On the misinterpretation of effect size

Adrian Simpson
School of Education, Durham University, UK

adrian.simpson@durham.ac.uk

ORCID: 0000-0002-3796-5506

I was partly heartened to read the editorial on effect sizes (Bakker et al., 2019). That editors of some of the major journals in mathematics education have brought together emerging concerns about the use and interpretation of effect size in education research for a wider audience is clearly of value. However, the editorial includes some misconceptions about the measure, cites some flawed arguments and perpetuates an interpretation of effect size (albeit in a more nuanced form) which is not helpful for the field.

The thrust of the editorial is to suggest alternative guidelines for using effect size, particularly for programme evaluation research, with a view to being able to answer the question "is the effect large enough to be useful?" The suggestions involve either using alternative benchmarks or, ideally, contexualising an effect size against comparable studies. The editorial outlines twelve issues for consideration in setting those benchmarks or in interpreting the effect size in context.

Underpinning the editorial appears to be a particular interpretation of effect size as a measure of the intervention: i.e. larger effect sizes indicate interventions which are more useful, effective or important. I have argued elsewhere that this interpretation is not just a misconception but is potentially damaging in leading teachers and policy makers to base decisions at least partly on a metric that does not measure relative effectiveness (Simpson, 2017).

I outline three key points here. First, the way in which many contextual benchmarks are constructed is flawed; second, in outlining twelve factors, the editorial omits at least one important issue and includes a factor which does not actually impact on effect size; third, interpreting effect size as a measure of the effectiveness of an intervention is normally a mistake, even when contextualized, and that interpreting it as a measure of the clarity of a study helps explain many of the twelve factors in the editorial. I conclude that we should stop adjusting, re-benchmarking or contextualising in order to force effect size to be something it is not. We should use it for comparative purposes only in rare and precisely defined situations, restricting its use to being a research design tool for replications.

1. Benchmarking

Cohen's (1962) introduction of effect size came in response to a need to determine if studies were well powered – giving a sufficient opportunity to detect an effect (if one was present). He set benchmarks in the form of exemplars of small, medium and large effect size values. Here (as in the editorial) effect size is taken to be standardised mean difference – for a typical randomised control trial, that is the difference between the mean score of the intervention group and the mean score of the comparison group on the chosen outcome measure, divided by some measure of the spread of those scores (such as the pooled standard deviation). Some of the issues covered here apply to other forms of effect size (such as raw mean difference or correlation coefficient), but others do not.

Cohen's exemplar benchmarks (originally 0.25 for 'small', 0.5 for 'medium' and 1.0 for 'large') appear to have been based on how 'noticeable' he felt a difference might be: using examples of people with different IQs he set a large difference as one which "would be so obvious as to virtually render a statistical test superfluous" (p. 147).

The editorial notes that Cohen's values seem inappropriate when one recognises that in recent years effect sizes, when averaged across many experiments in education, seem much smaller.

For example, Cheung and Slavin (2016) found an average effect size of 0.16 in their sample, and Lortie Forgues and Inglis (2019) found an average effect size of 0.06 in theirs.

There have been many more or less explicit attempts to set benchmarks for effect size since Cohen. For example, the Institute for Education Studies in the US (IES) sets a benchmark for an effect size of 0.25 for a finding to be designated 'substantively important' (IES, 2017) and Lortie-Forgues and Inglis (2019) use an effect size of 0.10 as the basis of their choice of definition of 'informative'. The editorial particularly highlights a recent, as yet unpublished paper (Kraft, 2019) which argues for new benchmarks for studies with a given context: pre-K through grade 12, evaluating effects on student achievement with broad, standardised outcome measures[1]. Kraft sets alternative values for boundaries between small and medium, and between medium and large effects at 0.05 and 0.20 respectively, again using the basis of the average of effect sizes across a large number of studies from his chosen context.

However, this idea of setting benchmarks on the basis of the average across a set of studies has two issues: one obvious and one more subtle.

First, they are based on averaging *signed* effect sizes. Given that typical trials are symmetrical this makes little sense: a trial comparing treatment A to treatment B (on a particular sample and outcome measure) with effect size x is also a trial of treatment B to treatment A (on that sample and measure) with effect size $-x$. If we declare x to be, say, a large effect, we must declare $-x$ to be a large effect as well (see Bergeron & Rivard, 2017, making a similar point about the way Hattie, 2009, calculates his 'hinge'). Of course, we might expect a positive bias, particularly in evaluation trials: researchers tend to propose intervention treatments they have reason to believe are better than comparison treatments. But a large negative effect should still be considered a large effect for the purposes of creating benchmarks.

