THE CENTRE FOR MARKET AND PUBLIC ORGANISATION

"Powered to Detect Small Effect Sizes": You keep saying that. I do not think it means what you think it means.

Michael Sanders and Aisling Ni Chonaire

April 2015

Working Paper No. 15/337

Centre for Market and Public Organisation
University of Bristol
The Priory Road Complex
Bristol BS8 1TU
http://www.bristol.ac.uk/cmpo/

Tel: (0117) 33 11195

E-mail: cmpo-admin@bristol.ac.uk

ISSN 1473-625X

CMPO Working Paper Series No. 15/337

"Powered to Detect Small Effect Sizes": You keep saying that. I do not think it means what you think it means.

Michael Sanders

Behavioural Insights Team and Harvard Kennedy School of Government & CMPO

Aisling Ní Chonaire

Behavioural Insights Team

April 2015

Abstract:

Randomised trials in education research are a valuable and increasingly common part of the research landscape. Choosing a sample size large enough to detect an effect but small enough to make the trial workable is a vital component. In the absence of a crystal ball, rules of thumb are often relied upon. In this paper, we offer criticism for commonly used rules of thumb and show that effect sizes that can be realistically expected in education research are much more modest than studies are powered to detect. This has important implications for future trials, which should arguably be larger, and for the interpretation of prior, underpowered research.

Electronic version:

Address for correspondence
CMPO
The Priory Road Complex, Bristol
BS8 1TU
www.bristol.ac.uk/cmpo/
cmpo-admin@bris.ac.uk
Tel +44(0) 117 33 11195

THE BEHAVIOURAL INSIGHTS TEAM

“Powered to Detect Small Effect Sizes”: You keep saying that. I do not think it means what you think it means.

Michael Sanders

Behavioural Insights Team and Harvard Kennedy School of Government & CMPO

Aisling Ní Chonaire

Behavioural Insights Team

Abstract:

Randomised trials in education research are a valuable and increasingly common part of the research landscape. Choosing a sample size large enough to detect an effect but small enough to make the trial workable is a vital component. In the absence of a crystal ball, rules of thumb are often relied upon. In this paper, we offer criticism for commonly used rules of thumb and show that effect sizes that can be realistically expected in education research are much more modest than studies are powered to detect. This has important implications for future trials, which should arguably be larger, and for the interpretation of prior, underpowered research.

Introduction

Public policy has been subject to a quiet revolution in the past decade, as evidence based policy has come increasingly to the fore, with growing prominence for think tanks and policy organisations that rely on empirical analysis, such as Brookings[1] in the United States and the Institute for Fiscal Studies[2] in the United Kingdom.

Recently, perhaps the noisiest part of this revolution has been in the rise of randomised controlled trials, both as a means for evaluating policy, but as the means by which policy is gradually improved by through a 'test, learn and adapt' framework (Haynes, Service, Goldacre, & Torgerson, 2012). If this is true, nowhere has been noisier in its adoption of these techniques than has the field of education research.

Conducting these field experiments costs money. In the UK, the Education Endowment Foundation (EEF) was created by the Department of Education and the Sutton Trust with an endowment of £125million (~$190,000,000)[3], to focus on reducing the attainment gap between the richest and poorest children in British schools. A range of foundations have funded similar experiments in the United States and elsewhere. It is not unusual for the cost of a single randomised trial in education to run into the hundreds of thousands, or even millions, of dollars.

One of the main challenges in running randomised evaluations in education settings is being able to recruit sufficient numbers of schools in order to be able to be confident in the robustness of the evaluation. This challenge is compounded by the fact that most trials will need to be run by comparing the performance of the schools themselves – rather than comparing the performance of different pupils – in order to avoid contamination effects.

[1] http://www.brookings.edu/.
[2] http://www.ifs.org.uk/.
[3] http://educationendowmentfoundation.org.uk/about/

Each extra unit of sample size increases the chance that an effect of a given size will be found to be statistically significant, at the cost of the total number of studies that can be run (Cohen, 1992).

