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

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 applied it to a set of original studies that examined the influence of social norms on towel reuse at hotels. While most of the original studies provided non-significant results (*p* > .05), BES provided "strong support" (p. 3) for the effect. Due to methodological limitations, we think that this conclusion is wrong and that BES suffers from several problems. Combining frequentist and Bayesian approaches, we: (1) illustrate the perils of pooling data; (2) assess publication bias, and (3) conduct a Bayesian meta-analysis.

Pooling of Data

The data used in SJW were obtained from experiments designed to investigate the frequency of towel reuse in hotels. To combine these data, conventional approaches first compute effect sizes from each experiment and then a meta-analytic estimate is obtained using a fixed or random effect model. In contrast, BES aggregates all observations into one large dataset, treating them as though they originated from the same study. This pooling is flawed, because it is susceptible to the well-known Simpsons' paradox. A classic example is the finding of gender bias in a pooled data analysis of admissions to UC-Berkeley (Bickel, Hammel, & O'Connell, 1975). When the results were analyzed separately for different departments, the pattern disappeared and the original effect was attributed to gender differences in number of applications to different apartments. For the present data, pooling is especially problematic because the balance between control and experiment group varies across samples. For example, a study with a high base rate of towel reuse had only a small control group and a large experimental group. When pooled with the other data, this sample added more towel reuse participants to the experimental condition than to the control condition, resulting in an incorrect overall estimate.

Assessment of Bias

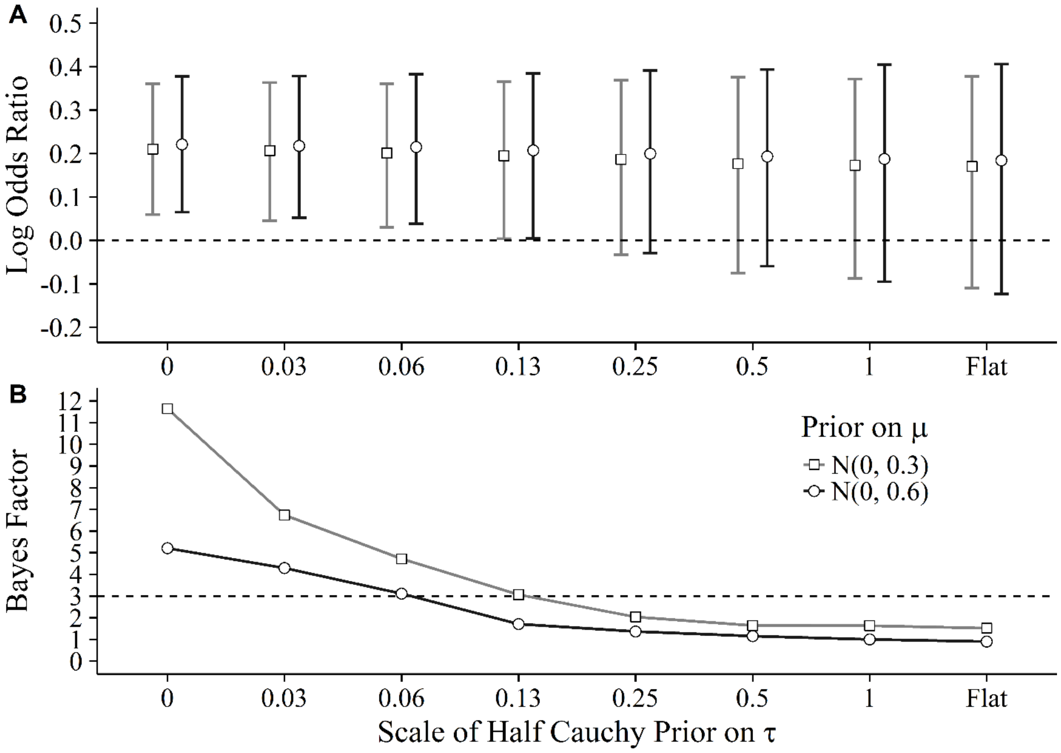
Although publication bias can seriously distort meta-analytical estimates, SJW did not consider this. In contrast, we examined it using two frequentist methods that can be used even with small sets of studies: (1) the incredibility index (IC-index; Schimmack, 2012); and (2) the Test of Insufficient Variance (Schimmack, 2015). The bias tests were applied to a cumulative meta-analysis of z-scores from a series of logistic regressions in each individual dataset. Median observed power for all 7 studies was 28% and the success rate was 29%, suggesting a credible rate of significant results. However, with 28% power, there was a 92% probability (IC-index = .92) of finding at least one non-significant result in the two studies of the seminal article, yet both reported significant results. TIVA also showed insufficient (< 1) variance for the original pair of studies, Var(z) = 0.05, $p < .18$. This suggests that the two studies from the seminal article are not representative and reported inflated effect sizes. In contrast, the full set of seven studies shows no signs of bias, Var(z) = 2.17, $p = .96$ (lower-tail), suggesting a meta-analysis can provide an unbiased effect size estimate.

Multilevel Bayesian Meta-analysis

Although Bayesian statistics has become popular in psychology, the approach gaining traction advocates and promotes the use of Bayes factors. Among proponents of Bayesian statistics, however, the exclusive use of Bayes factors is often criticized (Kruschke, 2011; Liu and Aitkin, 2008). An alternative approach is Bayesian multilevel modeling (Gelman et al., 2013). By incorporating prior information (via informative prior distributions) into the model, for instance, the between study standard deviation ($\boldsymbol{\tau}$) is not underestimated and the overall effect ($\boldsymbol{\mu}$) is not overestimated (Gelman, 2006). Using a multilevel Bayesian approach, we thus performed a meta-analysis on the same studies as SJW (see online supplementary material for details of model specification). In addition to obtaining a meta-analytic estimate and Bayes factor, we varied priors on both **τ** and **μ** to assess sensitivity in the point estimates, credible intervals, and Bayes factors.

In contrast to SJW, we assess the influence of our prior beliefs on the inference (see Figure 1). With a flat prior on **τ**, there is no effect according to the intervals and Bayes factors. As our prior becomes more informative, however, we see the intervals narrow (Figure 1A) and Bayes factors (Figure 1B) become larger. While the credible intervals are less sensitive to the prior, the Bayes factors continue to increase. Indeed, while the point estimate and intervals stabilize, the Bayes factor increase from no evidence to strong evidence in favor of the alternative. This is especially pronounced when a narrower prior on **μ** is used. Based on this sensitivity analysis, the evidence for an effect of social norms on towel reuse is therefore inconsistent and not near the Bayes factor of $BF_{10} = 37$ reported by SJW (see also Table B1 in the online supplement)

Figure. 1.



(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.

Discussion

SJW used a new meta-analytical tool (BES) to examine the effects of social norms on towel-reuse at hotels. They found strong evidence for the effect. Based on careful re-analysis of the data, we argue that the evidence was greatly overstated by SJW, and that their proposed method has serious limitations. Indeed, when data was correctly pooled and the sensitivity of priors considered, the evidence for the small effect is inconclusive.

With Bayesian methods becoming more common, the present re-analysis is also important for several reasons, including: (1) we showed an alternative approach 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, while Bayesian methods are suitable for meta-analysis, we strongly caution against BES and suggest that researchers use a multilevel approach, and include bias estimates, effect size estimates, credible intervals, as well as sensitivity analyses across a range of reasonable priors when Bayes factors are reported.

References

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–534. doi:10.1214/06-BA117A

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.

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

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/ August, 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.