British Journal of Political Science (2021), **51**, 454–459
doi:10.1017/S0007123419000796

British Journal of Political Science

COMMENT

Comment on Elff et al.

Daniel Stegmueller*

Department of Political Science, Duke University, Durham, NC, USA
*Corresponding author. E-mail: ds381@duke.edu

(Received 15 October 2019; revised 8 November 2019; accepted 28 November 2019; first published online 13 May 2020)

The article by Elff et al. makes a valuable contribution on how to provide post-estimation adjustments to statistical tests in hierarchical models estimated using maximum likelihood (ML) and small group sizes.

My article referenced by them is part of a larger group of articles investigating the performance of 'standard' ML estimation when group sizes are small (see, for example, Bell et al. 2014; Bryan and Jenkins 2015; Maas and Hox 2005). When writing the article (in early 2010), I focused on ML since I perceived it to be the estimation approach most commonly employed in empirical practice.[1] Like others, I found (using a data structure commonly employed in comparative politics) that working with very small group sizes (say, about ten countries) merits some caution.[2]

While diligence might be warranted when estimating models with limited group sample sizes using ML, no one is forced to abandon the approach and convert to a Bayesian viewpoint. A growing literature has discussed adjustments to statistical tests of coefficients on covariates in small group size settings when the model is estimated using (restricted) ML. Several studies suggest using the Kenward and Roger (1997) correction and assess its empirical performance. The general finding is that these corrections seem to perform rather well under small group sample sizes and even under some violations of normality and sphericity (for example, Arnau, Bono and Vallejo 2009; Arnau et al. 2014; Kowalchuk et al. 2004; Luke 2017; Schaalje, McBride and Fellingham 2002; Spilke, Piepho and Hu 2005). Thus, one author concludes a recent article introducing restricted maximum-likelihood (REML) estimation followed by Kenward-Roger (KR) corrected hypothesis tests, with the statement that the availability of 'frequentist corrections […] preclude researchers from necessarily having to resort to a Bayesian framework' (McNeish 2017, 666).[3]

Elff et al. forcefully echo these sentiments. In addition to their general exposition of the logic of post-estimation test correction, they provide ample simulation evidence using a data structure where the lower-level sample size is large and the second-level size is small. They thus provide Monte Carlo evidence that is much closer to data commonly found in observational political science applications (albeit still using a highly stylized data-generating process, a point I will return to below). They also discuss an immensely useful rule of thumb for approximating the degree of freedom used in corrected tests (allowing researchers to potentially skip more computationally expensive approximations, such as the one proposed by KR) and provide simulation evidence of its reliability. I agree with their central points, and I am sure that their contribution will be a major reference point for social scientists working with hierarchical models. Below, I clarify

[1]However, I do accept the author's (implied) criticism that then using the encompassing term 'frequentist' in the article title is a misnomer.

[2]To provide further 'historical' background, one impetus for writing the article was the often-heard objection to estimating hierarchical models using ML *at all* based on rules of thumb demanding group sizes of 30 or 50.

[3]Note that simulation studies performed in this literature usually employ data structures arising from experimental designs. The simulations performed by Elff et al. are much closer to data used in comparative politics applications.

some minor points of contention, provide some thoughts on open questions, and provide an argument for a continued role of Bayesian specifications. I conclude with some thoughts on best practices for estimating hierarchical models.

On the quality of unadjusted and adjusted ML estimates

While not the main focus of their article (or of this comment), Elff et al. point out that my original MC sample size of 1,000 is too small to assess the tail behavior of simulated quantities. They propose a more involved simulation scheme which uses several random seeds for the random number generator when creating MC draws. I agree with their suggestion. However, their discussion might create the impression that the confidence interval coverage bias found for ML estimates is simply the result of not using their proposed strategy of obtaining MC samples. In Table 1 I provide some evidence that this is not the case and that conducting hypothesis tests from (unadjusted) ML estimates using small group sample sizes is indeed rather anti-conservative. The results displayed in Table 1 are based on a simple data-generating process: a random-intercept model with a macro-level covariate (w_j), which is either continuous (Panel A) or dichotomous (Panel B). I follow Elff et al.'s prescription and use 10,000 MC samples based on ten random seeds.[4]

Lines 1 and 4 in Table 1 show large non-coverage for standard ML estimates. Even when looking at the right-tail of MC draws, they are about 10 percentage points too short. Note that switching to REML estimation does move the coverage of confidence intervals somewhat closer to their nominal value, but still leaves them about 4 percentage points too short (on average). A one-sided test of proportions indicates that one cannot reject the null hypothesis that the actual proportion of confidence intervals covering the true parameter value is smaller than 0.95.