Second, as the editorial points out, a key question which needs to be answered is whether an effect is real or could be attributed to chance. In setting benchmarks for effect sizes, the aim presumably is to set them for effects which researchers might classify as 'real'.

Imagine a team of geographers wanting to categorise geological features as 'hills' or 'mountains' by averaging heights. If they averaged across every feature in the terrain which deviated from level ground, no matter how small, they might include every rock or even every grain of sand. Done naïvely, this could reduce the average to the point that anything higher than a couple of grains of sand might get classified as a mountain. There needs to be some cut off point which identifies the class of things to be categorised and excludes what is little more than noise in the terrain.

The analogous argument applies to setting benchmarks for effect sizes from averaging across studies. In Cohen's (1988) discussion of his benchmarks as a "subjective average" of

[1] The fact that Kraft is somewhat inconsistent in his contextualising indicates how hard it is to pin down context. On the same page of his article (p. 20) he argues his benchmarks are for "effect sizes from causal studies of pre-K–12 education interventions evaluating effects on student achievement" and, later, his benchmarks are for "causal research that evaluates the effect of education interventions on standardized student achievement". Elsewhere he notes that still further contexualisation could be focussed on particular grade within pre-K-12 and on subject matter! I should also note that, given the unpublished nature of Kraft's paper, we need to be careful in referring to its arguments: there are multiple versions in circulation which make slightly different claims.

behavioural science effect sizes, he explicitly excludes those which are "so small that seeking them amidst the inevitable operation of measurement and experimental bias and lack of fidelity is a bootless task" (p. 13).

While there may be situations where it is appropriate to average all signed effect sizes, it seems inappropriate for setting benchmarks to categorise levels of real effects. Yet Kraft (2019) averages positive and negative effect sizes and sets no cut-off point – including effect sizes for studies which the researchers might conclude detected no effect and may represent little more than noise around zero.

To get a sense of the impact of these issues, of the 271 effect sizes extracted from the set of studies analysed by Lortie-Forgues and Inglis (2019), the (unweighted) mean of the signed effect sizes is 0.05, the mean absolute effect size is 0.09 and, of the 55 instances where the researcher might have concluded that the study detected an effect (by, for example, being statistically significant), the mean absolute effect size is 0.20.

2. Factors which impact on effect size

While the editorial promoted Kraft's benchmarks as "an improvement for policy decisions" (which I dispute below), it suggests that ideally there is a need for still further contextualisation. A substantial portion of the editorial is spent outlining twelve factors which influence effect size. However, this includes a factor which does not influence effect size and omits at least one other important factor.

2.1 Sample size – effect size association

Sample size – the ninth factor in the list – does not influence effect size, even if it has an association. Despite the apparent negative relationship between sample size and effect size, all other things being equal, increasing sample size does not reduce effect size. Identical experiments conducted with larger samples (taken at random from the same population) should be expected to have the same effect size. This is simple to see: recalculating effect size from a random subset of the participants in an experiment should have the same expected effect size (the means and standard deviations would be expected to be the same) albeit within a larger confidence interval.

There are other plausible ways to account for the observation that larger samples are often associated with smaller effect sizes.

First, where possible researchers designing studies in the expectation of small effect sizes might sensibly increase power by choosing larger samples. Assuming that there is a positive relationship between researchers' expectations of effect size at the design stage and the effect size determined from the experiment, we would expect larger sample sizes to be associated with smaller effect sizes: that is, rather than larger samples being a cause of smaller effect sizes, smaller effect sizes can be a cause of larger samples!

Second, larger experiments are likely to have differences in design other than the size of the sample. For example, larger experiments are more likely to be evaluations of programmes at scale and thus tend to use standardized measures and active, business-as-usual type comparison activities, while smaller experiments are more likely to use researcher designed

measures and perhaps more passive comparison treatments (whose impacts on effect size are discussed below).

Third, researchers who are constrained by circumstances to use smaller samples may choose to alter the design to enhance power. As Cohen (1973) put it, researchers should "strive toward developing the insights which lead to research procedures and instruments which make effects measurably large enough to be detected by experiments of reasonable size" (pp. 228–229). This might involve using more sensitive measures, a more homogenous sample or a larger difference in treatments (through a higher dose of the intervention treatment or a more passive comparison treatment). Each of these increases power by increasing effect size rather than sample size and gives another mechanism by which sample size and effect size can be negatively related without the need to posit some unknown mechanism by which increasing sample size independently causes smaller effect size.

