British Journal of Political Science (2020), page 1 of 15
doi:10.1017/S0007123419000097

British Journal of Political Science

ARTICLE

Multilevel Analysis with Few Clusters: Improving Likelihood-Based Methods to Provide Unbiased Estimates and Accurate Inference

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 9 December 2017; revised 17 January 2019; accepted 1 February 2019)

Abstract
Quantitative comparative social scientists have long worried about the performance of multilevel models when the number of upper-level units is small. Adding to these concerns, an influential Monte Carlo study by Stegmueller (2013) suggests that standard maximum-likelihood (ML) methods yield biased point estimates and severely anti-conservative inference with few upper-level units. In this article, the authors seek to rectify this negative assessment. First, they show that ML estimators of coefficients are unbiased in linear multilevel models. The apparent bias in coefficient estimates found by Stegmueller can be attributed to Monte Carlo Error and a flaw in the design of his simulation study. Secondly, they demonstrate how inferential problems can be overcome by using *restricted* ML estimators for variance parameters and a *t*-distribution with appropriate degrees of freedom for statistical inference. Thus, accurate multilevel analysis is possible within the framework that most practitioners are familiar with, even if there are only a few upper-level units.

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

Multilevel modelling has emerged as the standard tool for quantitative comparative research in the social sciences.[1] The predominant approach to estimating these models, which is also the default in most statistical software packages, is to use likelihood-based methods. However, several studies have raised serious concerns about the performance of these methods when there is a small number of clusters (that is, upper-level units), as is often the case in cross-national research in particular. Probably the most influential study on this topic is Stegmueller's (2013) article *How Many Countries for Multilevel Modelling?*[2] According to the results of Stegmueller's Monte Carlo

[1] Multilevel models as we understand them in this article are sometimes also referred to as 'hierarchical (linear) models' or 'mixed (effects) models'. These models are widely used to analyse multilevel data where lower-level observations are 'nested' in upper-level units (a.k.a. groups or clusters); the simplest example is a two-level structure in which lower-level units (e.g., citizens, students) are nested in one type of upper-level unit (e.g., countries, schools). In such data, there will typically be unexplained group-level variation in the level of the outcome variable and in the strength of lower-level relationships. In multilevel models, this variation is captured by specifying (latent) group-level 'random effects' (a.k.a. 'random intercepts' and 'random slopes'). Several textbooks provide thorough introductions to the approach (e.g., Snijders and Bosker (1999) or Gelman and Hill (2006)). Steenbergen and Jones (2002) provide an accessible introduction in article format.

[2] As of 22 October 2018, Google Scholar recorded 417 and Web of Science 186 citations of this article, making it the most-cited item to have appeared in the *American Journal of Political Science* in 2013 according to both sources.

simulations, ML techniques yield biased point estimates and dramatically anti-conservative inference for the coefficients of contextual variables (for example, of country-level characteristics) in few-cluster settings. That is, when there are fewer than 20 clusters, the point estimates of context effects seem to systematically misrepresent the true effect sizes, and actual coverage rates of confidence intervals are far below the nominal level, implying downward-biased *p*-values and over-rejection of the null hypothesis of no effect. With regard to inferential problems, several other simulation studies reach similar conclusions (see, for example, Bryan and Jenkins 2016; Maas and Hox 2005; Meuleman and Billiet 2009; Moineddin, Matheson and Glazier 2007). Thus the question of whether quantitative multilevel analysis requires a certain minimum number of clusters has long haunted the social sciences.

In this article we seek to make three main contributions to this debate, taking Stegmueller's influential study as our point of departure. The first is to show that Stegmueller's conclusions about the bias of maximum-likelihood (ML) coefficient estimators are incorrect. Statistical theory demonstrates that ML estimators provide unbiased estimates of (contextual and lower-level) effects in *linear* multilevel models. In the *generalized linear* case (for example, multilevel probit), substantial biases can occur when the *size* of the clusters is small, but in most social science applications this will not be a major concern. Our re-analysis of Stegmueller's Monte Carlo evidence supports these claims. Our second contribution is to demonstrate that likelihood-based techniques can achieve accurate inference for contextual effects in few-cluster settings – if researchers make two crucial and easily implemented improvements that we propose below. Monte Carlo simulations show that these improvements work well even with upper-level samples of only five cases, at least in the well-behaved set-up considered by Stegmueller. Our third and last contribution is to highlight some technical, yet important, details that affect the validity of Monte Carlo simulation evidence and that should be more widely known.

Stegmueller frames his findings in terms of comparing conventional 'frequentist' and Bayesian approaches to statistical inference.[3] His pessimistic assessment of (frequentist) likelihood-based methods is paired with a recommendation to adopt Bayesian Markov Chain Monte Carlo (MCMC) techniques in their stead. In contrast to Stegmueller, we do not want to claim that either the frequentist or the Bayesian approach to multilevel analysis is generally superior to the other. Our aim is merely to show how the problems of likelihood-based methods diagnosed by Stegmueller can be solved within a frequentist framework. We believe this is important because many practitioners are unfamiliar with Bayesian inference and may find MCMC techniques computationally expensive and conceptually challenging.

