Bayesian Evidence Synthesis is No Substitute for Meta-analysis: a Re-analysis of Scheibehenne, Jamil and Wagenmakers (2016)

Rickard Carlsson

Linnaeus University, Sweden

Ulrich Schimmack

University of Toronto, Mississauga, Canada

Donald R. Williams

University of California, Davis, USA

Paul-Christian Bürkner

University of Münster, Germany

Corresponding author:
Rickard Carlsson
Department of psychology
Linnaeus University
391 82 Kalmar, Sweden
Rickard.Carlsson@lnu.se

Scheibehenne, Jamil, and Wagenmakers (2016; SJW) recently introduced Bayesian evidence synthesis (BES). They used it to combine evidence from seven published studies that examined the influence of social-norm messages on hotel towel reuse rates. Although most of the original studies provided non-significant results (p-value > .05), BES provided strong support for the effect (Bayes factor = 37). We think that this conclusion is wrong. We demonstrate that BES is inherently flawed because it pools data in a way that is vulnerable to a Simpson's paradox, and that a Bayesian meta-analysis that avoids this problem produces weaker evidence.

Pooling of Data

Conventional meta-analyses first compute effect sizes from each experiment and then combine them to obtain an overall effect size estimate. In contrast, BES pools all observations into one large dataset—implicitly assuming all observations were obtained from a single study. This approach is flawed, because it is susceptible to the Simpson's paradox. A classic example is the spurious finding of gender bias in admissions to UC-Berkeley (Bickel, Hammel, & O'Connell, 1975). In the pooled analysis, women had lower admission rates compared to men. When the data were analyzed separately for each department, the pattern disappeared. The paradox occurred because women more frequently applied to departments with lower admission rates, consequently lowering their overall admission rate without the presence of gender bias.

The same problem plagues BES. The dependent variable in the seven studies was whether towels were reused or not. For the first two studies (log odds ratio = 0.381 and 0.305), BES showed a combined effect of log odds ratio of 0.340 and a Bayes factor (BF) of 22. When the third study with a lower effect (0.206) was added, the combined effect size

ironically *increased* to 0.361. The BF also increased to 274. In contrast, a meta-analysis of effect sizes with inverse variance weighting showed a decrease to 0.298. This discrepancy occurs because the studies had different allocations of participants to control and experimental condition, as well as different base rates of towel reuse in the control condition.

