer the researcher can match choice of sample to the people for whom the intervention's underlying mechanism is effective, the larger the mean difference. The more general the sample, including people for whom the intervention is ineffective, the smaller the mean difference (and also the wider the spread).

These two issues (differential effectiveness and range restriction) are easy to conflate. In the first, a particular intervention may disproportionately affect part of the population and so a researcher may target their study towards them. In the second, the researcher may increase experimental power (explicitly or implicitly) by restricting the range of the sample: in a population with similar ability, a small raw difference in achievement will stand out clearly simply because of the reduced spread.

These issues also interact: if one study conducted with lower achieving groups has a higher effect size than an otherwise identical study with wider ability groups, without adjusting for range restriction it is difficult to tell whether the intervention is better targeted at lower achieving pupils or whether restricted range has inflated effect size.

However, the key issue is that studies with the same intervention (and same comparison treatment and outcome measure) can have very different effect sizes with different samples. Again, this demonstrates that identifying relative effect sizes with relative effectiveness of interventions is a category error.

c) Tests

Perhaps the most obvious research design choice is the selection of outcome measure – most frequently a test which the participants take at the end of the treatments.

A simple thought experiment shows the sensitivity of effect size to the test, holding intervention, comparison treatment and sample constant. If we split a sample of pupils randomly in two and teach one group an isolated fact which no-one in the sample knew (e.g. 'oktatás' is the Hungarian word for 'education', assuming non-Hungarian speakers), then a test requiring the reproduction of that fact, would lead to a potentially infinite effect

size[5]. Minor changes to the test results in very different effect sizes: for example, translating 10 otherwise unknown Hungarian words (including 'oktatas') in a four-option multiple choice test would give an expected effect size around 0.6; with three options it would be around 0.4; with ten options it would be around 0.9 [6]. Removing the 'oktatás' question would result in an expected effect size of 0. Despite wildly different effect sizes, the intervention is the same in each case (as are the sample and comparison treatment).

In Gray and Allison's (1971) homework study with fraction arithmetic, the researchers designed a test of fraction material themselves, but could have made other design decisions. They could have tested a wider selection of mathematical topics; they could have selected from a bank of appropriate standardized tests (which would vary in the number of fraction arithmetic questions); they could (unadvisedly) have used a reading test[7].

Again, it is possible to see the impact of test selection on effect size in individual studies. In their evaluation of the 'response to intervention' literacy programme, Gorard, Siddiqui and See (2014) report d= +0.19 for the 'New Group Reading Test' and -0.09 for the 'Progress in English' test. In their evaluation of the Nuffield Early Language intervention, Sibieta, Kotecha and Skipp (2016) used a variety of different outcome measures, including a grammar test (d=0.29), an expressive vocabulary test (0.25) a letter sound knowledge test (0.12) and a word reading test (0.01). The same intervention, sample and comparison treatment results in very different effect sizes depending on the test.

Cheung and Slavin (2016) looked at 645 studies across twelve meta-analyses across a wide range of topics. They separated studies using tests designed by the researchers from studies with 'standardised tests' (tests designed by others to cover a particular area of the curriculum, often having been norm referenced, designed for particular age ranges etc.) Across the studies, effect sizes for standardized tests were around half those of researcher designed tests.

Few have explored reasons for this difference. Ruiz-Primo, Shavelson, Hamilton and Klein (2002) propose an explanation around curricular distance:

> At the *immediate* level, artifacts from the enactment of the curriculum provide achievement information. At the *close* level, assessments should be curriculum sensitive; they are close to the content and activities of the curriculum. At a *proximal* level, assessments should be designed considering the knowledge and skills relevant to the curriculum, but content (e.g. topics) can be different from the one studied in the unit. At a *distal* level, assessment may be based on state or national standards in

[5] If translating 'oktatás' was one of ten, otherwise unknown, Hungarian words on the test, nearly everyone in the intervention group would score 1 out of 10 (with little variance) and nearly everyone in the comparison group would score 0 (with little variance). A raw mean difference of 1, divided by a zero (or a very small) standard deviation. See Simpson (2018)

[6] For example, with four options, the intervention group would average 3.25 (one correct, nine guessed); the comparison group would average 2.5 (all ten guessed) with standard deviation around 1.4 (the intervention group's variation being slightly smaller – they guess fewer answers).