The article is structured as follows. We first summarize statistical theory and previous work to illustrate the conditions under which ML estimates of context effects are unbiased and to derive promising approaches for improving statistical inference. In the subsequent section, we revisit Stegmueller's Monte Carlo simulation evidence. We show how his misleading conclusions concerning the bias of ML estimates arose from a neglect of Monte Carlo error and certain technical aspects of the simulation setup. Crucially, we further demonstrate that restricted maximum-likelihood (REML) estimation, combined with a *t*-distribution with the appropriate degrees of freedom, resolves the inferential deficiencies diagnosed by Stegmueller. In a further section, we re-examine his empirical illustration (drawn from Steenbergen and Jones 2002) and find that REML and Bayesian estimation produce virtually identical results in this 'real-life' setting, provided that our recommendations are followed. We conclude with a summary of our recommendations and make suggestions for future research. In the interest of accessibility, we keep the technical details to a minimum in the main article. Interested readers can find a more extensive discussion of the theoretical and statistical underpinnings of our argument in the online appendix.

[3] The subtitle of his article is 'A Comparison of Frequentist and Bayesian Approaches' and the subsection motivating his analysis is entitled 'Frequentist versus Bayesian Multilevel Models'.

Theoretical Foundations of Accurate Estimation and Inference

Conditions for Unbiased Coefficient Estimates in Multilevel Models

A major conclusion of Stegmueller's study is that ML estimates of the coefficients of contextual (for example, country-level) variables are biased in multilevel models when there are few clusters. This would mean that the estimated coefficients of contextual variables would systematically misrepresent the true effect sizes (more technically, the coefficient estimates would not be equal to their true values in expectation). The existence of such biases would provide strong grounds against using conventional, likelihood-based techniques of multilevel analysis. However, it has long been known that ML estimates of *linear* multilevel model coefficients are unbiased under fairly general conditions (Kackar and Harville 1981). Here, we simply state this property and emphasize that the relevant conditions include those of Stegmueller's simulation study and most practical applications; in particular, the result does not depend on the size of the upper- or lower-level sample (see Appendix A.3 for the technical details).

The case of multilevel logistic regression and other *generalized linear* multilevel models (for example, multilevel probit) is slightly more complicated. For these models, small-sample biases do exist.[4] Crucially, however, the relevant sample size is the one at the *lower level*.[5] In most social science applications – and especially in the case of comparative cross-national analysis that motivated Stegmueller's study – the requirement of a large lower-level sample will typically be met. It is thus unlikely to account for any apparent biases in Stegmueller's Monte Carlo analysis where the minimum lower-level sample size was 2,500 cases (five upper-level units with 500 cases each).

Against this backdrop, we contend that Stegmueller's Monte Carlo results cannot be attributed to the inherent bias of ML estimators, an assertion that is further bolstered by the fact that he found larger apparent biases in the linear case – the case where we can, based on statistical theory, rule out any systematic bias. Our re-analysis below will indeed reconcile the Monte Carlo evidence with the theoretical results noted in this subsection. But before proceeding to the simulation results, we first turn to the issues of variance component estimation and statistical inference for the coefficients of contextual variables.

Bias Correction for Variance Parameter Estimators: Restricted Maximum Likelihood

We have argued that, from the viewpoint of statistical theory, one should not be too concerned about the potential biases of ML *coefficient estimates* in multilevel models: They are unbiased in linear multilevel models. In addition, the finite sample biases in generalized linear multilevel models should not matter much in the many social science applications in which clusters are sufficiently large. Yet ML estimates of *variance parameters* in multilevel models do have a small-sample bias even if the model is linear. More importantly, this bias *is* a serious concern in practice because it emerges when the number of upper-level units is small and does not vanish even if the lower-level sample is very large. The direction of the bias is such that variances at the upper level may be substantially underestimated. In case of a simple random-intercept model, this means that the estimated variance of the random-intercept term will be smaller on average than the true value.

The fact that an ML estimator of a variance parameter exhibits finite sample bias is actually a rather familiar phenomenon. Consider the ML estimator of the population variance σ^2 of a normally distributed variable x based on a sample of n observations $x_i, \ldots, x_n$:

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

[4] This is especially the case when approximation methods such as marginal quasi-likelihood or penalized quasi-likelihood are used; see Breslow and Clayton (1993); Breslow and Lin (1995); Lin and Breslow (1996).

[5] In fact, ML coefficient estimates exhibit a small-sample bias even in the case of standard logistic or probit regression without random effects. That is, coefficient estimates tend to be systematically and substantially larger (in absolute size) than the corresponding true values when the sample size is small; see Firth (1993); Kosmidis and Firth (2009); Zorn (2005).

This estimator is well known to be biased downwards, with the bias being equal to $-(1/n)\sigma^2$ (and thus decreasing with the sample size). In this case, the size of the bias is known exactly and the following bias-corrected estimator can be used instead of the ML estimator:

$$\hat{\sigma}^2_{\text{corr}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Another prominent example where ML estimation yields downward-biased variance estimates is the ML estimator of the error variance in a conventional linear regression model, which is just the mean squared residual. Thus it should not surprise that ML estimators of (upper-level) variance parameters in multilevel models are biased when the number of upper-level units is small.