Adding the KR correcting perfectly illustrates the point made by Elff et al.: the actual coverage of the adjusted confidence intervals is virtually identical to their nominal 0.95 value. In the case of a continuous covariate, the coverage rate is in fact indistinguishable from its nominal value (in the dichotomous case, the upper limit of the 95 per cent MC interval makes clear that the degree of non-coverage is negligible for any practical purposes). Finally, note that in this simulation, using the (computationally much simpler) $m - l - 1$ heuristic produces confidence intervals that perform equally well.

More realistic data-generating processes and suggestions for further work

The $m - l - 1$ heuristic is highly relevant, not only because it will likely be employed by many practitioners, but also since the authors use it to estimate 'REML-like' generalized linear model specifications. It is expected to work well in many settings, but it would be helpful to gather more evidence on its behavior when simulating more complex or 'realistic' data-generating processes (here I am echoing the criticism of Bryan and Jenkins (2015) that most MC studies do not look like real-world applications).

To give a first indication of how the accuracy of approximation varies with model features, Figure 1 plots histograms of 5,000 MC draws of KR-type degrees of freedom estimates in a more complex random-intercept random-slope model compared to the degrees of freedom suggested by the $m - l - 1$ heuristic. Panel A refers to a macro-level covariate, while Panel B refers to the main effect of a covariate with a random slope.[5]

[4]The initial value for the random number generator is obtained from random.org (again, following Elff et al.). The simulated model is given by $y_{ij} = \alpha_0 + \alpha_1 x_{ij} + \alpha_2 w_j + b_{0j} + \epsilon_{ij}$. The lower-level covariate is created within each group and ranges from 16 to 80 (think of an age variable). The variance of the random intercept, b_{0j}, is 0.2, while the variance of the residuals is 5. The macro-level variable w_j is drawn from a normal distribution with mean 0 and variance 0.7 in the continuous case; in the binary case it is constructed with the proportion of ones equal to 0.4. Fixed effects are: $\alpha_0 = 3$, $\alpha_1 = 1$, $\alpha_2 = 2$.

[5]The simulated model tries to capture a more involved empirical application with two random slopes and a 'cross-level interaction': $y_{ij} = a_0 + a_1 x_{1ij} + a_2 x_{2ij} + a_3 w_j + a_4 w_j x_{1ij} + b_{0j} + b_{1j} x_{1ij} + b_{2j} x_{2ij} + e_{ij}$. Panel A shows KR df estimates for a_3, Panel

https://doi.org/10.1017/S0007123419000796 Published online by Cambridge University Press

Table 1. Actual coverage of nominal 95 per cent confidence interval for a second-level covariate under different estimators and post-estimation corrections

	Monte Carlo estimate		
	Mean	95% CI	p(pr < 0.95)
A: Continuous w_j			
(1) ML	0.884	(0.877, 0.890)	0.000
(2) REML	0.915	(0.910, 0.921)	0.000
(3) REML + KR	0.949	(0.945, 0.953)	0.307
B: Dichotomous w_j			
(4) ML	0.879	(0.872, 0.885)	0.000
(5) REML	0.911	(0.905, 0.917)	0.000
(6) REML + KR	0.943	(0.939, 0.948)	0.001

Note: based on 10,000 Monte Carlo samples with ten random seeds. Sample size is 5,000 rows in 10 equally-sized groups. 10 × 1,000 MC draws. Ten random seeds initialized from master seed obtained from random.org. Final column entry is one-sided p-value for one-sample test of proportion.

Panel A shows that the $m - l - 1$ heuristic coincides almost perfectly with its KR-based counterpart. Panel B shows that the KR degree-of-freedom estimates are systematically larger. In this case the direction of the difference is such that tests using the t-distribution with $m - l - 1$ degrees of freedom will be more conservative than KR-based tests. But I am not sure that this direction is guaranteed in every application.