[7] But test-intervention match is not always obvious: note the use of mathematics tests to evaluate a philosophy intervention (Gorard, Siddiqui and See, 2015a) and a chess intervention (Jerrim, Macmillan, Micklewright, Sawtell & Wiggins, 2017)

> a particular domain. At a *remote* level, general measures of achievement should be used (p.371)

They found effect sizes between two and five times larger for close assessments than distal ones.

However, the characterization Ruiz-Primo et al. make is of curricular distance, rather than the mechanism impacted by the intervention: one might think of a fractions test as immediate and a public mathematics examination as distal, but an intervention which makes a very great difference to (say) mathematical reasoning ability may show up more clearly on a general mathematics examination than on a specific fractions test. So, rather than curriculum distance, it may make sense to talk of how closely a test matches the mechanism which the intervention impacts.

A test focused on what it is that the intervention does to the pupils (compared to the comparison) will lead to a larger effect size. However, even in an intervention with a very narrow outcome (such as improving procedural fraction addition), researchers may be constrained to use a standardized test. But they can still *select* to maximise power (and increase effect size) by choosing, say, a numeracy test rather than a more general mathematics test.

Again, there are other subtleties with test choice. Depending on the consistency of the test items, increasing the size of a test may increase effect size (though equally, adding irrelevant items will tend to decrease effect size by adding noise[8]). This is bound up with design decisions and researchers' freedoms and constraints: a long test may be impractical, piloting a test may have led to changes which increase its reliability (and therefore effect size) etc.

Selecting a test is a design decision which, for fixed intervention, sample and comparison treatment, can result in very different effect sizes. So, again, taking relative effect size as a measure of relative effectiveness of interventions is a category error.

Comparing effect sizes: individual studies

Just as the only way that Simon can draw the valid conclusion that Adam's elephant is really smaller than Belinda's on the basis of their photo-size is if all other components that impact on photo-size are equal, the only way of validly comparing the effectiveness of two interventions on the basis of effect size is to be sure that the test, sample and comparison activity are the same.

It is rare for researchers to directly compare individual studies by effect size to draw out policy advice. However, Gorard, Siddiqui and See (2017) compared seven literacy interventions[9]. While some had tests in common, some had populations which might be considered similar and some had similar comparison treatments, no pair of studies shared all of the three components needed to allow a valid comparison. For example, the

[8] Continuing the Hungarian example, a four-option multiple choice test with five unfamiliar words (including 'oktatás') might have an expected effect size of around 0.8, with twenty words we might expect 0.4.

[9] Though one, 'Philosophy for Children', was not obviously targeted at literacy.

Philosophy for Children intervention (Gorard, Siddiqui & See, 2015a) used a wide range of pupils (in-tact, mixed ability primary year 3-6 classrooms) and measured outcomes using gains between the national Key Stage 1 and 2 tests; Accelerated Reader (Gorard, Siddiqui & See, 2015b) used a narrower range (year 7 pupils with lower attainment) with post-test scores from the standardized *New Group Reading Test A*[10]. Similarly, Wasik & Slavin (1993) compare studies of five reading recovery programmes partly on the basis of effect size, no two of which are convincingly comparable on test, sample and comparison treatments.

