Whose Prior is it Anyway? A Note on "Rigorous Large-Scale Educational RCTs are Often Uninformative"

Adrian Simpson

School of Education, Durham University

School of Education

Durham University

Leazes Road

Durham

DH1 1TA

Email: adrian.simpson@durham.ac.uk

Acknowledgements

I am grateful to Matthew Inglis for giving generous amounts of time discussing LFI's approach and to Hugo Lortie-Forgues for discussions about Bayes factor calculations.

To appear in *Educational Researcher*

Abstract

A recent paper uses Bayes factors to argue a large minority of rigorous, large-scale education RCTs are ‘uninformative’. The definition of ‘uninformative’ depends on the authors’ hypothesis choices for calculating Bayes factors. These arguably over-adjust for effect size inflation and involve a fixed prior distribution, disregarding the trials’ varied intentions.

Whose Prior is it Anyway? A Note on "Rigorous Large-Scale Educational RCTs are Often Uninformative"

In a fascinating study, Lortie-Forgues and Inglis (2019 – hereafter LFI) conclude 40% of large scale RCTs commissioned by the Education Endowment Foundation (EEF) and the National Center for Education Evaluation (NCEE) are 'uninformative' – i.e. having Bayes factors between $\frac{1}{3}$ and 3.

A Bayes factor is the ratio of likelihoods for the observed data under two 'priors': competing hypotheses reflecting viewpoints about potential trial outcomes. LFI's priors follow a recommendation from Dienes (2014), contrasting a point null hypothesis with the hypothesis of effect sizes from the positive half of a normal distribution centered at zero, with the predicted effect size (PES) as the standard deviation. The positive half normal reflects beliefs that effects from zero to around twice the PES are plausible, with those below the PES around twice as likely as those above (see figure 1). Dienes & Mclatchie (2018) justify this latter feature, arguing that while PES should be theory driven, literature probably inflates average effect size (for example, through publication bias).

LFI's choice fixes PES for all EEF/NCEE studies, suggesting average effect size across education studies of around 0.4. However, they adjust this to 0.2 to account for their own view of the literature's effect size inflation, substantially increasing Dienes' distribution's existing

adjustment for this. LFI's supplementary material examines other fixed PES choices but only covering 0.1 to 0.3.

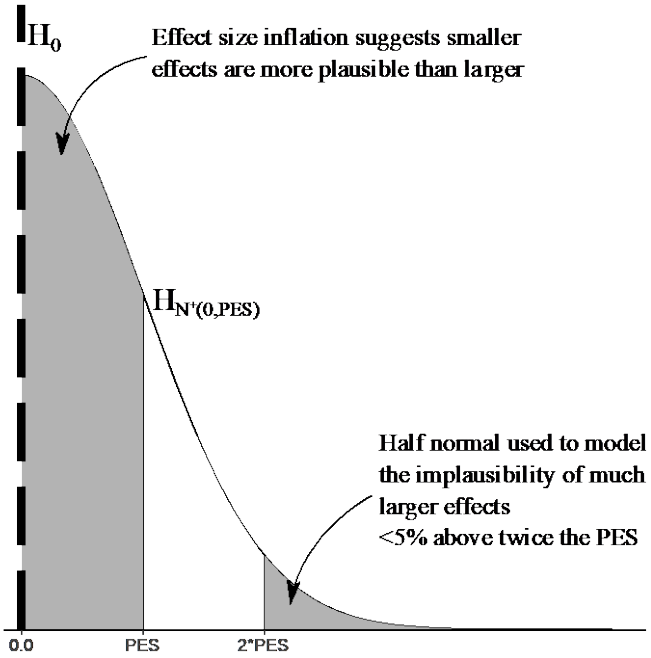

Figure 1: Competing hypotheses: a point null (dashed line) and positive half normal distribution (centre 0, with SD = predicted effect size)