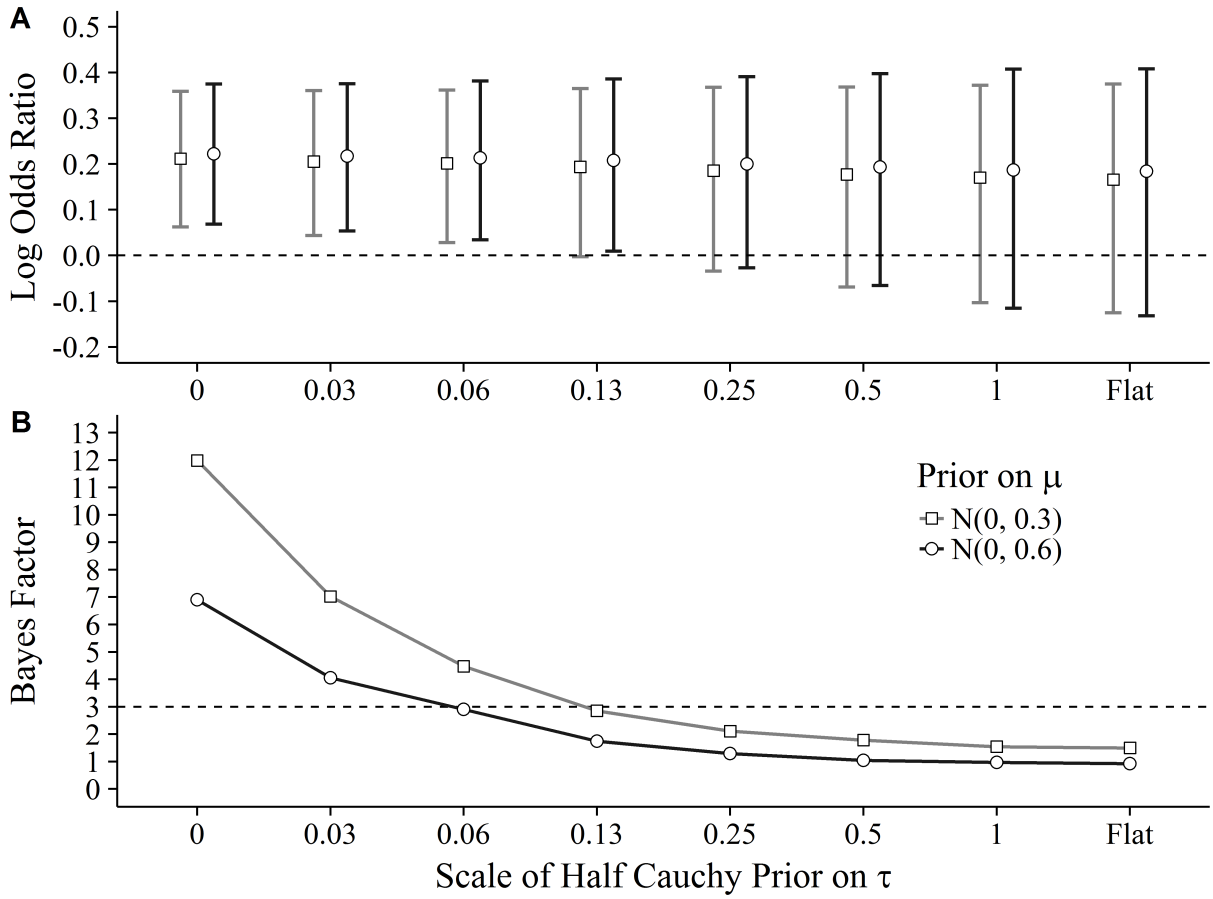
For the total set of studies, the difference was log odds ratio 0.247 vs. 0.226. Although this inflation in effect size is small, with large samples and small effects, even small levels of inflation can substantially affect Bayes factors. Further, because simple pooling can result in considerable inflation, BES will sometimes yield highly misleading evidence.

Bayesian Alternative

We performed a Bayesian meta-analysis without simple pooling of the data (Kruschke & Liddell, 2016), using the brms (Bürkner, in press) package in R. Aside from specifying priors and obtaining posterior distributions for all parameters, this method largely parallels a frequentist random effect analysis. When the number of studies is small, however, frequentist methods can underestimate between-study variability (Chung, Rabe-Hesketh, & Choi, 2013). In contrast, a Bayesian multilevel framework (Gelman, Carlin, Stern, Dunson, Vehtari & Rubin, 2013) allowed us to vary priors on both the between-study variability ($\boldsymbol{\tau}$) and the overall estimate ($\boldsymbol{\mu}$), through which we examined the sensitivity of Bayes factors to various model assumptions (Higgins, Thompson, & Spiegelhalter, 2009; Figure 1b). The strongest evidence for an effect was obtained with the a priori assumption of zero between-study variation (fixed effect assumption: $\boldsymbol{\tau} = 0$; Figure 1a). Although this suggest quite strong evidence (BF_{10} between 7 and 12) it is still substantially lower than the inflated $BF_{10} = 37$ reported by SJW. Moreover, small deviations from this assumption resulted in the evidence ranging from moderate (BF_{10} between 3 and 7) to inconclusive ($BF_{10} < 3$). Indeed, with a flat prior on $\boldsymbol{\tau}$ the intervals include zero which indicates non-significance. To conclude, the

estimate obtained from BES depends on the assumption of a fixed effect size and even a small amount of between-sample variability renders the evidence inconclusive.