Advising schools to select between these interventions on the basis of effect size is a mistake. At the individual study level, effect size can act as proxy for the effectiveness of the intervention only if all other components of the effect size are the same. While it may be argued that one might compare effectiveness of interventions on the basis of effect size provided these components are *similar* (rather than the more stringent requirement of being the same), the argument above shows that seemingly minor changes to the range of the sample, the comparison activity and the outcome measure can have large impact on effect size. So, those who argue we can accept comparing, combining or rank ordering interventions on the basis of effect size based on *similar* tests, comparison treatments and samples would need conversion factors relating some clearly defined metric for 'similarity' on each of these dimensions to impact on effect size. While, as noted above, such a conversion factor exists for range restriction, there is little evidence that it is used and it requires highly detailed knowledge about sample distributions (or reliance on still more assumptions). Further, it is not clear that one can, even in principle develop such conversions for the other factors.

Comparing Effect Sizes: Meta-analysis

While comparing interventions on effect size for individual studies is rare, comparing aggregated effect sizes is common and is the basis of the meta-meta-analyses used to direct 'evidence based education' policy. This is done in two ways: an individual study's effect size is compared against a standard, or the average effect size over one collection of studies is compared to the average effect size of another collection.

In the first case, the comparison of a single study to a group of studies is often implicit. Researchers sometimes report their study's effect size as 'small', 'medium' or 'large' depending on cut-off points derived from previously aggregated effect sizes (Levin, 1997). For example, Hattie (2009) averages all effect sizes in the collection of meta-analyses to calculate a 'hinge' (d=0.4), above which, it is argued, is a 'zone of desired effects'. This is used as a benchmark for the relative value of an individual study. Given the sensitivity of effect size to sample, test and comparison treatment, this is a mistake.

More often effect sizes are averaged in meta-analyses and then compared (e.g. Schneider and Preckel, 2017) or effect sizes from meta-analyses are further aggregated in themes and then ranked (e.g. Hattie, 2009; Higgins and Katsipataki, 2016). For example, the EEF toolkit combines seven meta-analyses under the heading of 'meta-cognition' and reports an effect size of 0.62 and combines eight meta-analyses under 'behaviour interventions' and reports

[10] There are further issues in comparing effect sizes between study designs which use 'gain scores' and which use end of treatment scores only (see Baguley, 2009)

an effect size of 0.25. This is promoted as evidence that schools are likely to find more potential for improving achievement with meta-cognition interventions than behavioural ones.

This argument is invalid: It may be possible to identify circumstances when relative aggregated effect sizes might be proxy for relative effectiveness of interventions, but to assume these hold without checking seems neglectful, and checking shows they do not hold.

Recall that the average photo-size of one set of creatures being larger than the average photo-size of another warrants a valid conclusion that the average real size of one group of creatures is larger only if an 'all other things being equal' (in distribution) assumption holds. Only if the *distribution* of lenses and distances for elephant photographers and princess photographers were the same – either systematically or randomly – in Catherine's and Douglas's collections, might the argument work.

This same requirements apply to meta-analysis. To make a valid comparison between averaged effect sizes stand as a valid comparison of effectiveness of the interventions, the other elements impacting on effect size have to be distributed equally across the meta-analyses (or meta-meta-analyses).

It is important to note that 'haphazard' is not enough, 'not systematically different' is necessary but not sufficient and even 'at random' is not sufficient. It is not enough to argue that a collection of studies has varying tests, samples and comparison treatments; to note that no deliberate attempt has been made to restrict on these components; to undertake moderator analyses of these components; nor to contend that the set of studies passes homogeneity tests (which are anyway tests of distributions of effect sizes, not of components of effect sizes).

It has to be the same distribution across all other components: either the same by design or by drawing at random, independently from the same distribution. If around three-quarters of the studies in one area use low achieving pupils, around three-quarters of the studies in the other meta-analysis need to be similarly range restricted. If 10% of the studies in one area use 'no teaching at all' as a comparison treatment, then around 10% of studies in other areas need to use 'no teaching at all'.

Berk and Freedman (2001) contend that 'statistical assumptions are empirical commitments': the assumptions on which an argument relies commit the arguer to a claim about the nature of the world which resulted in the data. For example, to validly argue that relative average photo-size is a proxy for relative average actual size, we commit to how the world of photography must be – that lenses and distances to subjects are equally distributed across different areas – and it is possible to at least sense check, if not empirically confirm, whether the world is like that. Even at the sense check level, the assumption appears absurd: wildlife photographers make design decisions systematically different from those of portraitists which are different again from micro-biologists.