Setting a small, narrow prior distribution of plausible effect sizes inevitably results in the large proportion of 'uninformative' trials, since the trials were generally designed to detect larger effect sizes. LFI appears to conclude, then, that many trials did not give information which they were never designed to give. Examining the reports and protocols of the EEF trials shows that only 12% of those reporting intended *minimum* detectable effect sizes (IMDES) were designed to detect effect sizes smaller than LFI's median (0.13). For the EEF trials, the median reported IMDES is 0.22, and, given this is the *minimum* the researcher would intend to detect, one would

expect them to *predict* effect sizes larger still. With a larger fixed PES, reflecting the general tendency of trial designers, fewer trials would be judged 'uninformative' (e.g. for PES=0.4, only 29% are 'uninformative'). Different reasonable choices of both null and alternative distributions results in even smaller proportions of 'uninformative' trials.

This demonstrates the importance of hypothesis choice to Bayes factors and the definition of 'informative':

> "Representing the predictions of the alternative hypothesis is the part of performing Bayes that requires most thought. … there is no default theory in science, a researcher still has the responsibility of determining whether the default representation used in any Bayes factor reasonably matches the predictions of their theory" (Dienes, 2014, p6)

So different theories may need different hypotheses. Beard, Dienes, Muirhead & West (2016) calculated Bayes factors for a collection of addiction trials, finding 56% of results 'uninformative' (though Beard et al. use 'data insensitive'). They used individual PESs determined separately using knowledge about each trial's underpinning literature, while LFI used a *fixed* prior distribution for plausible effect sizes across all trials, regardless of the trial's aims.

Effect size is misunderstood by many in Evidence-Based Education as intrinsic to the intervention: studies with higher effect sizes are deemed to involve more effective interventions than those in studies with lower effect sizes. This is not the case: the same intervention measured on more proximal outcomes, less active comparison treatments and

more homogenous samples gives higher effect sizes. I.e. effect size is a property of the trial as a whole and is sensitive to researchers' theory-led trial designs (Simpson, 2017). We should not expect the same effect size for 'universal school breakfast' against 'usual home breakfast' for heterogenous samples measured on public examinations as for a writing quality intervention against usual teaching for homogenous samples measured on researcher-selected extended writing tasks.

Some researchers believe they are looking for very small effect sizes while others believe they are looking for much larger ones, while LFI believes they should all be looking in the same narrow band. Across the EEF trials, intended minimum detectable effect sizes stated in reports and protocols vary considerably – from 0.08 to 1 – suggesting that designers' predicted effect sizes also vary considerably. LFI's fixed, narrow alternative hypothesis suggests the plausible range of effect sizes should be from 0 to around 0.4, with the large propensity below 0.2, disregarding researchers' intentions to test different theories with different designs.

If a researcher happens to accurately predict observed effect size, which is then found with reasonable precision, it should be judged positively against an appropriate definition of 'informative'. But Figure 2 shows that LFI's definition mischaracterises them in some circumstances. It illustrates the relationship between Bayes factors and alternative hypothesis SDs for different simulated, moderately precise trials designed with accurate PES, created to illustrate the impact of LFI's fixed alternative hypothesis overlooking design variation: t_1 and t_3

in figure 2 would be ‘uninformative’ for LFI, while judgements based on trial design intentions would characterise them as ‘informative’ in each case.

