Commentary

Psychological Science
1–4

DOI: 10.1177/0956797617724426
www.psychologicalscience.org/PS

Fixed or Random? A Resolution Through Model Averaging: Reply to Carlsson, Schimmack, Williams, and Bürkner (2017)

Benjamin Scheibehenne[1], Quentin F. Gronau[2], Tahira Jamil[2], and Eric-Jan Wagenmakers[2]

[1]Geneva School of Economics and Management, University of Geneva, and [2]Department of Psychological Methods, University of Amsterdam

Received 1/18/17; Revision accepted 7/14/17

In their Commentary, Carlsson, Schimmack, Williams, and Bürkner (2017) criticize our analysis of hotel towel reuse (Scheibehenne, Jamil, & Wagenmakers, 2016) because we collapsed data across different studies. In addition, Carlsson et al. apply a random-effects model and argue that allegedly minor changes in the parameter priors render the evidence inconclusive. These are two separate issues, and we will address each of them in turn.

Collapsing Data Is Not "Inherently Flawed"

Simpson's paradox describes a situation in which the association between two variables changes when conditioning on a third, such as gender, group membership, or other subpopulations within the data (Pearl, 2014).[1] Carlsson et al. argue that researchers should always control for such subpopulations if the analysis is potentially vulnerable to Simpson's paradox; hence, they consider our original approach of pooling the data across studies "inherently flawed" (p. 1). Controlling for third variables may often be worthwhile, but the decision relies on theoretical considerations, not on statistical criteria or rules of thumb (e.g., Arah, 2008; Pearl, 2014). Even for Carlsson et al.'s Berkeley example, Pearl (2009) showed that depending on the context (i.e., selective application patterns of qualified and unqualified candidates), a pooled analysis provides an unbiased result, whereas conditioning on college wrongly indicates a gender bias even if the colleges did not discriminate.

Nevertheless, our decision to pool the data was guided by pragmatic considerations and the fact that these were replication studies. It would have been prudent to make an informed choice, and we agree with Carlsson et al. that in the case at hand a disaggregated approach is more appropriate. Note that the primary goal of our work was to demonstrate how Bayesian techniques can be used to quantify and continually update evidence as new studies appear. The concepts of Bayesian evidence synthesis are entirely general and do not depend on the details of the statistical model or how the data were combined.

Fixed-Effect or Random-Effects Model?

Carlsson et al. suggest that our data required a random-effects model to allow heterogeneity across studies. However, the random-effects model is more complex than the fixed-effect model that assumes no heterogeneity. As mentioned in our original article, we felt that with only seven data points, the more complex random-effects model would overfit the data (e.g., Myung, 2000). This may seem at odds with the common recommendation in the meta-analysis literature to adopt random-effects models by default, as the fixed-effect assumption is a priori unlikely to hold (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2010). However, this recommendation has recently been challenged on theoretical grounds (Rice, Higgins, & Lumley, 2017). Also, information theory indicates that with only few data, the key goals of prediction and model generalizability are often better served by using simpler models, even when these models are known from the outset to be

Corresponding Author:
Benjamin Scheibehenne, University of Geneva, Geneva School of Economics and Management, 40 Boulevard du Pont d'Arve, CH-1211 Geneva, Switzerland
E-mail: benjamin.scheibehenne@unige.ch

"wrong" (e.g., Grünwald, Myung, & Pitt, 2005). In line with this insight and with established practices (e.g., Albert & Chib, 1997; Moreno, Vázquez-Polo, & Negrín, 2017; Sutton & Abrams, 2001), we conducted a Bayesian model comparison to test which model better predicted the observed data. The tested models were similar to those used by Carlsson et al.—that is, $\mu \sim N(0, 0.3)$—but in line with the directional hypothesis that social norms increase towel reuse, we implemented a one-sided test. More important, this also allowed a closer comparison with the results from our original one-sided analysis. In contrast, Carlsson et al. conducted a less informative two-sided test that almost halved the reported Bayes factors and thus inflated the differences between their results and ours. We consider this a modeling oversight.