It should be possible to sense check education meta-analyses and meta-meta-analyses for similar empirical commitments, even though the analysts neglect to do this. In this case, the assumptions require (among other things[11]) that the joint distribution of tests, comparison treatments and sample ranges is the same in the different collection of studies. At the sense check level, this is clearly unlikely and an exploration of some collections of studies below shows it does not hold: different educational areas have different conventions and freedoms within which researchers make design choices and this leads to very different distributions.

For example, researchers of feedback disproportionately use 'no feedback' comparison conditions. Meta analysts of phonics disproportionately collect studies which use literacy tests. The EEF toolkit designers took the decision to only look at meta analyses of setting and streaming with low achievers while averaging more widely in other areas. Comparing the effectiveness of interventions across these areas on the basis of relative average effect size is not a valid form of argumentation.

Schneider & Preckel (2017) rank (among many other meta-analyses) averaged effect sizes from a set of studies for intelligent learning systems (d=0.35, Steenbergen-Hu & Cooper, 2014) with that for testing aids (d=0.34, Larwin, Gorman & Larwin, 2013). To argue, then, that these types of intervention are about equally effective because the effect sizes are close requires that the other elements which impact on effect size (test, comparison treatment and sample) are distributed in the same way in each set.

They are not: take one component, the distribution of comparison treatments. Steenburgen-Hu & Cooper (2014) include studies with a wide variety of comparison activity (as noted above, from comparing to human tutoring, through to not teaching the topic at all), while Larwin, Gorman & Larwin (2013) include only studies where the comparison treatment involves no testing aids at all. These cannot be described as studies with the same distribution of comparison treatments. It would seem highly unlikely that any two meta-analyses would have the same distribution of comparison treatments, let alone that across all pairs of educational areas in a meta-meta-analysis, the distribution of comparison treatments would be equal.

Distributions of tests are also unequal. Means, Toyama, Murphy and Baki (2013) report that 26% of the tests in their meta-analysis of online and blended learning are tests of declarative knowledge; Höffler & Leutner (2007) report that 7% of their studies of use of instructional animation used tests of declarative knowledge. Springer, Stanne & Donovan's (1999) meta-analysis of small group teaching reported 60% of studies used mathematics tests, while Luiten, Ames & Ackerson's (1980) meta-analysis of advance organisers reports 22% of studies used mathematics tests. Yet these are ranked against each other in Schneider and Preckel (2017) purportedly to allow "for comparisons of the relative importance of a wide range of variables for explaining academic achievement in higher education, which can inform researchers, teachers, and policymakers" (p. 566)

[11] for example, that effects of interventions across different studies are constant (for fixed effect meta-analyses) or distributed normally around some mean (for random effect meta-analyses). These too are empirical commitments which go unchecked and are, prima facie, unlikely.

The same issue appears in the meta-analyses which underpin the EEF toolkit: e.g. Kulik & Kulik (1982) report that 82% of their studies on ability grouping use standardized tests; Paschal, Weinstein and Walberg (1984) report that 48% of their studies on homework use standardized tests. The assumption about the distribution of tests being the same across collections is clearly not met.

In terms of range restriction, of Lou, Abrami & d'Apollonia's (2001) set of studies on group and individual learning with technology, 13% use low ability participants; Luiten et al. (1980) report 26% of their studies involve low ability participants. At the meta-meta-analysis level, the EEF toolkit areas are combined sometimes deliberately with different ranges of participants: for 'setting and streaming', the average effect size comes only from meta-analyses where 100% of studies use low attainers[12]; of the studies in the meta-analyses which make up the 'summer school' area, around 70% had attainment as a criteria for the sample (most being low attainment); other areas, such as 'meta-cognition', predominantly use broader samples. The assumption that the distribution of the range of samples is equal across collections of studies is clearly not met.

