British Journal of Political Science (2020), page 1 of 3
doi:10.1017/S0007123419000802

British Journal of Political Science

RESPONSE

Rejoinder to Daniel Stegmueller's Comments

Martin Elff[1]*, Jan Paul Heisig[2], Merlin Schaeffer[3] and Susumu Shikano[4]

[1]Department of Political and Social Sciences, Zeppelin University, [2]Health and Social Inequality Research Group, WZB Berlin Social Science Center, [3]Department of Sociology, University of Copenhagen and [4]Department of Politics and Public Administration, University of Konstanz
*Corresponding author. E-mail: martin.elff@zu.de

(Received 5 November 2019; accepted 28 November 2019)

Keywords: multilevel analysis; cross-national comparison; comparative politics; methodology; statistical inference; maximum likelihood

We are grateful to the *BJPolS* editors for giving us the opportunity to respond to Daniel Stegmueller's comment on our article. We also appreciate that he agrees with our main conclusions: that some straightforward corrections can largely resolve the alleged shortcomings of frequentist estimators for multilevel models with small upper-level samples, at least in the set-ups we study, and that there is therefore no pressing need 'to resort to a Bayesian framework' (McNeish 2017, 666). Nevertheless, two points of contention remain.

The first one concerns the relevance and interpretation of our simulation study. Stegmueller (2020) seems to agree with two important improvements of our own simulation study over the one presented in his *AJPS* article (Stegmueller 2013): the use of different random number seeds for different experimental conditions and the increase in the number of Monte Carlo replications. Yet he also claims that our 'discussion might create the impression that the confidence interval coverage bias found for maximum-likelihood (ML) estimates is simply the result of not using' these improvements of the Monte Carlo analysis (Stegmueller 2020). This interpretation is misleading. In our article, we make very clear that the crucial ingredients for improving confidence interval coverage (and statistical inference more generally) are to use restricted maximum likelihood (REML) estimation and a *t*-distribution with limited degrees of freedom (approximated using the Satterthwaite/Kenward-Roger methods or the m–l–1 rule). This was exactly the main point of the simulation study that we report in Figure 3 of our article. The additional simulation evidence that Stegmueller presents in Table 1 in his comment is thus simply a corroboration of it.

Our critique of the design of Stegmueller's Monte Carlo simulation had a different focus. Commenting on Figure 2 of his *AJPS* article, Stegmueller states that 'maximum likelihood estimates are sharply biased upward when the number of countries is fewer than 20 in a hierarchical linear model' (Stegmueller 2013, 753). This finding contrasts with the relevant statistical literature (Jiang 1999; Kackar and Harville 1981), which implies that ML estimates of coefficients in linear multilevel models are *not* biased, and that this holds independently of the number of clusters. We therefore concluded that Stegmueller's finding of such a bias could only be the result of a design flaw in his Monte Carlo analysis. Because the same random seed was used across all experimental conditions, the (pseudo-)random deviations of the average coefficient estimates from their true values tended to go in the same direction in all settings. The main purpose of our extended re-run of Stegmueller's Monte Carlo analysis was to demonstrate that an increase in the Monte Carlo sample size or the variation of random seeds across simulation settings could correct

for this problem. Since this design flaw is far from obvious – otherwise Stegmueller surely would have been able to avoid it – our goal was to 'sensitise both the producers and readers of Monte Carlo simulations to the importance of such technicalities' (Stegmueller 2020).

In Figure 1 of his comment, Stegmueller (2020) presents another simulation study which has important implications for improving statistical inference with multilevel models. He shows that the degrees of freedom of *t*-statistics computed using the Kenward and Roger (1997) method tend to be larger than those obtained with the $m-l-1$ heuristic for models with random slopes on individual-level covariates. The differences found by Stegmueller are certainly not dramatic, but this nevertheless opens up the question of whether the $m-l-1$ heuristic leads to conservative inferences or the Kenward-Roger method leads to anti-conservative inferences. While it is plausible to expect the former, this question deserves further investigation in additional simulation studies.

The second point of contention concerns the use of Bayesian inference as a 'robustness' check. Stegmueller suggests that 'a Bayesian specification […] be estimated as a complementary robustness specification' (Stegmueller 2020). We agree that a congruence of frequentist point estimates and Bayesian posterior modes can underline the validity of the results of a given analysis. In fact, the Bernstein-von Mises theorem implies that any Bayesian posterior mode and the ML estimator of a correctly specified model will converge to each other as the sample size goes to infinity. Yet Stegmueller's term 'robustness specification' is a bit infelicitous. Robustness checks as they are commonly understood in political science involve variations in model construction – for example, which control variables are used, what functional form is assumed for the influence of important covariates, what distribution is assumed for the error term in the model, etc. But the question of whether to use Bayesian or frequentist techniques for estimation arises *after* model construction. If conventional robustness checks indicate that the results depend crucially on details of the model specification, then this indicates that the assumptions leading to the specification of a model should be empirically checked and that, if necessary, further data should be gathered. It is not clear what conclusions one should draw from a discrepancy between frequentist and Bayesian estimates, because it does not have to stem from problems with model specification. It may also result from sample size limitations (so that the Bernstein-von Mises theorem does not apply) or from the prior distribution being so informative that it pulls the posterior mode away from the ML estimate (McNeish 2016). Besides, while Hausman (1978) specification tests rest on comparisons between results yielded by different estimators, their construction and theory are based on frequentist arguments. Thus they provide little guidance for a comparison between frequentist and Bayesian estimators.

To conclude, Stegmueller's results, our own findings and the broader related literature indicate that (a) the accuracy of inference may suffer if standard assumptions about the sampling distribution of estimators (such as asymptotic normality) do not apply and (b) such violations can easily go unnoticed if practitioners use the default settings of statistics packages without further reflection. These issues deserve greater attention in applied work. The main contribution of our article is to show that some straightforward and easy-to-implement steps can greatly improve the performance of frequentist estimators for multilevel models in few-cluster settings. The improvements we suggest are not the only way to improve inference. An alternative is to use the increased power of contemporary computers and rely on computationally intensive methods. Markov Chain Monte Carlo methods have the advantage of avoiding parametric approximations to a Bayesian posterior, which might explain their good performance in Stegmueller's *AJPS* article. Those who do not want to subscribe to Bayesian methodology (be it for philosophical or practical reasons) can turn to bootstrap techniques for statistical inference. As noted in our main article, future work should explore the performance of such computational methods, our proposed methods and possible alternative approaches. We believe that the value of such work is nicely illustrated by Stegmueller's *AJPS* article, our own study and the ensuing exchange in this issue.

References

Hausman J (1978) Specification tests in econometrics. *Econometrica* **46**, 1251–1271.

Jiang J (1999) On unbiasedness of the empirical BLUE and BLUP. *Statistics & Probability Letters* **41**, 19–24.

Kackar RN and Harville DA (1981) Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics-Theory and Methods* **10**, 1249–1261.

Kenward MG and Roger JH (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983–997.

McNeish D (2016) On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal* **23**, 750–773.

McNeish D (2017) Small sample methods for multilevel modeling: a colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research* **52**, 661–670.

Stegmueller D (2013) How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science* **57**, 748–761.

Stegmueller D (2020) Comment on Elff et al. *British Journal of Political Science*, forthcoming.

Cite this article: Elff M, Heisig JP, Schaeffer M, Shikano S (2020). Rejoinder to Daniel Stegmueller's Comments. *British Journal of Political Science* 1–3. https://doi.org/10.1017/S0007123419000802