Commentary

ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Psychological Science
1–4
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797616684682
www.psychologicalscience.org/PS
SAGE

Bayes Factors From Pooled Data Are No Substitute for Bayesian Meta-Analysis: Commentary on Scheibehenne, Jamil, and Wagenmakers (2016)

Rickard Carlsson[1], Ulrich Schimmack[2], Donald R. Williams[3], and Paul-Christian Bürkner[4]

[1]Department of Psychology, Linnaeus University; [2]Department of Psychology, University of Toronto, Mississauga; [3]Psychology, University of California, Davis; and [4]Institute of Psychology, University of Münster

Received 9/2/16; Revision accepted 11/25/16

Scheibehenne, Jamil, and Wagenmakers (2016) recently introduced a new analytical technique for Bayesian evidence synthesis. They used it to combine evidence from seven published studies that examined the influence of social-norm messages on rates of hotel towel reuse. Although most of the original studies provided nonsignificant results (*p*s > .05), Bayesian evidence synthesis provided strong support for the effect (Bayes factor, BF = 36.89). We think that this conclusion is wrong. We demonstrate that Bayesian evidence synthesis is inherently flawed because it pools data in a way that is vulnerable to Simpson's paradox and that a Bayesian meta-analysis that avoids this problem produces weaker evidence than what Scheibehenne et al. reported.

Pooling of Data

In conventional meta-analyses, effect sizes from each experiment are first computed and then combined to obtain an overall effect-size estimate. In contrast, Bayesian evidence synthesis pools all observations into one large data set—the implicit assumption being that all observations were obtained from a single study. This approach is flawed because it is susceptible to Simpson's paradox. A classic example of this paradox is the spurious finding of gender bias in admissions to the University of California, Berkeley (Bickel, Hammel, & O'Connell, 1975). In the pooled analysis, women had lower admission rates compared with 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, which consequently decreased their overall admission rate without the presence of gender bias.

The same problem plagues Bayesian evidence synthesis. The dependent variable in the seven studies examined by Scheibehenne et al. was whether towels were reused or not. For the first two studies (log odds ratios = 0.381 and 0.305), Bayesian evidence synthesis showed a combined effect of log odds ratio of 0.340 and a BF of 22. When the third study, which had a lower effect size (log odds ratio = 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 in the log odds ratio to 0.298. This discrepancy occurred because the studies had different allocations of participants to control and experimental conditions, as well as different base rates of towel reuse in the control condition.

For the total set of studies, the difference in log odds ratios obtained with Bayesian evidence synthesis and inverse variance weighting was 0.247 versus 0.226, respectively. Although this inflation in effect size is small, with large samples and small effects, even small levels of inflation can substantially affect BFs. Further, because simple pooling can result in considerable inflation, Bayesian evidence synthesis will sometimes yield highly misleading evidence.

Corresponding Author:
Rickard Carlsson, Department of Psychology, Linnaeus University, 391 82 Kalmar, Sweden
E-mail: Rickard.Carlsson@lnu.se

