The value of consensus priors: A response to Simpson

The value of consensus priors: A response to Simpson

Hugues Lortie-Forgues
University of York

Matthew Inglis
Loughborough University

Department of Education
University of York
York
YO10 5DD
United Kingdom
Email: hugues.lortie-forgues@york.ac.uk

Abstract

In this response we first show that Simpson's proposed analysis answers a different and less interesting question than ours. We then justify the choice of prior for our Bayes factors calculations, but also demonstrate that the substantive conclusions of our article are not substantially affected by varying this choice.

The value of consensus priors: A response to Simpson

In our article 'Rigorous Large-Scale Educational RCTs Are Uninformative: Should We Be Concerned?' we demonstrated that a surprisingly large number of educational trials are uninformative. Simpson makes two criticisms of our article. First, he suggests that we should have used a subjective Bayes approach to setting priors by defining our alternative hypotheses based on the personal beliefs of trial designers. Second, he argues that our primary analysis relies upon too narrow a range of plausible effect sizes, and that therefore we should have used a more diffuse prior. We disagree on both counts.

Simpson's comment focuses on our use of Bayes factors, rather than any of the frequentist analyses we reported. Bayes factors compare two competing hypotheses: typically a null hypothesis of no effect and an alternative hypothesis which asserts that the true effect size is from some prior distribution. Because Bayes factors are inherently comparative, it is important that readers are interested in the two hypotheses being compared. So the choice of prior distribution for the alternative hypothesis is important.

There are two broad philosophical approaches to choosing priors in Bayesian statistics (e.g., Baguley, 2012; Berger, 2006; Spiegelhalter & Rice, 2009). The so-called subjective Bayes approach, which is apparently strongly favored by Simpson, views the selection of priors as a matter of capturing the beliefs of an individual, usually the researcher conducting the study, about the probable outcomes of the study. In contrast, the objective Bayes approach involves selecting priors that capture the range of plausible effects that are likely to be endorsed by researchers not directly involved in the study, and which "apply over a wide range of contexts and domains" (Baguley, 2012, p. 396). In our analysis we adopted a broadly objective approach by using a similar prior to those default priors widely used in psychological research and implemented in software packages such as JASP (e.g., de Vries and Morey, 2013; Rouder, Speckman, Sun & Morey, 2009; Wagenmakers et al., 2018a, 2018b).

While we would not dismiss the subjective Bayes approach favored by Simpson, it answers a different, and in our view less interesting, question. If you believe the primary purpose of rigorous large-scale educational RCTs is to calibrate the personal beliefs of the researcher who conducted the trial, then Simpson's proposal is the more appropriate analysis. It would allow such researchers to draw the conclusion described by Morey, Wagenmakers and Rouder (2016) as "if you were me, you would believe this" (p. 17). If instead, like us, you believe that rigorous large-scale educational trials should be of interest to the wider

research community, then priors should be based on the beliefs of that wider community. Morey et al. (2016) directly critiqued the position favored by Simpson by saying that "In the context of a scientific argument it is much more useful to have priors that approximate what a reasonable, but somewhat removed researcher would have in the situation." (p. 18). They referred to this flavor of the objective Bayes approach as using 'consensus priors'.

But does the specific consensus prior we used in our primary analysis, a half normal distribution centered at 0 with a standard deviation of 0.2, HN(0,0.2), actually approximate what a reasonable, but somewhat removed, researcher would believe? Recall that this prior suggests that effect sizes found in rigorous large-scale RCTs that test effective interventions will mostly fall between 0 and 0.4, with effects below 0.2 around twice as likely as effects above 0.2. Simpson argues that, because Hattie's (2009) "hinge point" is 0.4, we should have used a prior that captures a wider range of plausible effects. He further suggests that we reached our HN(0, 0.2) prior by adjusting Hattie's figure from 0.4 to 0.2 to account for our "own view of the literature's effect size inflation". This is incorrect. Because Hattie's estimate includes a combination of randomized, non-randomized, and correlational studies, it does not provide useful information about the range of effects we might expect in rigorous large-scale RCTs. Instead, as stated in our article, we based our prior of HN(0,0.2) on Cheung and Slavin's (2016) analysis of how different study characteristics influence effect sizes. The mean effect size of the randomized studies in their sample was 0.16, although they noted that smaller effects than this might be expected in studies with large samples and distant independent measures (features present in the Education Endowment Foundation (EEF) and National Center for Educational Evaluation and Regional Assistance (NCEE) trials we studied). Because Cheung and Slavin's estimate includes both studies that tested effective interventions and those that tested ineffective interventions (i.e. it mixed true nulls and true alternatives), and because our prior needed to only capture a plausible range of effects associated with effective interventions (i.e. we were modeling the alternative hypothesis), we felt, and still feel, that HN(0,0.2) was a reasonable choice.

However, because other researchers might disagree, we also conducted a robustness check. In other words, we repeated our analysis with a variety of different priors, and included the findings in the supplemental materials to the original paper. An expanded version of this robustness check analysis is given in Table 1.

If you believe that most researchers are as optimistic as Simpson, in the sense that they believe that EEF/NCEE-style RCTs which test effective interventions will find standardized effect sizes larger than 0.4 a third of the time, then you will be most interested in

the HN(0, 0.4) line of Table 1. If you agree with us, and believe that most would think that effects larger than 0.4 are very unlikely in these kinds of trials, then you will be more interested in the analysis we conducted (shown in row HN(0, 0.2)).

However, crucially Table 1 reveals that our substantive conclusion is not dramatically altered regardless of which prior you favor. In every case, a surprisingly large proportion of the EEF and NCEE trials in our sample are uninformative (in the sense that they have Bayes factors between 3 and 1/3). We believe this should concern the educational research community.

References

Baguley, T. (2012). *Serious Stats: A Guide to Advanced Statistics for the Behavioral Sciences*. Basingstoke, UK: Palgrave Macmillan.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis, 1*(3), 385-402.

Cheung, A., & Slavin, R. E. (2016). How methodological features of research studies affect effect sizes. *Educational Researcher, 45*(5), 283–292.

de Vries, R. M., & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods, 18*(2), 165-185.

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

Morey, R. D., Wagenmakers, E.-J., & Rouder, J. N. (2016). Calibrated Bayes Factors Should Not Be Used: A Reply to Hoijtink, van Kooten, and Hulsker. *Multivariate Behavioral Research, 51*(1), 11-19.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237.

Spiegelhalter, D. J., and Rice, K. (2009) Bayesian Statistics. *Scholarpedia, 4*(8), 5230.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35-57.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... & Meerhoff, F. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58-76.

Table 1

The percentage of trials in our sample that were uninformative (had Bayes factors between 3 and 1/3), that favored the null, H_0 (had Bayes factors below 1/3), and that favored the alternative, H_a (had Bayes factors above 3), under various different priors for the alternative hypothesis.

H_a Prior Mean	H_a Prior SD	% uninformative	% favoring H_0	% favoring H_a
Half Normal Priors, HN(0, SD)				
0	0.2	40	38	23
0	0.3	35	45	20
0	0.4	29	52	18
0	0.5	28	56	16
Normal Priors, N(0, SD)				
0	0.2	57	27	16
0	0.3	43	43	14
0	0.4	38	51	11
0	0.5	33	57	10
Normal Priors, N(0.2, SD)				
0.2	0.1	29	49	22
0.2	0.2	38	45	18
0.2	0.3	37	48	15
0.2	0.4	30	55	14