It is not just that we can find pairs of meta-analyses which obviously vary in distributions of comparison, test and range features, it is also clear that it we should not expect to get the same distributions in *any* pair of meta-analyses, since the nature of the area under investigation often restricts the distribution of these factors differentially.

For example, it would be unethical to evaluate a behaviour intervention where the comparison activity involves no behaviour intervention. However, many laboratory based studies of feedback interventions compare to 'no feedback' treatments (Simpson, 2017). Ranking 'feedback' above 'behaviour interventions' on the basis that relative averaged effect sizes act as proxy for relative effectiveness of interventions is clearly invalid. Interventions which target pedagogical approaches to an academic topic directly (such as feedback, meta-cognition and homework) are more likely to use researcher-designed tests focused on the subject content; while interventions less directly aimed at a pedagogical approach for a particular academic topic (such as behaviour interventions, school uniforms or aspiration interventions) are more likely to use general, standardised tests. We cannot therefore conclude that direct forms of instruction are more effective than indirect approaches simply because average effect sizes are larger.

Not only do the assumptions required for using relative averaged effect sizes as proxy for relative effectiveness of interventions fail a sense check, the empirical commitments they entail are unrealistic.

What are effect sizes?

The argument above shows that relative effect sizes are not measures of relative effectiveness of interventions at individual study or meta (or meta-meta) analysis levels. Effect sizes refer to the study process as a whole (just as photo-size of a creature refers to

[12] Interestingly, many of the studies (including one whole meta-analysis) are about within-class grouping, so not obviously 'setting and streaming' at all.

the photography process as a whole and is not an appropriate measure of the actual size of the creature).

They may be better thought of as a measure of the *clarity* of the study (Simpson, 2017) – they indicate how clear the difference between the treatments was on the sample, as measured on the test. With different measures, the difference may be clearer. More homogenous samples may make the difference stand out more. More active comparison treatments may make the difference less detectable.

Effect size was developed for this notion of detectability, through power analysis. Indeed, for a fixed sample size and significance level, effect size and the chance of detecting a difference (if one exists) are essentially the same. The discussion above has important implications for power analyses. Lipsey (1990) and Levin (1997) suggest careful research design decisions are a valid alternative to increasing sample size for increasing power (and therefore increasing effect size).

It is often recommended that researchers look to studies of similar interventions to find an estimate of the effect size to determine power and sample size (e.g. Murphy, Myors & Wolach, 2014). But the argument above highlights that the intervention is only one component. It is possible that, since the researchers intend to conduct a study in the same area as others, they may use similar tests, comparisons and samples, but one should check this is the case and base power analyses on studies which are similar not just on interventions, but also on these other components.

This problem can be seen in large scale (and expensive) evaluation programmes. In commissioning studies of interventions, the EEF have pilot, effectiveness and efficiency stages: each phase likely to be commissioned (and receive extra funding) partly on the basis of previous effect sizes. But later phases are more likely to require standardised tests and, perhaps, wider samples. First, by equating promise of the *intervention* with the effect size of the *study*, they make the category error discussed above; second, by ignoring the impact that changes to test, sample and comparison treatment has on effect size, studies often end up underpowered, with few detecting effect sizes as large as those assumed in power analyses. Across the 32 effectiveness reports published by the EEF at the time of writing, not one found an effect size as large as the minimum detectable effect size (MDES) used at the planning or protocol stage – on average the reported effect size was under two-fifths of the MDES.

Reporting effect size is important for future researchers. By selecting studies as near to their planned design (not just the intervention) as possible, they can conduct valid power analysis. Effect sizes are not, however, suitable for policy making.

The Role of Context, Resources and Intentions

Photographers make decisions on the basis of context, resources and intentions. They may be able to get close to some subjects, but not others; they may not be able to afford a telephoto lens; they may be aiming to submit the photo for publication (where a wildlife editor will want considerable context and a portrait editor may not). A picture of a newly

isolated microbe may be taken so the image dominates the photograph; while almost any picture of a rare white rhino is worth taking even if there is no time to get close.