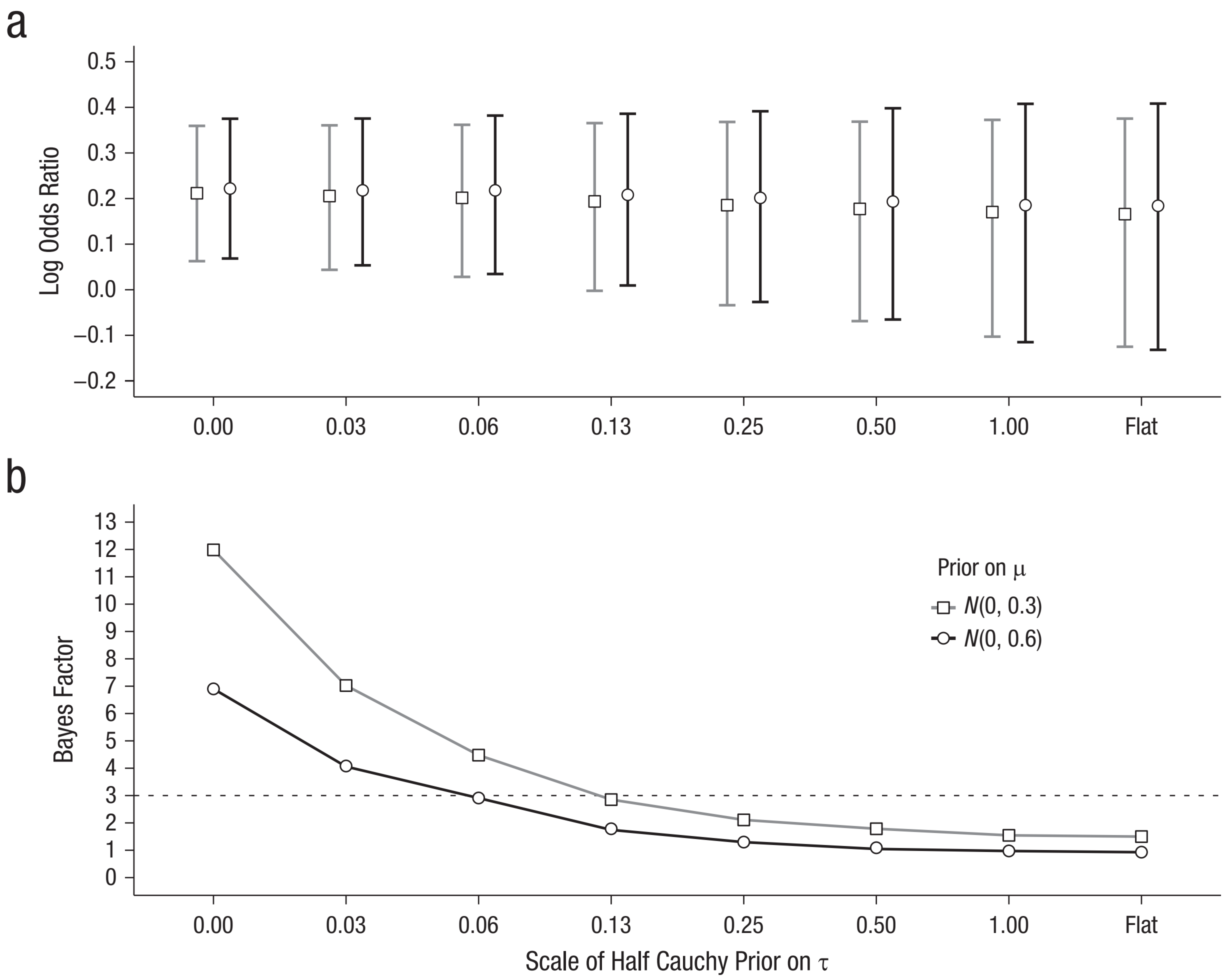

Fig. 1. Results of the Bayesian meta-analysis. The graphs show (a) 95% credible intervals of the meta-analytic log odds ratio μ, as well as (b) Bayes factors measuring evidence in favor of a nonzero effect for different prior distributions of μ (the effect) and τ (the between-studies variability). The dashed line in (b) signifies the threshold for moderate evidence (Bayes factor > 3).

Bayesian Alternative

We performed a Bayesian meta-analysis without simple pooling of the data (Kruschke & Liddell, 2017) using the brms package (Bürkner, in press) in R statistical software. Aside from specifying priors and obtaining posterior distributions for all parameters, this method largely parallels frequentist random-effects analysis. When the number of studies is small, however, frequentist methods can underestimate between-studies variability (Chung, Rabe-Hesketh, & Choi, 2013). In contrast, a Bayesian multilevel framework (Gelman et al., 2013) allowed us to vary priors on both the between-studies variability (τ) and the overall estimate (μ), through which we examined the sensitivity of BFs to various model assumptions (Higgins, Thompson, & Spiegelhalter, 2009; Fig. 1b). The strongest evidence for an effect was obtained with the a priori assumption of zero between-studies variation (fixed-effect assumption: $\tau = 0$; Fig. 1a). Although this suggested quite strong evidence favoring the alternative over the null hypothesis (BF_{10}s between 7 and 12), it was still substantially lower than the inflated BF_{10} of 36.89 reported by Scheibehenne et al. Moreover, small deviations from this assumption resulted in the evidence ranging from moderate (BF_{10}s between 3 and 7) to inconclusive (BF_{10}s < 3). Indeed, with a flat prior on τ, the intervals include zero, which indicates nonsignificance. To conclude, the estimate obtained from Bayesian evidence synthesis depends on the assumption of a fixed effect size, and even a small amount of between-samples variability renders the evidence inconclusive.