In future work, it would be of interest to see how the $m - l - 1$ heuristic behaves in extended model setups. I am thinking specifically of two popular variants of hierarchical models. First, what is sometimes called the 'correlated random effects' model, where averages (or other transforms) of (some or all) lower-level covariates are included as second-level predictors (Raudenbush 1989; Wooldridge 2010). Secondly, the class of models involving crossed random effects, which arise when lower-level units are simultaneous members of different groups. A specific example is hierarchical models applied to age-period-cohort analysis (for example, Yang and Land 2008), where the numbers of cohorts is often small. Thinking about implementation of the heuristic and assessing its empirical performance in these models – especially when estimating these models using 'REML-like' approaches with limited dependent variables – would be quite helpful.[6]

What about Bayesian inference?

So what role remains for Bayesian analyses of hierarchical models?[7] Maybe none, as Elff et al. or McNeish (2017) might argue. An alternative view, rooted in the idea of consilience (Whewell 1858, 83–96), might ask for a Bayesian specification to be estimated as a complementary robustness specification. In any given application of an estimator to a fixed set of data (that is, a real research application, not a Monte Carlo simulation) the congruence between different approaches (employing different aassumptions and approximations) can lend additional weight to one's

B for a_1. Covariates are distributed as before, x_2 is drawn from a normal distribution with a variance equal to 5. The data has the structure of an 'unbalanced' panel with within-group sample sizes between 250 and 500. The variances of the random intercept and the two random slopes are 0.6, 0.2. and 0.3; their covariances are 0.05 and 0.1.

[6]One of the reasons I did not consider REML estimates in my original article was the fact that I could not provide a reliable equivalent for generalized linear models. In a number of trial runs (conducted in early 2010), 'REML-like' estimation using PQL produced rather high levels of non-convergence (in more than 10 percent of Monte Carlo runs). It is encouraging that an approach relying on H-likelihood based estimation seems to solve that issue (the authors do not report issues with non-convergence). For a helpful discussion of the H-likelihood (with an admittedly Bayesian bent) see Meng (2009). The discussion provided by the authors in Appendix A.6 is also quite illuminating.

[7]Assuming someone remains agnostic to the siren song of the Bayesian interpretation of probability.

https://doi.org/10.1017/S0007123419000796 Published online by Cambridge University Press

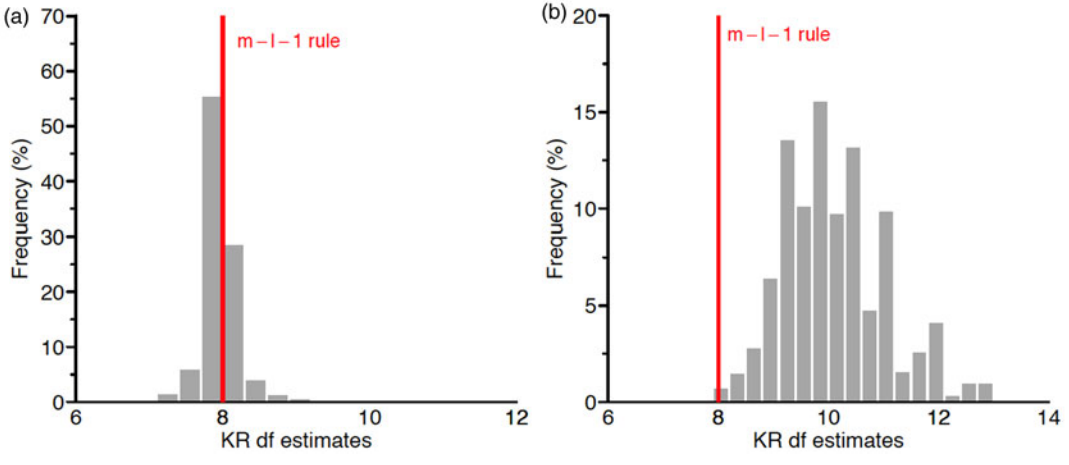

Figure 1. Comparison of KR degrees of freedom approximation with $m - l - 1$ heurstic
Note: Panel A shows results for the second-level covariate. Panel B shows results for a first-level covariate with a random slope. Based on 5,000 MC samples.

results. The last few years have seen tremendous advances easing the application of Bayesian inferential procedures. In particular, the development of the Stan language (Carpenter et al. 2017) and its associated tools have abstracted away some implementation difficulties. Hierarchical models are now available in 'pre-packaged' form (Bürkner et al. 2017) so that researchers can specify models using the familiar R formula interface. Bayesian inference is also possible in *Stata* using a simple prefix statement.[8] Note that I am not advocating that these tools should be deployed without attention to the underlying details. But they are now available with much more computational ease and can serve (without any claim of superiority) as a simple plausibility or robustness check, for example, when using the $m - l - 1$ heuristic in 'REML-like' models without access to KR or Satterthwaite approximations or when estimating models with complex variance structures.