Figure 1



Note: (A) Credible intervals of the meta-analytic log odds ratio μ, as well as (B) Bayes factors measuring evidence in favor of a non-zero effect for different prior distributions of μ and τ.

Assessment of Bias

SJW acknowledged that their results could have been inflated by publication bias, but do not assess the presence of publication bias. In contrast, we used the Incredibility index (Schimmack, 2012) and the Test of Insufficient Variance (Schimmack, 2015) to estimate bias. Both tests found no evidence of publication bias. Thus it does not appear that publication bias inflated the evidence for an effect of social-norms messages on towel reuse.

Discussion

An important goal for psychologists is developing methods that can synthesize evidence across multiple studies. The new method SJW introduced, Bayesian Evidence Synthesis (BES), provided strong evidence for an effect of social norm messages on towel reuse in hotels. We showed that BES is vulnerable to the Simpson's paradox and that a multilevel model produced weaker evidence than BES. Whereas BES assumes zero between-study variability, a multilevel model does not make this assumption and allows for examining the influence of heterogeneity on Bayes factors. Indeed, allowing for some variability substantially reduced the evidence in favor of an effect.

Bayesian approaches, especially those using Bayes factors, are becoming more popular in psychology. Even among some proponents of Bayesian methods, however, using Bayes factors as the main criteria for evidence has been criticized (Kruschke, 2011; Liu & Aitkin, 2008). Accordingly, the present analysis is important for several reasons: (1) we provided a Bayesian alternative to simple pooling of data; (2) we demonstrated the value of modeling and of conducting sensitivity analyses; and (3) we elucidated how differing prior distributions can substantially influence the degree of evidence and even the presence of an effect.

In conclusion, we strongly caution against BES and suggest that researchers wanting to use Bayesian methods adopt a multilevel approach. Like other methods, a Bayesian meta-analysis will produce biased results if the data are biased. We therefore recommend that results should be reported together with a bias analysis. In addition, since Bayes factors are sensitive to prior specification (Liu & Aitkin, 2008), they should be reported with sensitivity analyses across a range of reasonable priors.

References

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, (4175). 398.

Bürkner, P.C. (in press). brms: An R Package for Bayesian Multilevel Models using Stan. *Journal of Statistical Software*.

Chung, Y., Rabe-Hesketh, S., & Choi, I. H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine*, *32*(23), 4071–4089. http://doi.org/10.1002/sim.5821

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Boca Raton: CRC Press, Taylor & Francis Group.

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *172*(1), 137–159. http://doi.org/10.1111/j.1467-985X.2008.00552.x

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299-312. doi:10.1177/1745691611406925

Kruschke, J., & Liddell, T. (2016). *The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Planning from a Bayesian Perspective*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2606016

Liu, C., & Aitkin, M. (2008). Bayes factors: prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362-375. doi:10.1016/j.jmp.2008.03.002

Scheibehenne, B., Jamil, T., & Wagenmakers, E. (2016). Bayesian Evidence Synthesis Can Reconcile Seemingly Inconsistent Results: The Case of Hotel Towel Reuse. *Psychological Science (0956-7976)*, *27*(7), 1043-1046.

Schimmack, U. (2012). The Ironic Effect of Significant Results on the Credibility of

Multiple-Study Articles. *Psychological Methods*, *17*(4), 551-566.

Schimmack, U. (2015). *The Test of Insufficient Variance (TIVA): A new tool for the detection of questionable research practices.* R-Index Website. Downloaded from https://replicationindex.wordpress.com/2014/12/30/the-test-of-insufficient-variance-tiva-a-new-tool-for-the-detection-of-questionable-research-practices/ October, 2016

Open Practices

All data, code and supplementary analyses have been made publicly available via the Open Science Framework and can be accessed at http://osf.io/krshq.