Our Bayesian model comparison involves the fixed-effect model without heterogeneity and a range of random-effects models that differ in the extent to which they predict heterogeneity (i.e., models with a more diffuse prior on τ predict more heterogeneity). Thus, the approach encompasses the models contrasted by Carlsson et al. Figure 1 displays the results obtained through bridge sampling (Meng & Wong, 1996).[2] The upper panel shows that the fixed-effect model (white diamonds) yields strong evidence for the effectiveness of social norms (Bayes factor favoring the alternative over the null hypothesis, or $BF_{10} = 24$). For the random-effects models (white triangles), the evidence for the alternative hypothesis decreases as the models assume more between-studies heterogeneity. This replicates Carlsson et al.'s findings. Crucially, however, the comparison reveals that the fixed-effect model dominates the random-effects models (black squares), and increasingly so as the latter assume more heterogeneity. In other words, random-effects models with diffuse priors on τ may yield different results from the fixed-effect model, but these random-effects models are overly complex and do not generalize well. The simple fixed-effect model predicts the observed data best.

The disparate results can be reconciled in a Bayesian model-averaging approach that weights the models' estimated effect sizes with their respective posterior probabilities (e.g., Gronau et al., 2017; Hoeting, Madigan, Raftery, & Volinsky, 1999; Moreno et al., 2017).[3] This approach has the advantage of avoiding the discrete choice for any particular model, and it accounts for uncertainty about which single model is best suited to predict the data (Sutton & Abrams, 2001). The upper panel of Figure 1 shows the results of this model as gray circles, indicating consistent evidence for the alternative over the null hypothesis ($BF_{10} \approx 15$) irrespective of the assumptions regarding between-studies variance. Although quantitatively lower, the findings are qualitatively consistent with our original results. The lower panel of Figure 1 shows the corresponding posterior log odds ratios for both models and the weighted estimate (based on a two-sided analysis). Again, this weighted estimate most closely resembled the fixed-effect model. The results were robust to the choice of alternative (informed) priors.[4]

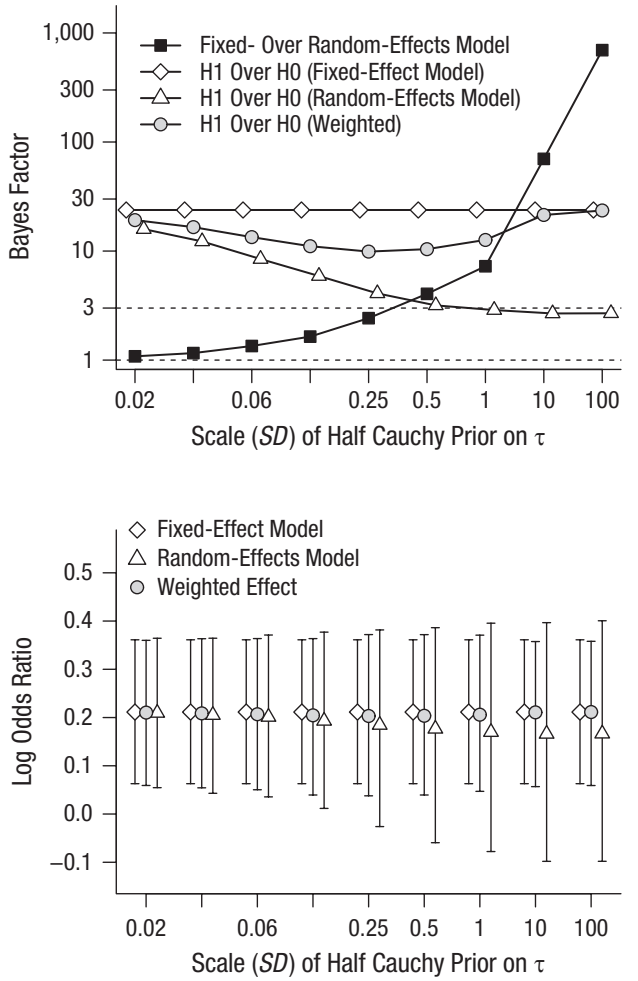

Fig. 1. Results of the Bayesian model comparison. The upper panel shows the Bayes factor (log scale) as a function of prior values on the between-studies variance τ, separately for three models comparing the alternative hypothesis (H1) with the null hypothesis (H0) and the model comparing the fixed-effect and random-effects models. The lower panel shows posterior log odds ratios as a function of prior values on the between-studies variance τ and model type. Error bars are 95% highest-posterior-density intervals. In the upper panel, the dashed line at 1 indicates no evidence for either hypothesis, and the dashed line at 3 marks the cutoff below which evidence for H1 is sometimes considered "hardly worth mentioning" (Jeffreys, 1961, p. 432).