Note that these mechanisms undermine the use of funnel plots and related methods which many meta-analysts use to evaluate publication bias: since we would expect experiments with larger samples to have smaller effect sizes, we would expect funnel plots to be asymmetrical even without missing studies (Terrin, Schmidt, Lau, & Olkin, 2003).

2.2 Measure Design

While the editorial notes that the alignment of the measure to the difference in treatments influences effect size, this is not the only aspect of measure design which impacts on it. As Cheung and Slavin (2016) note, averaged across many studies, effect sizes from 'standardised' measures are around twice those from 'researcher designed' measures. Different ways of categorising measures as well as different sets of studies have unveiled similar findings, albeit with different ratios: Li and Ma (2010) found that effect sizes from studies of computer technology use on mathematics with standardised outcome measures were around 50% higher than those with non-standardised measures. These and many other associated issues might be artefacts of the more general issue of the distance between a measure and the difference in treatments.

If one imagines an intervention which improves learners' fraction addition performance compared to the comparison treatment (and has no other impact), then a measure of fraction addition performance will lead to a larger effect size than that from a measure of more general fraction arithmetic performance, which would be larger than that from a measure of mathematics performance, which would be larger than that from a measure of general educational performance (mixing multiple subject areas).

Standardised tests are more likely to include questions which are unrelated to the difference in treatments than researcher designed ones (though researchers can still *select* a standardised test to align more closely with the difference in treatments thus inflating effect size, which demonstrates that simply adjusting for the standardised/researcher-designed dichotomy is not sufficient). At the most extreme, one can obtain infinite effect sizes for the most proximal of measures. Simpson (2019) outlines a simple thought experiment where the intervention group is taught a single fact while the comparison group is not: An immediate test of recall of that one fact could easily result in completely dissociated outcomes and thus an infinite effect size – for something which may be as trivial as teaching a single fact.

However, the alignment of the measure and the difference in treatments is not the only aspect of measure design which impacts on effect size. For example, a delay between the end of the treatments and the test is likely to decrease effect size. The number of items on a test can also impact on effect size (Simpson, 2017).

To take one example in more depth: The format of the measure can also substantially alter the effect size. In addition to noting the possibility of an infinite effect size for a perfectly aligned measure, Simpson (2019) goes on to consider different formats: If instead of an open response format, a test uses a multiple choice format, the effect size will be deflated (correct guessing both increases the comparison group's mean score and increases the spread of the scores) where the amount of deflation may depend on the number of choices (fewer choices results in a larger chance to guess correctly and wider spread). In addition to the thought experiment, the same paper goes on to give real examples of experiments which use two outcome measures which result in substantially different effect sizes.

3. What effect size measures

Many people in the evidence-based education community treat effect size as a platonic, fixed quantity associated with the particular intervention. From this viewpoint they argue for the use of effect size as a way to distinguish between more or less effective interventions and thus to make recommendations for policy or practice (e.g., Slavin, 1996; Hattie, 2009; Higgins & Katsipataki, 2016). However, they often recognise that there are factors (such as the twelve in the editorial) which can cause the discovered effect size to deviate from this fixed value.

Some with this viewpoint argue that we can recover the platonic quantity ("*the* effect size of the intervention") by adjusting for possible causes of the deviations (e.g., Hunter & Schmidt, 2004). Others seem to argue for contextualising: replacing one such quantity ("the effect size of the intervention") with another ("the effect size of the intervention in this type of context"). For example, Kraft's (2019) focuses on effect sizes from causal evaluation studies of pre-K–12 education interventions measuring student achievement. The contextualising approach still appears to allow for interventions to be compared to each other – directly or to benchmarks – provided the study contexts are sufficiently similar. The editorial can be interpreted as taking this approach, at least for evaluation studies.

I argue that neither of these approaches is adequate. The effect size is not a direct measure of the intervention at all. It is a measure of the experiment as a whole. Effect size should be seen as a measure of the *clarity* of the between group difference in the study: a measure of how well the signal of the experimental finding stands out against the noise of other factors. It can be inflated or deflated by research design choices that remove or add noise (or boost/diminish signal) without changing the intervention. Researchers may intuitively design studies to remove noise where possible to obtain higher effect sizes, just as a photographer intuitively adjusts zoom, focal length and their distance from the subject to achieve a pleasingly sized photograph of an object. Just as the size of the object in a photograph does not directly relate to the size of an object in real life, the effect size in an experiment does not directly relate to the effectiveness of the intervention. Just as a photographer may be restricted in making their adjustments depending on context, a researcher may be restricted in design decisions which help remove noise (Simpson, 2018).