Sample size is therefore at a premium – in most cases, delivering an intervention to an additional student or school is costly. For a researcher with a finite budget (all researchers that the authors have ever encountered), there is therefore a trade-off between the volume of research that is conducted in terms of the number of studies, and the number of participants (or participating schools), in each study.

There are also ethical concerns arising from overpowered studies – those which recruit more participants than are required to detect their effect size of interest with reasonable power. List (2011) argues that the ethics of a natural field experiment should weigh the potential benefit of the research conducted against the burden placed on the participants. If the burden-per-participant is fixed, and a study has some non-zero value, the expected realised value of the study is a function of this value and the probability that this value will be realised by the present study – which can be approximated for by the power of the study's test. Taking this crude, mechanistic view, we can see experiments either as ethical or unethical depending on whether the expected value minus the total expected burden is greater or less than zero. The relationship between statistical power and sample size means that the merits of a trial on this metric will follow a parabola shape, with both very small nor very large trials requiring extra consideration. Ethical research practice therefore requires that effort is made to balance these two factors. It is worth acknowledging that this parabola is necessarily the result of heavily stylisation – if an intervention is essentially costless relative to the control, and/or if researchers are genuinely agnostic as to the magnitude and directions of effects, the upper limit on ethical sample sizes may not exist.

For these reasons, ex ante power calculations are a vital part of running a randomised trial in general, and particularly in the high-stakes, complicated world of running trials in education. The ability to conduct and have confidence in these calculations is therefore an indispensable tool for researchers. As noted above and in Hutchison & Styles (2010), in many education trials, randomisation occurs at the level of the school, which makes the calculation of power more complicated to achieve in advance. This is particularly true where good baseline data is not available to facilitate the calculation of the intra-cluster-correlation rate (Kerry & Bland, 1998). The growing use of fixed effects models, stepped wedges designs (Hussey, & Hughes, 2007), or trials with split-level randomisation (individual/class/school) (Silva et al., 2015), all add layers of complexity to this problem.

Added to this, it is often unclear what kind of effect size a study should be powered to detect. In practical terms and in relation to subsequent implementation, if an intervention is free or very cheap, even a modest effect size is worth detecting, while if an intervention is very expensive, only a large effect size would justify its wider use. In some fields other than education, for example government subsidies to facilitate small business growth, the clear correspondence between the intervention (money spent in subsidies), and the outcome measure (money raised in gross value added), makes the judgement of what effects are adequate rather straightforward (see Sanders, 2014 for an example of calculations in this field). More often than not, this is not the case for education, where the long term gross value added of even a modestly impactful intervention can be vast, and regular changes in the political priorities and expenditure can make today’s implausibly expensive intervention tomorrow’s centrepiece national policy.

This ambiguity makes rules of thumb useful and widely used. The most pervasive of these are provided by Cohen (1988), who in his review of social science research deemed that an effect with a Cohen's D (standard deviation change) of 0.2-0.3, 0.5, and 0.8 corresponded to small, medium and large effect sizes respectively. There are good reasons to be sceptical of these figures when considering a single sub-field of social science, such as education research, rather than the average across all social science. Lipsey & Wilson (1993), in their review of educational interventions, find that even effective interventions "rarely produce effect sizes greater than half a standard deviation". Wiliam (2008), in his international comparison of educational attainment and its sensitivity finds that the effect of a year's education on standardised tests is roughly 30% of a standard deviation, or Cohen's d of 0.3 – in the range established as a "small" effect. Hence, if a study's sample size is calibrated to have reasonable confidence of detecting an effect of this size, and manages to do so, the likelihood is either that the intervention is incredibly effective or, as shown by Simonsohn, Nelson and Simmons (2014) to be likely, the researchers have simply gotten lucky.

In this paper we attempt to pragmatically pursue a better rule of thumb to use when conducting sample size calculations in education research, by estimating the distribution of effect sizes in previous studies. The next section describes our methodology for drawing together papers. This is followed by the results of this search, and the implications of these for sample size calculations. Finally, we offer brief conclusions.