Assessment of Bias

Scheibehenne et al. acknowledged that their results could have been inflated by publication bias but did not assess the presence of publication bias. In contrast, we used the incredibility index (Schimmack, 2012) and the Test of Insufficient Variance (Schimmack, 2014) to estimate bias. Both tests showed no evidence of publication bias. Thus, it does not appear that publication bias inflated the evidence for an effect of social-norm messages on hotel towel reuse.

Discussion

An important goal for psychologists is developing methods that can synthesize evidence across multiple studies. The new method that Scheibehenne et al. introduced, Bayesian evidence synthesis, provided strong evidence for an effect of social-norm messages on towel reuse in hotels. We showed that Bayesian evidence synthesis is vulnerable to Simpson's paradox and that a multilevel model produced weaker evidence than Bayesian evidence synthesis. Whereas Bayesian evidence synthesis assumes zero between-studies variability, a multilevel model does not operate under this assumption, which allows researchers to examine the influence of heterogeneity on BFs. Indeed, in the case of Scheibehenne et al.'s data, allowing for some variability substantially reduced the evidence in favor of an effect.

Bayesian approaches, especially those using BFs, are becoming more popular in psychology. Even among some proponents of Bayesian methods, however, using BFs as the main criteria for evidence has been criticized (Kruschke, 2011; Liu & Aitkin, 2008). Accordingly, the present analysis is important for several reasons: (a) We provided a Bayesian alternative to simple pooling of data, (b) we demonstrated the value of modeling and of conducting sensitivity analyses, and (c) 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 Bayesian evidence synthesis and suggest that researchers wanting to use Bayesian methods adopt a multilevel approach. In line with 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, because BFs are sensitive to prior specification (Liu & Aitkin, 2008), they should be reported with sensitivity analyses across a range of reasonable priors.

Action Editor

D. Stephen Lindsay served as action editor for this article.

Author Contributions

R. Carlsson analyzed the data, wrote the section of the manuscript on Simpson's paradox, and drafted an outline of the manuscript. U. Schimmack analyzed the data and wrote the section of the manuscript on bias tests. D. R. Williams and P.-C. Bürkner conducted the Bayesian meta-analysis and wrote the sections of the manuscript pertaining to it. All the authors jointly discussed and agreed on analytical choices (e.g., priors) for the different analyses. All the authors helped revise the manuscript.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

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. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797616684682. This article has received the badge for Open Data. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

References

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*, 398–404.

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*, 4071–4089. doi: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, FL: CRC Press.

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 A: Statistics in Society*, *172*, 137–159. doi: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*, 299–312. doi:10.1177/1745691611406925

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*. Advance online publication. doi:10.3758/s13423-016-1221-4

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

Scheibehenne, B., Jamil, T., & Wagenmakers, E.-J. (2016). Bayesian evidence synthesis can reconcile seemingly inconsistent results: The case of hotel towel reuse. *Psychological Science*, *27*, 1043–1046. doi:10.1177/0956797616644081

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*, 551–566.

Schimmack, U. (2014, December 30). *The Test of Insufficient Variance (TIVA): A new tool for the detection of questionable research practices* [Web log post]. Retrieved from https://replicationindex.wordpress.com/2014/12/30/the-test-of-insufficient-variance-tiva-a-new-tool-for-the-detection-of-questionable-research-practices/