Researchers too work in societies where contexts, resources and intentions matter. For many studies, the aim may be to show an intervention 'worked' or not – in the sense of finding a statistically significant difference between intervention and comparison groups (provided one exists). Improving power by increasing sample size is one possibility, but can be resource intensive and may not always be available, so selecting more homogenous samples or more focussed tests may be sensible design decisions. Funders may add restrictions on acceptable design and the culture of the researchers' discipline may reduce their freedoms.

So researchers with an eye on a journal publication may choose tests they have designed themselves, to increase power; while those asking for funding from the EEF will tend to propose standardised tests. Laboratory based research may allow for 'no treatment' comparisons; field research may not. Researchers interested in social justice may be more likely to conduct studies with low achievers, while researchers interested in scaling a promising intervention to evaluate broader applicability will likely use a wide sample. Context, resources and intentions all impact on design and design impacts on effect size, separately from the intervention. Moreover those contexts, resources and intentions vary systematically between educational areas; one cannot just assume this systematic variation away because doing so is convenient.

Conclusions

Understanding that effect size is a property of the whole study, not just the intervention may help us avoid reifying concepts such as the 'effectiveness of the intervention' which may have little straightforward meaning.

For example, the meaning of 'the effect of feedback' is far from clear. 'Evidence based education' proponents often claim inspiration from pharmacology, yet we would dismiss arguments about 'the effect of aspirin' that drew on studies of aspirin against a placebo, aspirin against paracetamol, aspirin against an antacid and aspirin against warfarin on measures as diverse as blood pressure, wound healing, headache pain and heart attack survival. Yet we are asked to believe in an 'effect for instructional technology' based on comparing interactive tutorials to human tutoring, to reading text, to 'traditional instruction' and to not being taught the topic at all, on outcomes as diverse as understanding the central limit theorem, writing simple computer programs, completing spatial transformations and filling out account books[13].

The assumptions for comparing meta-analyses and meta-meta-analyses require that design factors are distributed equally across different areas of education, yet clearly researchers in different areas make systematically different decisions. In Berk's (2011) deconstruction of meta-analysis he notes that even small deviations from assumptions invalidate the

[13] It could be argued that these are all measures of achievement and standardising makes them comparable. The argument applied to aspirin would be that the outcomes are all measures of health, yet standardising does not somehow make them sensibly comparable.

inferential argument, leaving "statistical malpractice disguised as statistical razzle-dazzle" (p.199)

Freedman (2009) lists specious defences used by those who would razzle-dazzle policy makers. Among these are "you can't prove the assumptions are wrong"[14], "If we don't do it someone else will" and "the decision maker has to be better off with us than without us". It is possible to read many of these defences in the work of prominent authors in the 'evidence based education' movement (e.g. Higgins and Katsipataki, 2016; Schneider & Preckel, 2017; Hattie, 2017).

Amongst Freedman's defences, the most telling may be "where's the harm?". Harm comes in presenting policy conclusions as 'evidence based' when the evidential chain of argument is broken. The 'evidence based education' movement promotes a doctrine based on providing "solid evidence of effectiveness" leading to "genuine, generational progress instead of the usual pendulum swings of opinion and fashion" (Slavin, 2002).

As Henry Fielding (1749) said of the religious doctrine relating virtue and happiness, this is "a very wholesome and comfortable doctrine, and to which we have but one objection, namely, that it is not true."

Acknowledgements
I am grateful to Doug Newton and Terry Wrigley for wise advice with earlier drafts of this paper.

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported?. *British Journal of Psychology*, *100*(3), 603-617.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research, 85,* 89–99.

Berk, R. (2011). Evidence-based versus junk-based evaluation research: Some lessons from 35 years of the evaluation review. *Evaluation Review*, *35*(3), 191-203.

Berk, R. A., & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. *Law, punishment, and social control: Essays in honor of Sheldon Messinger*, *2*, 235-254.