Methodology

To correct for biases in estimators, we select only randomised controlled trials or similar interventions. Although this is a stricter criteria of quality than may be found in a meta-analysis, it has a broader purpose here. Hutchison and Styles (2010), in their description of sample size calculations, state that “it is important to recall that the terms small, medium, and large are relative, not only to each other, but to the area of behavioural science, or even more particularly, to the specific content and research method being employed in any given investigation” (p.40). We argue that context is also important, as there may be peculiarities to the experimental sample that makes an intervention more or less likely to have an effect. This may be particularly the case for experimental studies compared with quasi-experimental analysis of secondary datasets, as experiments often require not just opting in on the part of participating schools but often the expenditure of a considerable amount of effort and/or money in the execution of the experiment. Belot and James (2013; 2014), investigate both empirically (2013) and theoretically (2014), selection into experiments, and discover theoretically ambiguous opt-in decisions by schools that are unusually either optimistic or pessimistic. This implies that inclusion of non-randomised studies, although helpful for ex-post consideration of average effects across a field of study, may introduce bias when attempting to predict reasonable effect sizes for ex-ante power calculations.

A further selection issue emerges when we consider only those studies published in peer reviewed journals. To the extent that there is a bias in publishing skewed towards only those papers with significant results (Pashler, & Wagenmakers, 2012; Ioannidis, 2005), including only peer reviewed studies will tend to bias our understanding of the distribution of effect sizes upwards. In field experiments, this can occur for two, compounding reasons.

First, the oft noted fact that it is harder to publish a null result than a positive one, which has been covered elsewhere (Rosenthal, 1979; Ioannidis, 2005). Second, to the extent that researchers rely on rules of thumb, such as those of Cohen (1988), which overestimate the likely distribution of outcome measures, the more studies will be run that are underpowered to detect significant effects at conventional levels. Due to the asymptote of statistical power at very small effect sizes, the likelihood is that a greater proportion of studies with smaller effects will fail to achieve statistical significance than studies with larger effects, exacerbating the issue of publication bias. Making extensive use of studies funded by the EEF, for which independent evaluation and conditions of grant awards allow us to see the universe of all completed studies, offers a partial solution to this problem. A slightly wider picture still is attained through the inclusion of working papers (papers that are released by academics prior to completing peer review), which are not subject to the same degree of publication bias.

Procedure

As outlined above, the authors identified relevant studies, both published peer-reviewed articles and working papers, through searches of journal databases and electronic search engines. Leading researchers in the area of educational randomised control trials and field experiments, who have been shown to have an impact on education in public policy settings were also identified. Examination of their research along with reference lists of sourced papers were also examined. As mentioned previously, the EEF publish all of their projects regardless of their statistical significance, and were included as part of the analysis. Meta analyses were excluded as they provide an aggregated mean figure for a set of studies, and their inclusion therefore biases the distribution of effect sizes in our data towards the mean.

The selection criteria was based on three principles: (1) participants were randomly allocated to a treatment or to a control condition; (2) participants were enrolled in primary or secondary school education; and (3) the research outcomes related to educational attainment.

As summarised above, papers were selected on the basis that there was a randomisation of participants and that they were conducted in primary or secondary school educational settings. All papers were vetted to ensure this experimental condition was satisfied. In practical terms, this meant that the experiments took part in a school environment, or that the participants were school students. Finally, the authors were concerned with research where educational attainment, such as test scores or grades, represented the outcome measure. Proxy measures of attainment, such as school attendance or matriculation did not qualify for selection within this study. Furthermore, studies related to pre-school interventions were not included.

As is common in experimental research, the vast majority of papers included a number of effect sizes. In an effort to harness a representative set of effect sizes, the authors systematically focused on the primary outcome measures of the papers, based on the studies' own descriptions, where this measure was relevant to all, or the largest number of participants. Where test scores/grades were secondary outcome measures rather than primary measures, these were included in our collection of effect sizes. In addition, where analysis was conducted with and without covariates, effect sizes related to the inclusion of covariates were included/collected.