While unsurprising, biases in the variance parameters can be very consequential even if the variance estimates themselves are not of substantive interest. The reason is that they play a central role in the computation of standard errors for coefficient estimates and related quantities, just as the error variance plays a role in the computation of standard errors in conventional linear regression. If the variance parameters are underestimated, the standard errors of the coefficient estimates will also be underestimated and statistical inference will be anti-conservative.

How, then, can one correct the downward bias of estimated variance parameters in multilevel models? Unlike in the simpler case of the sample variance considered above, there is no generally applicable way to calculate the exact magnitude of the bias in the context of multilevel modelling. Fortunately, however, the *restricted (or residual) maximum-likelihood estimator* (REML) introduced by Patterson and Thompson (1971) can at least greatly reduce the bias and even completely eliminate it in some situations. The construction and derivation of this modification is technically quite involved. Details are available in Appendix Sections A.4 and A.6.

Suffice it to state here that the REML estimator for linear multilevel models has been available as an option in standard software for quite some time, including in programs such as *MLwiN*, the *R* packages *nlme* and *lme4*, *Mplus*, or *Stata*. REML and ML tend to produce similar coefficient estimates, but REML typically yields considerably larger estimates of upper-level variances when there are few clusters. As noted above, these differences in the estimated variance components (that is, random-effects variances) can have important consequences for statistical inference about the coefficients. As we will demonstrate in our re-analysis of Stegmueller's Monte Carlo evidence, the use of the simple, uncorrected ML estimator is indeed one of the reasons why frequentist estimation showed such poor inferential performance in his study.

The REML estimator was developed by Patterson and Thompson (1971) for *linear* multilevel models with normally distributed random effects, and it is not quite clear how the bias correction generalizes to non-linear multilevel models and multilevel models with non-normal random effects (Breslow and Clayton 1993).[6] The performance of Breslow and Clayton's (1993) REML-type modification in the case of a multilevel probit model therefore is another topic of our simulation study. More specifically, we will investigate a variant of the Breslow-Clayton approach that is also referred to as extended quasi-likelihood (EQL) estimation (Lee and Lee 2012).

Improving Hypothesis Tests and Confidence Intervals: Using a t-Distribution with Appropriate Degrees of Freedom

In order to obtain accurate hypothesis test results, *p*-values and confidence intervals, it is not enough to have accurate parameter estimates and standard errors. One also needs to select the appropriate sampling distribution for the test statistics. The standard Wald test statistic for the statistical significance of coefficient estimates is the estimate divided by its standard error. In the case of a linear regression model with *n* observations, *k* independent variables, a constant, and normally distributed errors, this test statistic (also referred to as the *t*-statistic) is known to

[6] See also Drum and McCullagh (1993); Liao and Lipsitz (2002); Noh and Lee (2007).

have a t-distribution with $n - k - 1$ degrees of freedom. For most statistical models other than linear regression with normally distributed errors, the exact sampling distribution of this test statistic is not known and can be approximated at best.

Software for ML estimation usually reports p-values or confidence intervals based on the assumption that the distribution of test statistics can be approximated by the standard normal distribution. This assumption is motivated by the *asymptotic normality* of ML estimators: under certain conditions, ML estimates can be shown to approach the standard normal distribution as the sample size gets large.[7] In the case of multilevel modelling, however, it may be very misleading to rely on asymptotic normality to compute p-values and confidence intervals, in particular for the effects of contextual variables. For example, consider the case of a multilevel analysis of data clustered in 10 upper-level units of size 200 each. The effect of a contextual variable – a variable that represents properties of the upper-level units and therefore does not vary within these units – on a lower-level outcome can be estimated using a multilevel model. An alternative would be to estimate a group-level (for example, a country-level) regression of the group means of the outcome variable on the group means of the independent variables (Heisig, Schaeffer and Giesecke 2017; Lewis and Linzer 2005). Such a group-level (or 'means-as-outcomes') regression would not sacrifice any information about the relationship of interest: All information contributed by the 200 observations within each upper level will be captured by the group means, because the contextual variable cannot explain within-group differences in the outcome.[8]

What is the sampling distribution of the Wald statistic for the coefficient of the contextual predictor in the group-level regression? Provided that the group means are normally distributed, elementary probability theory implies that the Wald statistic of the contextual predictor follows a t-distribution with $10 - 2 = 8$ degrees of freedom. This simple analogy suggests that the t-statistic for the contextual predictor in the corresponding multilevel model might also follow, at least approximately, a t-distribution with 8 degrees of freedom rather than a standard normal distribution.[9] A similar argument can be made for the case of cross-level interactions: For a multilevel model with a cross-level interaction between a contextual and an individual-level independent variable, the coefficient of the interaction term is conceptually related to the coefficient on the contextual variable in a group-level ('slopes-as-outcomes') regression where the outcomes are the within-group slopes of the individual-level variable.[10] This correspondence again suggests that a t-distribution with $10 - 2 = 8$ degrees of freedom should be assumed for the test statistic of the cross-level interaction term. It further suggests that this distribution should also be assumed for the main effect of the lower-level variable included in the interaction, because this term corresponds to the constant term in the group-level slopes-as-outcomes regression. More generally, these considerations motivate what we call the '$m - l - 1$ rule', where m refers to the number of upper-level units and l refers to the number of contextual variables, that is, to the number of observations and independent variables in a group-level approximation of the contextual relationships in a multilevel model, respectively.