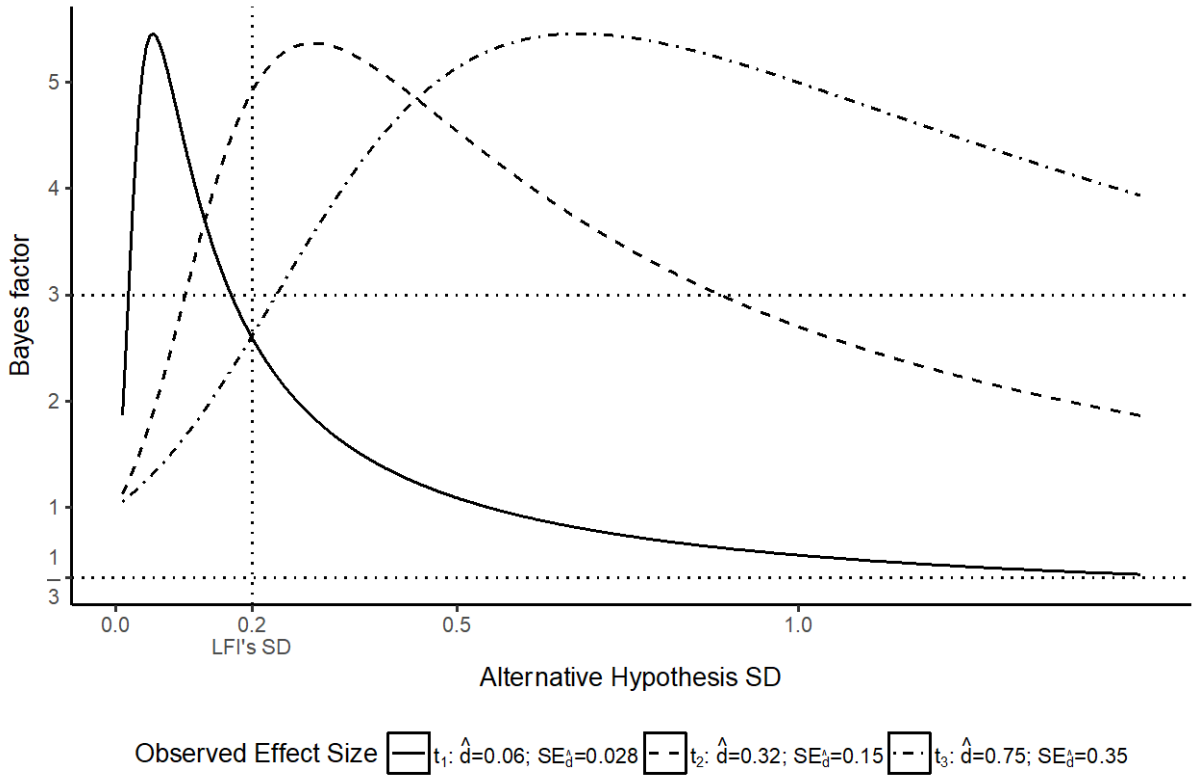

Figure 2: The impact of the alternative hypothesis distribution SD, for three simulated, moderately precise trials with accurate PES (where $\hat{d}$ is the observed effect size)

LFI’s meaning for ‘informative’ is subjective; determined by their choice of prior hypotheses. Setting the alternative more restrictively than researchers intended makes finding many trials ‘uninformative’ inevitable. Fixing it for a field ignores researchers’ intentions in designing different trials for different theories.

LFI suggests, among other things, resource may be being misdirected to a flawed research approach (which others argue is anyway inadequate for policy, e.g. Deaton and Cartwright, 2018) and draws crucial attention to the Bayes factor's potential for identifying the evidential value of trial results.

Certainly, researchers designing evaluation studies need to avoid the mistakes which follow from incorrectly taking effect size to be a property of the intervention: they are led to predict effect size from previous trials of similar interventions, ignoring differing factors like measure, comparison treatment and sample homogeneity. This is likely to lead both to evaluation trials being underpowered and to people drawing incorrect comparisons about the relative effectiveness of different interventions.

But even when researchers avoid this error and then wish to follow LFI's sensible lead in using Bayes factors, the field should not use a fixed, small, narrow standard which defines 'informative' without regard to intentions. Trials should be judged 'informative' or not in their own terms which will vary according to researchers' design choices, using hypotheses tailored for the theory being tested: Whose prior is it anyway?

References

Beard, E., Dienes, Z., Muirhead, C., & West, R. (2016). Using Bayes factors for testing hypotheses about intervention effectiveness in addictions research. *Addiction*, 111(12), 2230-2247.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2-21.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, *5*, 781.

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic bulletin & review*, *25*(1), 207-218.

Lortie-Forgues, H. and Inglis, M. (2019). Rigorous Large-Scale Educational RCTs are Often Uninformative: Should We Be Concerned? *Educational Researcher*, in press.

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