Effect sizes related to sub-sample analysis were only included where full sample outcomes were not reported. In the case of EEF studies, where a significant part of their functionality is to increase educational attainment for children of poor families, as identified via "free school meal" status, sub-sample outcomes were included along with full-sample outcomes.

Results and Implications

As can be seen in figure 1, overleaf, the effect sizes reported in the studies we have reviewed are much more conservative than implied by a rule of thumb following Cohen (1988). The accompanying table provides considerable further insight. The median effect size reported in the 113 studies reviewed is a Cohen's D of 0.1 – considerably smaller than even the 0.3 suggested as a small effect. Only 6 studies show effect sizes greater than 0.5, and only 3 show "large" effect sizes of greater than 0.8. The mean effect size is 0.17.

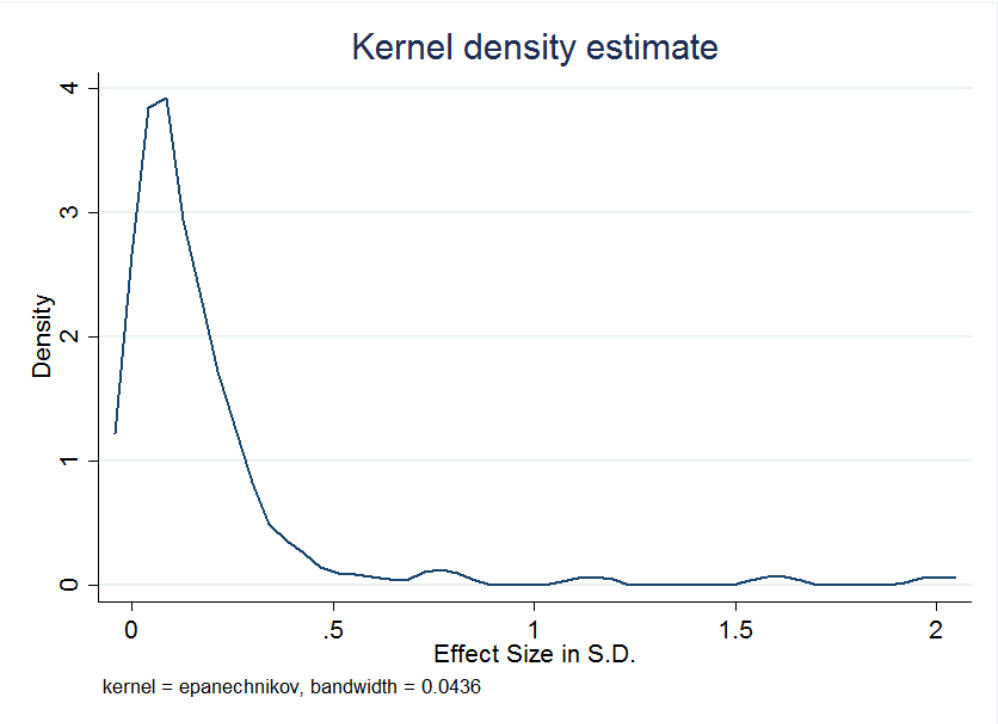

Figure 1: Distribution of effect sizes in education research

Table 1: Distribution of effect sizes	
Distribution:	Effect Size
10%	0.012
25%	0.04
50%	0.1
75%	0.20
90%	0.33
87%	0.3 (“Small”)
95%	0.5 (“Moderate”)
97.5%	0.8 (“Large”)

A fairly systematic failure to design trials with adequate power to detect observed effect sizes is observable. Of 83 interventions for which statistical significance is reported, 53 (63.8%) do not have a significant effect on attainment at conventional levels, of which 51 (61.4%) are not reported as statistically significant at the 10% level.

Impacts:

Although this is of some theoretical interest, our findings are primarily of importance as a practical consideration. We have argued that failure to understand the likely distribution of effect sizes might lead researchers to conduct underpowered research. Armed with this distribution, we can estimate the extent to which studies designed using the rule of thumb will be underpowered. The results of these calculations can be found in the table below.