Several textbooks on multilevel modelling mention the $m - l - 1$ approximation (for example, Raudenbush and Bryk 2002), but it does not seem to be widely used in practice. In particular, Stegmueller relied on the normal approximation throughout his study. The same holds for

[7] On the conditions for asymptotic normality, see Lehmann and Casella (2011).

[8] A simple (unweighted) group-level regression would ignore one potentially important piece of information – the extent of within-cluster (i.e., lower-level) errors in the dependent variable. Due to differences in this within-cluster variability, some group means may be estimated more reliably than others. Lewis and Linzer (2005) therefore argue that it will usually be preferable to estimate the group-level regression using a feasible generalized least squares approach that gives greater weight to more reliable estimates.

[9] The analogy further suggests that the t-statistic for the overall constant approximately follows this distribution (because it, too, can be seen as a parameter in the group-level regression). We do not pursue this issue further, however, because the overall constant is almost never of substantive interest.

[10] For details, see Lewis and Linzer (2005); Heisig, Schaeffer and Giesecke (2017).

other studies that have found frequentist inference to be anti-conservative in few-cluster settings (Bryan and Jenkins 2016; Maas and Hox 2005; Moineddin, Matheson and Glazier 2007). It seems quite likely that the use of the normal approximation is a major reason why these studies reached such sobering conclusions concerning the accuracy of frequentist inference. When the number of clusters is small, a t-distribution selected according to the $m - l - 1$ rule will lead to much larger p-values and wider confidence intervals than the assumption of asymptotic normality. For example, the critical values for a two-sided test at the 5 per cent significance level are ±1.96 based on the normal, yet ±2.31 based on a t-distribution with 8 degrees of freedom. The fact that these values are also used to identify the limits of 95 per cent confidence intervals implies that the interval based on a t-distribution with 8 degrees of freedom will be approximately 1.18 times as wide as the normal-based alternative.

An alternative to the $m - l - 1$ rule is to use one of several other approximations of the distribution of test statistics that have been proposed in the statistical literature (for overviews, see Li and Redden 2015; Schaalje, McBride and Fellingham 2002). One of the most promising approaches involves a generalization of Satterthwaite's (1946) method, as developed for single-constraint t-tests by Giesbrecht and Burns (1985) and extended to multiple-constraints F-tests by Fai and Cornelius (1996). The most advanced approach is the approximation of Kenward and Roger (1997), yet this method is computationally more demanding than the Giesbrecht-Burns approach and leads to highly similar results for single-constraint tests, which include standard t-tests of the null hypothesis that a given parameter equals zero (Li and Redden 2015). Since the latter type of test is by far the most common in the social sciences, we focus on the Giesbrecht-Burns method in this article. In keeping with widespread practice, we simply refer to the method as the Satterthwaite method/approximation (rather than the Giesbrecht-Burns method) hereafter. This is also the label most commonly used for implementations of the method in statistics packages, including *SAS*, the *R* package *lmerTest* (Kuznetsova, Brockhoff, and Christensen 2017), and *Stata* (since version 14).

A crucial advantage of the Satterthwaite method is that it can provide approximate degrees of freedom for complex multilevel designs (for example, cross-classified structures) where the $m - l - 1$ rule is not applicable. But when dealing with the simple hierarchical structures that predominate in the social sciences (and that are the focus of Stegmueller's analysis) the $m - l - 1$ heuristic may perform quite well and is appealingly simple to implement. We describe in the Appendix how this rule can be applied in practice, using current statistical software. Moreover, the rule can be readily used in the case of generalized linear multilevel models such as multilevel logit and probit. The Satterthwaite approximation is not currently available in all major statistics packages. For generalized linear multilevel models it is, to our knowledge, only available in *SAS* as part of the *GLIMMIX* procedure, but it may become available for *R* or *Stata* in the near future.

In sum, the above discussion points to another possible explanation for the inferential problems detected in Stegmueller's simulation study. Not only did Stegmueller use ML rather than REML estimation, which resulted in downward-biased standard errors; he also relied on the assumption of normally distributed test statistics. The above discussion suggests that a heavier-tailed t-distribution with limited degrees of freedom – approximated using the $m - l - 1$ rule or the Satterthwaite method – may be more appropriate for conducting inference about coefficients that effectively describe group-level relationships. To investigate the importance of this suggestion and of the other claims made above, we now turn to our Monte Carlo simulation study.

Improved Estimation and Inference in Likelihood-Based Multilevel Analysis: Monte Carlo Evidence

We now revisit Stegmueller's influential Monte Carlo analysis. We first demonstrate that the bias he finds in ML parameter estimates is spurious. It can be attributed to an insufficient number of simulated data sets in combination with the choice of a random number seed that accidentally produces extreme results. Moreover, Stegmueller repeatedly started the Monte Carlo simulations for each experimental condition with the random number seed 12345, thereby inadvertently creating a misleading impression of *systematic* bias. In the second step of the Monte Carlo analysis, we show that using REML estimation and the *t*-distribution with appropriate degrees of freedom resolves the inferential problems diagnosed by Stegmueller.