Concluding Comments

Carlsson et al. did not show that our approach was "inherently flawed," but they did show that the interpretation of empirical data depends on the statistical lens employed. As we anticipated in our original article, the lenses proposed by Carlsson et al. are overly

complex and are contraindicated by the data. Viewed constructively, the debate highlights the hidden uncertainty associated with the selection of statistical models (Jeffreys, 1961; Silberzahn & Uhlmann, 2015). Here, we showed how this uncertainty can be accounted for by Bayesian model averaging, which produces relatively constant evidence in favor of the effectiveness of social norms. If new data are included in the future, it is expected that the random-effects model will receive higher weight and hence will increasingly drive the outcome. We believe Bayesian model averaging is a promising tool for meta-analysis and for reconciling statistical disagreement.

Action Editor

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

Author Contributions

All the authors contributed to the statistical analyses, the interpretation of the data, and the writing of 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.

Funding

This research was supported by a grant to the first author from the Swiss National Science Foundation (100014_149846), a grant to the last author from the European Research Council, and a grant from the Berkeley Initiative for Transparency in the Social Sciences, an initiative of the Center for Effective Global Action (CEGA).

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/hjt65/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617724426. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

1. Collapsing the data differs from conducting a fixed-effect meta-analysis in that only the former is potentially affected by Simpson's paradox.
2. See the folder labeled "online supplementary material (R and JAGS script)" on the Open Science Framework (https://osf.io/hjt65) for the underlying R and Just Another Gibbs Sampler (JAGS) code.
3. In line with common practice in Bayesian model averaging, each model receives equal weights a priori (Fragoso & Neto, 2015).
4. If the number of studies is small, the prior on τ can be particularly influential. Hence, alternative priors have been proposed in the literature on Bayesian meta-analyses, and there is no broad consensus on which prior is suited best to model the data (Sutton & Abrams, 2001). See the folder labeled "z_analyses_alternative_priors_between-study_heterogeneity" on the Open Science Framework (https://osf.io/hjt65) for a detailed analysis.

References

Albert, J., & Chib, S. (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *Journal of the American Statistical Association*, *92*, 916–925.

Arah, O. A. (2008). The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect: Covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology*, *5*, Article 5. doi:10.1186/1742-7622-5-5

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97–111.

Carlsson, R., Schimmack, U., Williams, D. R., & Bürkner, P.-C. (2017). Bayesian evidence synthesis is no substitute for meta-analysis: A reanalysis of data from Scheibehenne, Jamil, and Wagenmakers (2016). *Psychological Science*, *28*, 1–5.

Fragoso, T. M., & Neto, F. L. (2015). Bayesian model averaging: A systematic review and conceptual classification. *Cornell University Library arXiv.org*. Retrieved from https://arxiv.org/abs/1509.08864

Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*, 123–138.

Grünwald, P. D., Myung, I. J., & Pitt, M. A. (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, England: Oxford University Press.

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*, 831–860.

Moreno, E., Vázquez-Polo, F. J., & Negrín, M. A. (2017). Bayesian meta-analysis: The role of the between-sample heterogeneity. *Statistical Methods in Medical Research*. Advance online publication. doi:10.1177/0962280217709837

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, England: Cambridge University Press.

Pearl, J. (2014). Comment: Understanding Simpson's paradox. *The American Statistician*, *68*, 8–13.

Rice, K., Higgins, J. P. T., & Lumley, T. (2017). A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society A: Statistics in Society*. Advance online publication. doi:10.1111/rssa.12275

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

Silberzahn, R., & Uhlmann, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature*, *526*, 189–191.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, *10*, 277–303.