Final notes on best practice

Surely, best practice when working in a frequentist setting should be to estimate the model using REML and to present hypothesis tests either using the t-distribution with $m - l - 1$ degrees of freedom or the KR correction (which is now widely available in *R* and *Stata*). Elff's *iimm R* package makes comparing different strategies as easy as possible. When conducting more involved analyses, for example when using generalized linear models or when employing complex variance structures, such as crossed random effects or autoregressive structures, the original KR correction might not necessarily work well (Kenward and Roger 2009, 2591). I think it would be prudent to check one's analysis using an alternative (computationally expensive) strategy – be it Bayesian or frequentist, as in, for example, a parametric bootstrap. Computational performance is such that these alternative specifications can be run easily while writing one's article.

A more ambitious version of this proposal is to conduct one's own application-specific Monte Carlo simulation, taking estimated coefficients and variance-covariance matrix as the baseline data-generating process against which to compare one's chosen estimator(s). Such a data-specific simulation is easily implemented using standard software (for example, the simulate prefix in *Stata*, or the *MonteCarlo* package in *R*) and can be run while writing the article. Providing information on the expected non-coverage rate of one's tests would provide more transparent communication of the limits (or strengths) of a given set of results.

References

Arnau J et al. (2014) Should we rely on the Kenward–Roger approximation when using linear mixed models if the groups have different distributions? *British Journal of Mathematical and Statistical Psychology* **67**(3), 408–429.

Arnau J, Bono R and Vallejo G (2009) Analyzing small samples of repeated measures data with the mixed-model adjusted F test. *Communications in Statistics: Simulation and Computation* **38**(5), 1083–1103.

Bell BA et al. (2014) How low can you go? *Methodology* **10**(1), 1–11.

Bryan ML and Jenkins SP (2015) Multilevel modelling of country effects: a cautionary tale. *European Sociological Review* **32**(1), 3–22.

Bürkner P-C et al. (2017) Brms: an R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**(1), 1–28.

Carpenter B et al. (2017) Stan: a probabilistic programming language. *Journal of Statistical Software* **76**(1), 1–32.

Kenward MG and Roger JH (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**(3), 983–997.

Kenward MG and Roger JH (2009) An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis* **53**(7), 2583–2595.

Kowalchuk RK et al. (2004) The analysis of repeated measurements with mixed-model adjusted F tests. *Educational and Psychological Measurement* **64**(2), 224–242.

Luke SG (2017) Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods* **49**(4), 1494–1502.

[8]Thus, the minimum addition to estimate a hierarchical linear model in *Stata* using a Gibbs sampler is 'bayes, gibbs:'. This uses a set of default priors, which are of course open to debate in any given application.

https://doi.org/10.1017/S0007123419000796 Published online by Cambridge University Press

Maas CJM and Hox JJ (2005) Sufficient sample sizes for multilevel modeling. *Methodology* **1**(3), 85–91.
McNeish D (2017) Small sample methods for multilevel modeling: a colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research* **52**(5), 661–670.
Meng X-L (2009) Decoding the h-likelihood. *Statistical Science* **24**(3), 280–293.
Raudenbush SW (1989) Centering predictors in multilevel analysis: choices and consequences. *Multilevel Modelling Newsletter* **1**(2), 10–12.
Schaalje GB, McBride JB and Fellingham GW (2002) Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics* **7**(4), 512.
Spilke J, Piepho H-P and Hu X (2005) A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural, Biological, and Environmental Statistics* **10**(3), 374–389.
Whewell W (1858) Novum Organon Renovatum. London: JW Parker and Son.
Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd edn. Cambridge, MA: MIT Press.
Yang Y and Land KC (2008) Age-period-cohort analysis of repeated cross section surveys. Fixed or random effects? *Sociological Methods & Research* **36**(3), 297–326.

Cite this article: Stegmueller D (2021). Comment on Elff et al.. *British Journal of Political Science* **51**, 454–459. https://doi.org/10.1017/S0007123419000796