Like Stegmueller, we concentrate on the case where the intra-class correlation (ICC) equals 0.10. Moreover, we focus on the case where the contextual variable has a simple additive effect on the lower-level outcome (a 'direct context effect' in the terminology of Heisig, Schaeffer and Giesecke 2017). Additional results, presented in Appendix B.2, show that all results similarly hold for the three constitutive terms of a cross-level interaction (that is, the interaction term and the main effects of the lower- and upper-level predictors). The additional results also show that our conclusions do not change when we consider ICCs of 0.05 and 0.15.

We present results in the same fashion as Stegmueller (2013) to facilitate comparisons, but additionally visualize Monte Carlo sampling variability via 95 per cent confidence intervals. As demonstrated below, it is important to consider sampling variability in Monte Carlo studies. The basic idea of Monte Carlo analysis is to learn about the properties of an estimator by applying it to a large number of simulated data sets, sampled from a known data-generating process (DGP) with random components. For example, investigating the potential bias of an estimator would involve comparing the average point estimate across many simulated data sets to the true value in the underlying DGP. For an unbiased estimator, this average will approach the true value as the number of simulations gets large. Intuitively, it would be premature to dismiss an estimator as biased just because it happens to be off target in one or two simulated data sets. This is because Monte Carlo results are subject to sampling error, much like conventional parameter estimates. This statistical uncertainty needs to be taken into account when drawing conclusions from Monte Carlo experiments. Stegmueller's failure to do so, and the relatively small number of replications in his study (1,000), seem to be the major reasons why he drew misleading conclusions concerning the bias of ML-based estimators of multilevel model parameters.

Figure 1 visualizes the results concerning the bias of point estimates. We conducted this part of the analysis using a free trial version of *Mplus* and Stegmueller's publicly available replication files.[11] We focus on results for the linear model in the left panel. Conclusions are similar for the probit case in the right panel. Our exact replication (black triangles) obviously reproduces Stegmueller's findings. ML point estimates appear to suffer from systematic upward bias, especially when the number of clusters is 15 or less. However, we also find that the estimated biases are subject to considerable Monte Carlo uncertainty, which Stegmueller did not report. That said, the result that linear multilevel models produce positively biased point estimates is significant at the 5 per cent level when the number of clusters is 15 or less (that is, the 95 per cent confidence intervals do not include zero). Appendix Figure B.1 shows that Stegmueller's original results were never significant for any of the three constitutive terms of the cross-level interaction (that is, the main effects of the lower- and upper-level variables and their interaction).

Given these additional findings and the theoretical results from the previous section, one might suspect that the statistically significant Monte Carlo estimates in Figure 1 are Type-I errors (that is, that they belong to the 5 per cent of cases where a correct null hypothesis of no bias is rejected). To investigate this possibility, we separately introduce two modifications that should not qualitatively change the results if they were indicative of systematic upward bias, but affect them in foreseeable ways if the opposite holds true. The dark grey line with circles shows results based on 10,000 instead of 1,000 replications (leaving everything else the same). The increased Monte Carlo sample size systematically shrinks the estimated magnitude of parameter bias towards zero, just as one would expect if the large deviations found in the original analysis

[11] We used the replication files available on the *American Journal of Political Science* dataverse (Stegmueller 2012).