We take the simplest case of an individually randomised trial, although the findings can be generalised to other cases.

Table 2: Impacts on Power

Effect Size predicted	Power	Sample Size required	Power to detect d = 0.17 (mean)	Power to detect d= 0.10 (median)
0.3	80%	175	35%	15%
0.3	90%	234	45%	19%
0.5	80%	63	15%	8.6%
0.5	90%	85	20%	9.9%
0.8	80%	25	9.0%	6.3%
0.8	90%	33	10.4%	6.9%

Power calculations conducted in R using the pwr library. Power calculations are first conducted for each level of effect size (expressed in cohen's D), to produce required sample size per cell. These figures are then used to calculate power taking the effect size as given at the values found in our data.

Caveats

As with any review of a large body of empirical literature, it is necessary to make caveats about the reliability of our findings. This is particularly the case where we explicitly seek to study the distribution of effect sizes, where one cannot rely on the large sample to erode the effect of any outliers. The most obvious caveats have been mentioned already – that there may be a "file drawer" problem. We investigate this in the figure below, in which we compare effect size with sample size.

The downward curve seen, with effect size decreasing as sample size increases (p=0.086, or 0.020 if unusually large studies are excluded) is broadly suggestive of a file-drawer type publication bias problem, as noted by Ioannidis (2005). However, when comparing multiple studies and/or different interventions, it is not clear that looking at a funnel plot in this way is appropriate. Unlike in medical trials in which the same drug might be tested again and again, the interventions in our data are different, and researchers may have priors about the size of effects likely to be induced by their intervention. Hence, this downward curve may not suggest publication bias but actually that researchers' priors about the power of their interventions are ordinally correct, even if they are not correct in absolute terms. To investigate this, we separate out those studies which are funded by the EEF. As described earlier, all studies funded in this way are published by the EEF, and therefore the same publication biases cannot exist. Here, we find a similar pattern of decline in effect size for studies with larger sample sizes for the 42 EEF studies in our data (p=0.06), with the effect actually slightly larger in absolute terms than for the entire sample. This is broadly consistent with a hypothesis of intelligent design.