The editorial notes that explanations of many of the twelve highlighted factors "are not always so clear," (Bakker et al., 2019, p. 4) but using the interpretation of effect size as the clarity of the study should help explain them and also highlights other influential factors.

For example, in the discussion of research design, accounting for individual differences (using matched designs or within-subject designs) reduces noise. Using more homogenous samples reduces noise. This interpretation of effect size also explains some of the issues discussed above: The more questions which are unrelated to the difference in treatments which are included in an outcome measure, the more noise is added; replacing an open answer format with a multiple choice format adds noise which comes from guessing correctly; increasing the number of options in a multiple choice response reduces the chance of correct guessing (presuming the options are adequate distractors) and so reduces noise; increasing the delay between the treatments and the outcome measure increases noise.

The eleventh item on the editorial's list – how easily one can influence the dependent variable – might also be explained in these terms. While Savelsbergh et al. (2016) may have found larger effect sizes for achievement variables than for affect variables, this might be because measures of affect may be inherently noisier than achievement measures, since affective constructs may be harder to define and resulting measures may be less reliable. Moreover, in the particular case of Savelsbergh et al., they note they were more likely to include researcher-designed tests for achievement, while accepting only previously validated measures of affect.

To take one further example – the seventh item in the editorial – in more depth, including only participants who received the treatment (rather than all those who researchers intended to be treated) removes noise. This example is simply an extreme manifestation of the impact of treatment fidelity. The more variation there is in the treatment which the intervention group receives, the more their outcome scores might be expected to vary and thus the smaller the effect size.

However, this also applies to the comparison group: the more their treatment varies, the more their outcome scores might be expected to vary and the smaller the effect size. This raises questions about whether simply describing a comparison group as 'business as usual' or not treating the fidelity of the comparison treatment as seriously as that of the intervention group is a mistake. For example, the project reports from the Educational Endowment Foundation, which might be considered the state-of-the-art in rigorous process evaluations and reporting, often go on at great length about the intervention treatment fidelity but hardly mention comparison treatment fidelity. In one particular example – the report of the Shared Maths Project (Lloyd et al., 2015) – four pages are devoted to the intervention treatment in the methodology, with five lines describing the comparison treatment; while in eleven pages describing a process evaluation, the comparison treatment is not mentioned.

4. The Uses of Effect Size

The editorial concludes with a recommendation to compare effect sizes with those from other studies undertaken in similar contexts. However, it is likely to be vanishingly rare that two studies would overlap sufficiently on sample homogeneity, measure proximity, measure length, question design, comparison treatment, test delay, intervention treatment fidelity, comparison treatment fidelity and so on to ever allow relative effect size to be a proxy for the relative effectiveness of the interventions. Contextualising to ever finer situations could be

seen as similar to making the myriad of adjustments proposed in Hunter and Schmidt's (2004) approach to meta-analysis – which even the authors note will normally be impossible to undertake.

Given the rarity with which we could ever conclude that one intervention is more effective than another on the basis of effect sizes from two different studies (or that one *type* of intervention is better than another *type* of intervention on the basis of effect size from two meta-analyses[2]), it would also be a mistake to continue to propose a role for effect size in making decisions about policy and practice. Moreover, even if we could deal with those concerns, we would still have the difficulties associated with transporting results from the study context to policy and practice contexts (see Joyce & Cartwright, 2019).

It is worth asking why effect size should be reported at all. If one believes that studies with bigger effect sizes indicate "better bets" for effective interventions, then of course it makes sense to report them. But as indicated above, this is a misconception: Larger effect sizes can come from experiments conducted more precisely and where more effort was expended to reduce noise. Huge effect sizes can come from interventions which are objectively trivial. Unless we can show compatibility on all contextual factors we will still be left unable to determine whether an intervention with a smaller effect size in one study was less effective than another intervention with a larger effect size in another study or whether the first study was noisier than the other on some unadjusted or unconsidered factor, and we will still be inviting readers to make policy decisions on potentially invalid comparisons.

