Fixed or Random? A Resolution Through Model-Averaging. Reply to Carlsson, Schimmack, Williams, and Burkner.

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

1) University of Geneva

2) University of Amsterdam

1) Corresponding Author:

University of Geneva

Faculty of Economics and Management (GSEM)

40 bd du Pont D'Arve

CH-1211 Geneva

benjamin.scheibehenne@unige.ch

In their comment, Carlsson, Schimmack, Williams, and Bürkner (in press; henceforth CSWB) criticize our analysis (Scheibehenne, Jamil, & Wagenmakers, 2016) because it collapses data across different studies. In addition, CSWB apply a random-effect model and argue that allegedly minor changes in the parameter priors render the evidence inconclusive. These are two separate issues and we address each of them in turn.

Collapsing Data is Not "Inherently Flawed"

Simpson's paradox describes a situation where the association between two variables changes upon conditioning on a third, such as gender, group membership, or other subpopulations within the data (Pearl, 2014).[1] CSWB 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". 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 CSWB'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 CSWB 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

[1] Collapsing the data differs from a fixed-effect meta-analysis, with only the former being potentially affected by Simpson's paradox.

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 was combined.

Fixed-Effect or Random-Effects Model?

CSWB suggest that the data require 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 "wrong" (e.g., Grünwald, Myung, & Pitt, 2005). In line with this insight and 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 CSWB (i.e., $\mu \sim N[0,.3]$), but in line with the directional hypothesis that social norms *increase* towel reuse, we implemented a one-sided test. More importantly, this also allows a closer comparison to the results from our original one-sided analysis. In contrast, CSWB conducted a less informative two-sided test which almost halves the reported Bayes factors and thus inflates 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 CSWB.

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 ($BF_{10} = 24$). For the random-effects models (white triangles), the evidence for H1 decreases as the models assume more between-study heterogeneity. This replicates CSWB. Crucially, however, the model 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 et al., 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 (Sutton & Abrams, 2001). The upper panel shows the results as grey circles, indicating consistent evidence for H1 ($BF_{10} \approx 15$) irrespective

[2] See osf.io/hjt65 for the underlying R and JAGS code.

[3] In line with common practice in Bayesian model-averaging, each model receives equal weights a-priori (Fragoso & Neto, 2015).

of the assumptions regarding between-study variance. Although quantitatively lower, the findings are qualitatively consistent with our original results. The lower panel 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 resembles the fixed-effect model. The results are robust to the choice of alternative (informed) priors.[4]

Concluding Comments

CSWB did not show that our approach is "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 CSWB 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, producing relatively constant evidence in favor of the effectiveness of social norms. If new data is 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.

[4] If the number of studies is small, particularly the prior on τ can be influential. Hence, alternative priors have been proposed in the Bayesian meta-analyses literature and there is no broad consensus on which prior is suited best (Sutton & Abrams, 2001). See 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*(439), 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*(1), 1.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed□effect and random□effects models for meta□analysis. *Research Synthesis Methods*, *1*(2), 97-111.

Carlsson, R., Schimmack, U., Williams, D., & Bürkner, P. C. (in press). Bayesian Evidence Synthesis is No Substitute for Meta-analysis: A Re-analysis of Scheibehenne, Jamil and Wagenmakers (2016). *Psychological Science*.

Fragoso, T. M., & Neto, F. L. (2015). Bayesian model averaging: A systematic review and conceptual classification. *arXiv Preprint* arXiv: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*. 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*. Third Edition. 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.* doi: 10.1177/0962280217709837.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190-204.

Pearl, J. (2014). Comment: understanding Simpson's paradox. *The American Statistician*, *68* (1), 8-13.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Rice, K., Higgins, J., & Lumley, T. (2017). A re□evaluation of fixed effect(s) meta□analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society), online first*. 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*(7), 1043-1046.

Silberzahn, R., & Uhlmann, E. L. (2015). 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*(4), 277-303.

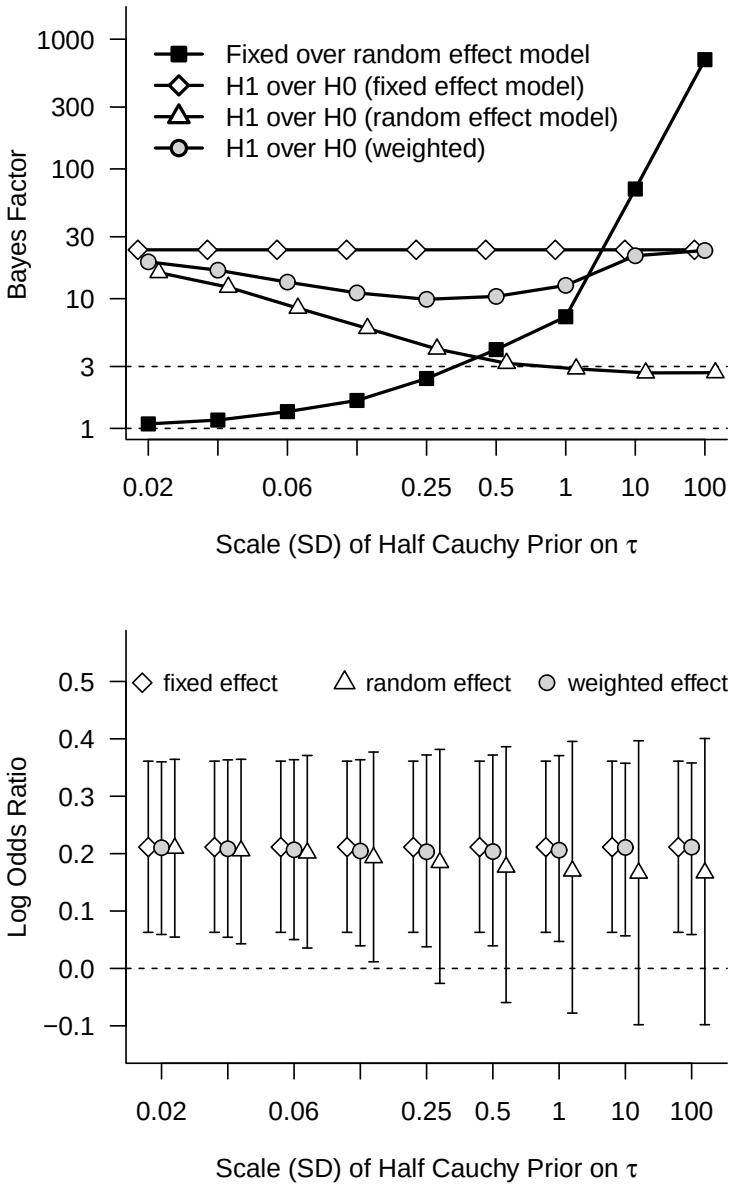

Figure 1: The upper panel shows the Bayes factor (log scale) of the fixed over the random-effects models (black squares) and of H1 over H0 for both type of models (white diamonds and white triangles) across different prior values on the between-study variance τ. The round grey dots indicate the model-averaged Bayes factor of H1 over H0. The lower panel shows the posterior log odds ratios under the respective models. Error bars are 95% highest posterior density intervals.

Author Note:

This research was supported by a grant to the first author from the Swiss National Science Foundation (100014_149846), an ERC grant to the last author, and a grant to the from the Berkeley Initiative for Transparency in the Social Sciences, a program of the Center for Effective Global Action (CEGA), with support from the Laura and John Arnold Foundation.