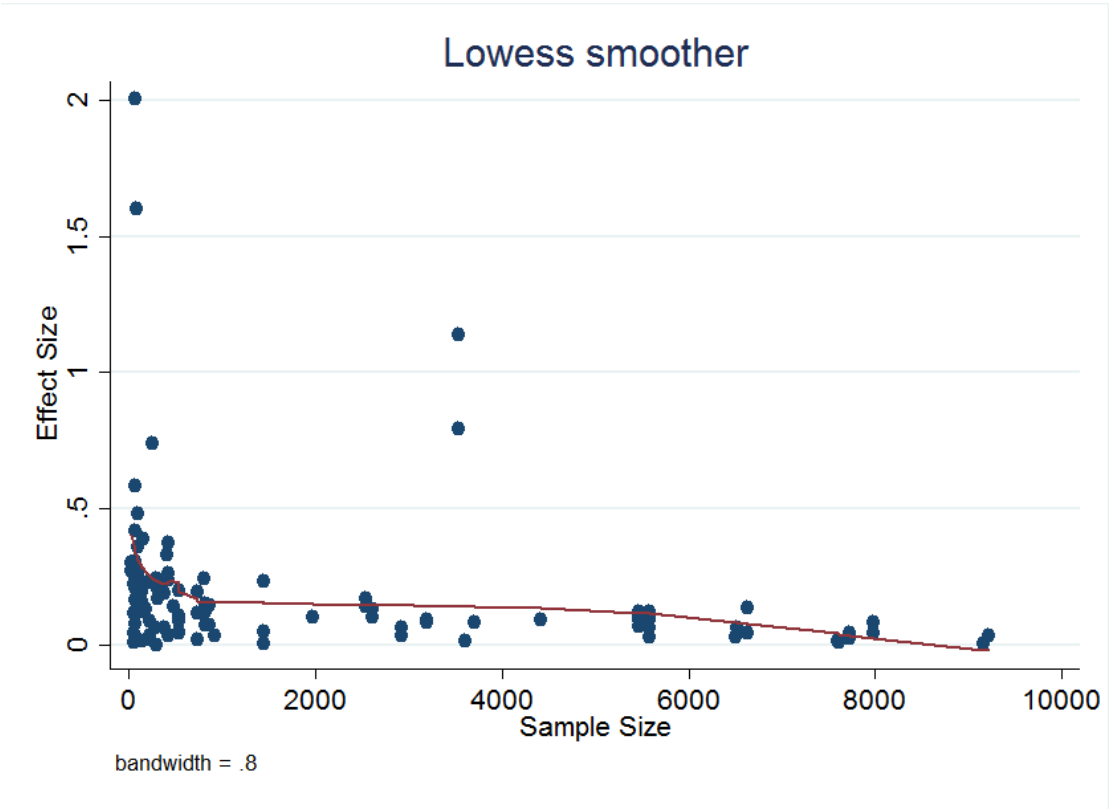

Figure 2: Observed effect sizes and sample size

We should also briefly consider two other possible sources of bias. First, that researchers may be attempting to actively "myth bust" – that is, to dispel widely held untruths about education, and so be deliberately selecting interventions that are likely to fail. Secondly, it is possible that there is some kind of negative publication bias, as this is a relatively new field and studies with positive results may take longer to publish. Given the balance of evidence in this area, we find these sources of bias relatively unlikely.

Conclusions

We have presented an analysis of the effects of over 100 interventions in randomised trials in education. Unlike a traditional meta analysis, we are primarily concerned with the *distribution* of effect sizes, rather than the mean, of these interventions.

Our principal contribution is to show that firstly that sample size calculations based on the rules of thumb proposed by Cohen (1988), are typically over-ambitious, and secondly to propose an alternative rules of thumb based on realised effect sizes. Interestingly, we have also found that although there is considerable and substantial underpowering of randomised trials in education, there is a negative correlation between observed effect size and sample size. As this cannot be explained purely be publication bias, this suggests that researchers' intuition about effects are at least ordinally correct, but unduly optimistic in their absolute value.

Bibliography

Belot, M., & James, J. (2013). Partner Selection into Policy Relevant Field Experiments (No. 236). Edinburgh School of Economics, University of Edinburgh.

Belot, M., & James, J. (2014). A new perspective on the issue of selection bias in randomized controlled field experiments. Economics Letters, 124(3), 326-328.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (Second ed.). New Jersey: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. Psychological bulletin, 112(1), 155.

Haynes, L., Goldacre, B., & Torgerson, D. (2012). Test, learn, adapt: developing public policy with randomised controlled trials. Cabinet Office-Behavioural Insights Team.

Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. Contemporary clinical trials, 28(2), 182-191.

Hutchison, D., & Styles, B. (2010). A guide to running randomised controlled trials for educational researchers. Slough, UK: NFER.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Kerry, S. M., & Bland, J. M. (1998). The intracluster correlation coefficient in cluster randomisation. Bmj, 316(7142), 1455-1460.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. American psychologist, 48(12), 1181.

List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *The Journal of Economic Perspectives*, 3-15.

Nelson, L. D., Simonsohn, U., & Simmons, J. P. (2014). P-curve fixes publication bias: Obtaining unbiased effect size estimates from published studies alone. Available at SSRN 2377290.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science A Crisis of Confidence?. Perspectives on Psychological Science, 7(6), 528-530.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

Sanders, M. (2014). "Growth Vouchers." AEA RCT Registry. January 28. https://www.socialscienceregistry.org/trials/227/history/982

Silva, A, Sanders, M, and Ni Chonaire, A (forthcoming) *"Raising Aspiration – Changing the perceived costs and benefits of going to university."* Working Paper

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534.

Wiliam, D. (2008). 'International comparisons and sensitivity to instruction', Assessment in Education: Principles, Policy and Practice, 15, 3, 253–257.