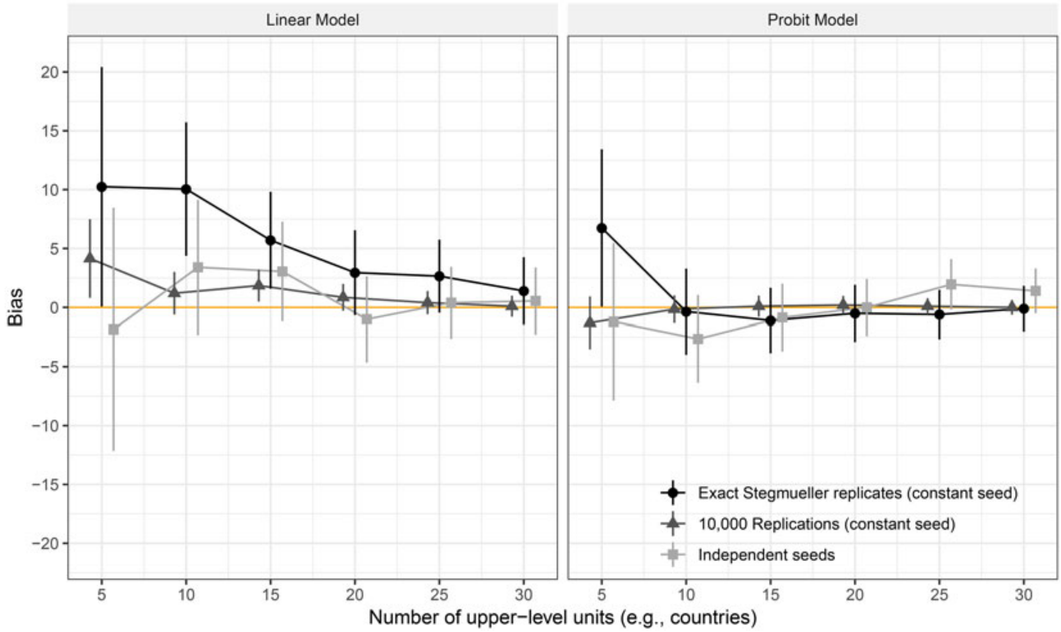

Figure 1. Performance of ML point estimates of upper-level covariate effects in multilevel linear and probit models
Note: The figure displays relative biases of ML point estimates (in percent of the true effect size). Vertical lines depict 95 per cent Monte Carlo confidence intervals for these results. The horizontal zero line denotes the reference of no bias. Black triangles replicate the results presented in the left column ('Estimate') of Figure 2 on page 754 in Stegmueller (2013). We additionally present two modifications of Stegmueller's analysis. The first (black circles) increases the number of replications from 1,000 to 10,000, leaving everything else the same. The second (grey squares) follows Stegmueller in using only 1,000 replications, but specifies different random number seeds for the different experimental conditions. The Monte Carlo confidence intervals are computed on the base of the standard deviation of the estimates across Monte Carlo replications divided by the square root of the Monte Carlo sample size (the number of simulation replications), and the 2.5 and 97.5 percentiles of the standard normal distribution (i.e. −1.96 and +1.96).

were due to chance. However, the direction of the estimated biases remains consistently positive (and statistically significant in the 5 and 15 cluster conditions).

Our second modification swipes this pattern away. Stegmueller repeatedly used the seed 12345 to initialize Monte Carlo sampling for each single experimental condition. A random number seed initializes a specific pseudo-random sequence of numbers. While the sequence 'behaves' like a truly random sequence statistically, a given random number seed, such as 12345, always initializes the exact same sequence. Specifying a random number seed thus guarantees the reproducibility of results, but it comes at the risk of creating unwanted interdependencies and repeatedly producing similar chance findings by using non-independent seeds. This indeed seems to be the reason why Stegmueller's simulation results are suggestive of systematic upward bias in the parameter estimates, especially for the linear case. To illustrate this, our second modification replaces the repeated use of the seed 12345 with different and independent seeds for each experimental condition.[12] The grey squared line shows that any systematic pattern of positive bias disappears when we use a different random number seed for each condition (like the original

[12] We generated the random number seeds for the different experimental conditions using www.random.org, an online resource that exploits atmospheric noise to generate numbers that are truly random (rather than pseudo-random). The exact values of the seeds are documented in the *Mplus* replication files. Alternatively, one could set a random seed only once and successively simulate all experimental conditions. Yet, Stegmueller's replication files do not allow such a setup, as each experimental condition is based on a separate *Mplus* '.inp' file. In another re-examination implemented in *R*, we ran the simulations in this alternative way. The results are similar to the ones based on independent seeds which are displayed here. They are available from the authors upon request.

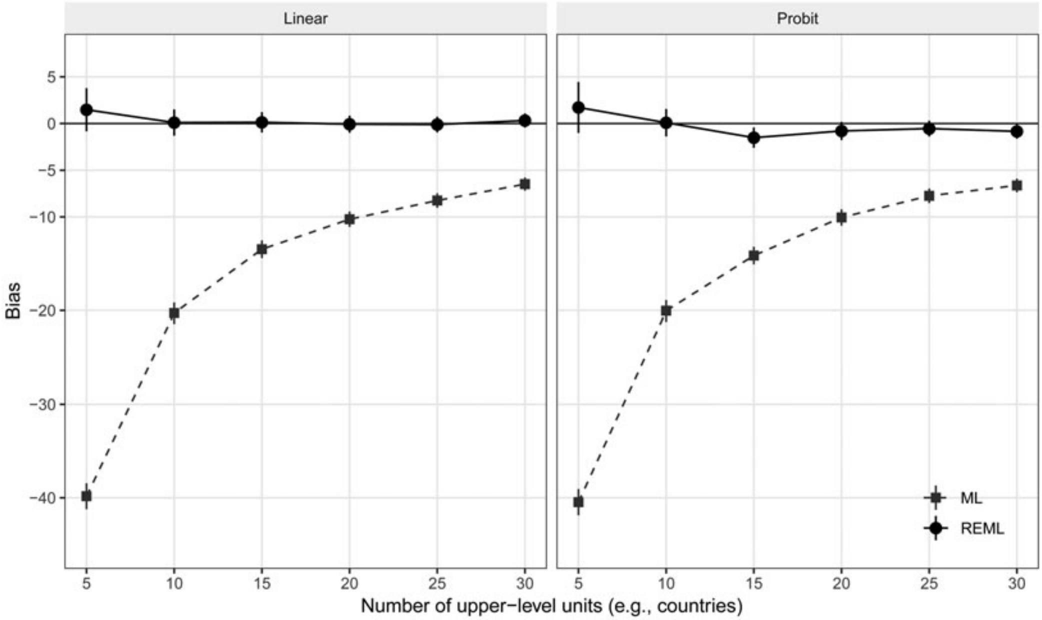

Figure 2. Performance of likelihood-based estimators of random intercept variances in multilevel linear and probit models
Note: The figure displays relative bias (in percent of the true parameter size) in variance estimates for the random intercept. Vertical lines depict 95 per cent confidence intervals. The horizontal zero line denotes the reference of no bias. The Monte Carlo sample size is 5,000. The confidence intervals are constructed analogously to those in Figure 1.

analysis, this modification uses 1,000 simulated data sets per condition). In addition, all simulated 'biases' are statistically insignificant and more moderately sized than in the original analysis.