Thus, apart from those vanishingly rare occasions of studies with identical designs, or the absurdly unlikely case of meta-analyses with identically distributed design attributes, effect size cannot answer the question “is the effect large enough to be useful?” nor be used for policy or practice decisions.

Effect size does retain a potentially valuable role for researchers in terms of replications (Simpson, 2018). A researcher seeking to directly replicate a study with a small effect size might do a number of things. They might increase sample size (which will not change the expected effect size, but increases the chance of detecting the effect if it is real). Alternatively, they might reduce noise by altering other aspects of the research design (for example, by choosing a more homogenous sample, a more proximal test, a different comparison treatment etc.) which increases power by increasing effect size. Similarly, a researcher might use previous research and an understanding of the impact of changes of context to adjust their expected effect size and consequently modify their design. For example, Verkoeijen and Bouwmeester (2014) undertook a close replication of Kornell and Bjork’s (2008) study of the interleaving/blocking effect but discussed carefully the issue of using a less homogenous sample, adjusted their expected effect size and thus increased their sample size accordingly.

Instead of repeatedly adjusting effect sizes in search of a fixed, platonic quantity ("*the* effect size of the intervention") or trying to pin down context so accurately as to create a more nuanced quantity ("the effect size of the intervention in this type of context") we should

[2] Note that averaging across studies, as in meta-analysis, does not allow these factors to “wash out” and somehow make the resulting effect sizes comparable measures of the types of intervention. To do this would require that the factors (sample homogeneity, measure proximity, measure length, question design etc.) are *distributed* equally across the sets of studies for that comparison to be valid and that too is vanishingly unlikely (and normally goes unchecked by meta-analysts and meta-meta-analysts in education).

accept that comparing effect sizes between experiments does not help us determine which is the more useful, effective or important intervention. In almost all cases, we need to abandon the idea that effect size is a measure of the effectiveness of the intervention and that larger effect sizes indicate better interventions, even if we try to account for context. Instead, we can think of effect size as a measure of the clarity of a study as a whole: the extent to which any signal of an effect stands out against the noise of other research design factors.

While the editorial plays a valuable rule in introducing readers to some concerns about effect size, it perpetuates some misconceptions. It is important to recognize that (with rare exceptions) following the editorial's advice will not make effect size any more useful in making educational policy decisions and there is every possibility that policy makers and practitioners will continue to be misled.

References

Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics, 102,* 1–8.

Bergeron, P. J., & Rivard, L. (2017). How to engage in pseudoscience with real data: A criticism of John Hattie's arguments in visible learning from the perspective of a statistician. *McGill Journal of Education/Revue des sciences de l'éducation de McGill,* 52(1), 237–246.

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), 145–153.

Cohen, J. (1973). Brief notes: statistical power analysis and research results. *American Educational Research Journal*, *10*(3), 225–229.

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

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

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.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

IES (2017) *What works clearinghouse: Procedure handbook v4.0.* Retrieved from Institute for Education Studies https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Joyce, K. E., & Cartwright, N. (2019). Bridging the gap between research and practice: Predicting what will work locally. *American Educational Research Journal*, online first.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592.

Kraft, M. (2019). Interpreting effect sizes of education interventions. (EdWorkingPaper: 19-10). Retrieved from Annenberg Institute at Brown University: http://edworkingpapers.com/ai19-10

Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review*, *22*, 215–243.

Lloyd, C., Edovald, T. Morris, S., Kiss, Z., Skipp, A. & Haywood, S. (2015) *Durham Shared Maths Project: Evaluation report and executive summary*, London: Educational Endowment Foundation.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*(3), 158–166.

Savelsbergh, E. R., Prins, G. T., Rietbergen, C., Fechner, S., Vaessen, B. E., Draijer, J. M., & Bakker, A. (2016). Effects of innovative science and mathematics teaching on student attitudes and achievement: A meta- analytic study. *Educational Research Review*, 19, 158–172.

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). Princesses are bigger than elephants: Effect size as a category error in evidence-based education. *British Educational Research Journal*, 44(5), 897–913.

Simpson, A. (2019). Separating arguments from conclusions: The mistaken role of effect size in educational policy research. *Educational Research and Evaluation*, 25(1-2), 99–109.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational researcher*, *15*(9), 5-11.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*(13), 2113–2126.

Verkoeijen, P., & Bouwmeester, S. (2014). Is spacing really the "friend of induction"? *Frontiers in Psychology*, *5*, 259.