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of d for range restriction and unreliability. *Organizational Research Methods*, *4*(1), 46-61.

Campbell, S. & Harper, G. (2012). *Quality in policy impact evaluation: understanding the effects of policy from other influences (supplementary Magenta Book guidance)*. London: HM Treasury.

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

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

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

[14] It is not, of course, for the reader to disprove the assumptions, but for the analyst to justify them. In this case, however, we can indeed prove the assumptions are wrong.

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

Fielding, H. (1749) *The History of Tom Jones, A Foundling*, London: Andrew Millar.

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

Freedman, D. (2009) *Statistical Models: Theory and Practice*, Cambridge: Cambridge University Press.

Gorard, S., Siddiqui, N., & See, B. H. (2014). *Response to Intervention: Evaluation Report and Executive Summary*, Education Endowment Foundation.

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

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

Gorard, S., Siddiqui, N., & See, B. H. (2017). What works and what fails? Evidence from seven popular literacy ‘catch-up’schemes for the transition to secondary school in England. *Research Papers in Education*, *32*(5), 626-648.

Gray, R. F. & Allison, D.E. (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.

Hattie, J. (2017). Educators are not uncritical believers of a cult figure. *School Leadership & Management*, *37*(4), 427-430.

Higgins, S., & Katsipataki, M. (2016). Communicating comparative findings from meta-analysis in educational research: some examples and suggestions. *International Journal of Research & Method in Education*, *39*(3), 237-254.

Höffler, T. N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction, 17,* 722–738.

Jerrim, J., Macmillan, L., Micklewright, J., Sawtell, M., & Wiggins, M. (2017). Does teaching children how to play cognitively demanding games improve their educational attainment? Evidence from a Randomised Controlled Trial of chess instruction in England. *Journal of Human Resources*, online first.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746-759.

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.

Kulik, C., & Kulik, J. (1982). Effects of Ability Grouping on Secondary School Students: A Meta-analysis of Evaluation Findings. *American Educational Research Journal*, *19*(3), 415–428.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*(863), 1-12.

Larwin, K. H., Gorman, J., & Larwin, D. A. (2013). Assessing the impact of testing aids on post-secondary student performance: A meta-analytic investigation. *Educational Psychology Review, 25,* 429–443.

Levin, J. R. (1997). Overcoming feelings of powerlessness in “aging” researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging*, *12*(1), 84–106.

Lipsey, M. W. (1990). Design sensitivity: Statistical power for experimental research (Vol. 19). Sage.

Lou, Y., Abrami, P. C., & d'Apollonia, S. (2001). Small group and individual learning with technology: A meta-analysis. *Review of educational research*, *71*(3), 449-521.

Luiten, J., Ames, W., & Ackerson, G. (1980). A meta-analysis of the effects of advance organizers on learning and retention. *American Educational Research Journal, 17,* 211–218.

Means, B., Toyama, Y., Murphy, R., & Baki, M. (2013). The effectiveness of online and blended learning: A meta-analysis of the empirical literature. *Teachers College Record, 15,* 1–47.

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Routledge.

NAO (2015) *Funding for disadvantaged pupils. London:* National Audit Office

NFER (National Foundation for Educational Research). (2014). *Catch-Up Numeracy: Evaluation Report and Executive Summary.* Education Endowment Foundation.

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

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369-393.

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological bulletin*, *143*(6), 565.

Shute, V. J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments*, *1*(1), 51-77.

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

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy 32*(4): 450-466.

Simpson, A. (2018). Separating arguments from conclusions: The mistaken role of effect size in educational policy research. *Educational Research and Evaluation*, online first.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational researcher*, *31*(7), 15-21.

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research, 69,* 21–51.

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology, 106,* 331–347.

Wasik, B. A., & Slavin, R. E. (1993). Preventing early reading failure with one-to-one tutoring: a review of five programs. *Reading Research Quarterly*, *28*(2), 179–200.