In sum, the impact of the two modifications on the simulation results is fully consistent with, and thus substantiates, the theoretical result that ML point estimates are unbiased. Our analysis also illustrates how certain technical issues can crucially affect the validity of Monte Carlo evidence.

To illustrate how the inferential problems reported by Stegmueller and others (for example, Bryan and Jenkins 2016; Maas and Hox 2005) may be overcome, we re-examine his simulations using *R* because *Mplus* does not provide an implementation of the Satterthwaite approximation (for details on the *R* implementation, see Appendix B.1). In addition to using a different software package, we also specified independent random number seeds for the different experimental conditions and ran 5,000 rather than 1,000 replications per condition.[13] In all other respects, the following simulations are identical to Stegmueller's.

Figure 2 shows how well REML estimators can improve on ML estimators of variance parameters. Stegmueller did not report Monte Carlo results for the variance components, so these results have no correspondence in his article or the accompanying online appendix. We show these results here because they forcefully demonstrate the importance of using REML rather than ML estimation in few-cluster settings. Whereas ML estimates are increasingly biased as the number of upper-level units declines, REML generally achieves a tremendous degree of bias reduction. The extended quasi-likelihood implementation of REML-like estimation for probit models appears to be slightly less accurate than REML in the linear case, but also performs very well.[14] In any case, the potential inaccuracies are negligible compared to the expected downward bias of ML estimates.

[13] See note 1.

[14] We used the *R* package *hglm* by Rönnegård, Alam and Shen (2015) for estimation, which is based on work by Lee and Lee (2012); for details, see Appendices A.6 and B.1.

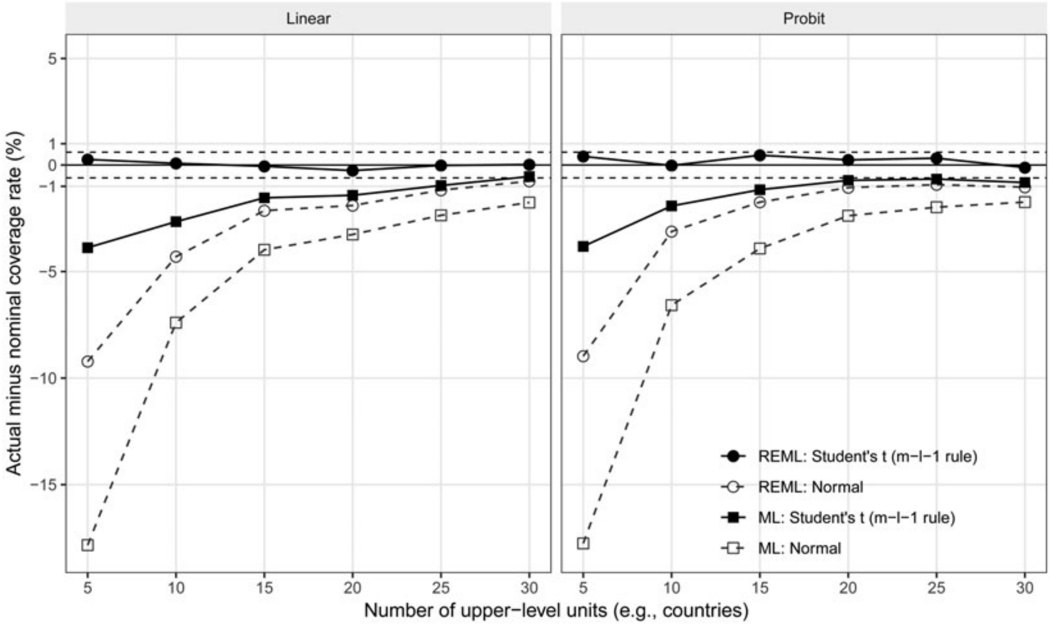

Figure 3. Performance of likelihood-based confidence intervals for upper-level covariate effect in multilevel linear and probit models

Note: The figure shows percentage point deviations of actual coverage rates from the nominal value of 95 per cent. The horizontal zero line denotes the reference of accurate coverage (i.e., actual equals nominal coverage rate) based on 5,000 Monte Carlo replications. The dashed horizontal lines indicate 95 per cent test intervals. For an accurate estimator of the 95 per cent confidence interval (i.e., one that has an actual coverage rate of 95 per cent), the estimated actual coverage rate should fall into the test interval 95 per cent of the time. In this sense, estimated coverage rates falling outside the test interval constitute statistically significant evidence against an accurate coverage rate. The test intervals are constructed to range from the 2.5 percentile to the 97.5 percentile (thus containing 95 per cent of the probability mass) of a binomial distribution with success probability p(1) = 0.95 and size parameter n = 5,000. This figure corresponds to the right-hand panel 'CI non-coverage' of Figure 2 on page 754 in Stegmueller (2013).

Figure 3 shows the extent of undercoverage for two-sided 95 per cent confidence intervals. The figure shows by how much (in percentage points) the actual coverage rate across the 5,000 replications (that is, the proportion of confidence intervals that include the true value of the parameter) differs from the nominal coverage rate of 95 per cent. The optimal value is zero, indicating that the actual coverage rate of the confidence intervals equals the nominal rate. To represent Monte Carlo uncertainty without sacrificing readability, we include dashed horizontal lines that indicate the limits of a two-sided 95 per cent confidence interval for a random variable that has a binomial distribution with success probability p(1) = 0.95 and size parameter n = 5,000. Non-coverage estimates that fall between the two dashed lines are not significantly different from zero at the 5 per cent level. As before, we focus on the linear case in the left panel, but the results are highly similar in the probit case (right panel).

The dashed line with hollow squares shows results based on ML estimation and the normal approximation; it nicely replicates Stegmueller's finding of confidence intervals that are much too narrow, particularly for small upper-level samples. If Stegmueller had used REML instead of ML, the extent of the problem would have been much smaller. The dashed line with hollow circles shows a maximum bias of about −9.2 percentage points rather than the −17.8 percentage points found for the combination of ML estimation with the normal distribution. Constructing confidence intervals from the appropriate t-distribution, however, seems even more important than using REML. The solid line with filled squares shows that the maximum bias declines to a mere −3.9 percentage points when we stick with ML estimation but construct confidence intervals based on a t-distribution with $m - l - 1$ degrees of freedom. Yet, the most important result of

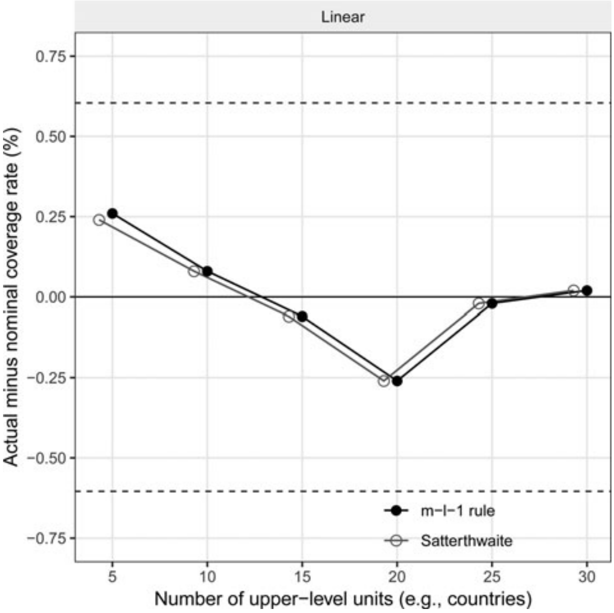

Figure 4. Performance of degrees of freedom approximations for the sampling distribution of test statistics in multilevel linear models
Note: The figure shows percentage point deviations of actual coverage rates from the nominal value of 95 per cent. The horizontal zero line denotes the reference of accurate coverage (i.e., actual equals nominal coverage rate) based on 5,000 Monte Carlo replications. The dashed horizontal lines indicate 95 per cent test intervals that are constructed in the same way as in Figure 3. This figure has no correspondence in Stegmueller (2013).

our analysis is that if we combine REML estimation with the $m - l - 1$ rule, inference is almost perfectly accurate. A few readily implementable choices can thus effectively address the inferential problems diagnosed by Stegmueller and others. Notably, the right panel in Figure 3 shows that the combination of (REML-like) EQL estimation with the $m - l - 1$ rule also leads to correct statistical inference in the multilevel probit case.

Figure 4 compares the performance of the $m - l - 1$ rule and the Satterthwaite approximation for the linear case.[15] Reassuringly, both methods perform very similarly. In all cases, the deviations of the actual from the nominal coverage rate of 95 per cent are close to the ideal value of zero and can reasonably be attributed to Monte Carlo error (note the much larger scale of Figure 4 compared to Figure 3). Appendix Figure B.4 shows that this result does not change if we use a more complex data-generating process that includes an additional lower-level predictor with substantial between-cluster variance. Our analysis thus suggests that both the $m - l - 1$ rule and the Satterthwaite approximation perform well in practice. In combination with REML, they can fully resolve the inferential deficiencies diagnosed by Stegmueller. That said, future research should investigate their performance under more complicated data-generating processes as encountered in applied research. As a first step in this direction, we now revisit the empirical example from Stegmueller's study.

An Empirical Application: Estimating The Determinants of Support for the European Union

As a first step towards evaluating our recommendations under more complex and realistic conditions, we follow Stegmueller and replicate Steenbergen and Jones' (2002) model of citizen support for the European Union (EU). Support for the EU is modelled as a function of a country's trade balance and tenure in the union, with GDP and monetary inflation as controls. For trade balance and inflation, Stegmueller finds that ML estimation leads to a rejection of the

[15] We are not aware of an *R* implementation of the Satterthwaite approximation for multilevel probit models. To our knowledge, *SAS* is currently the only major statistics package that provides the approximation for multilevel probit regression and other generalized linear multilevel models.

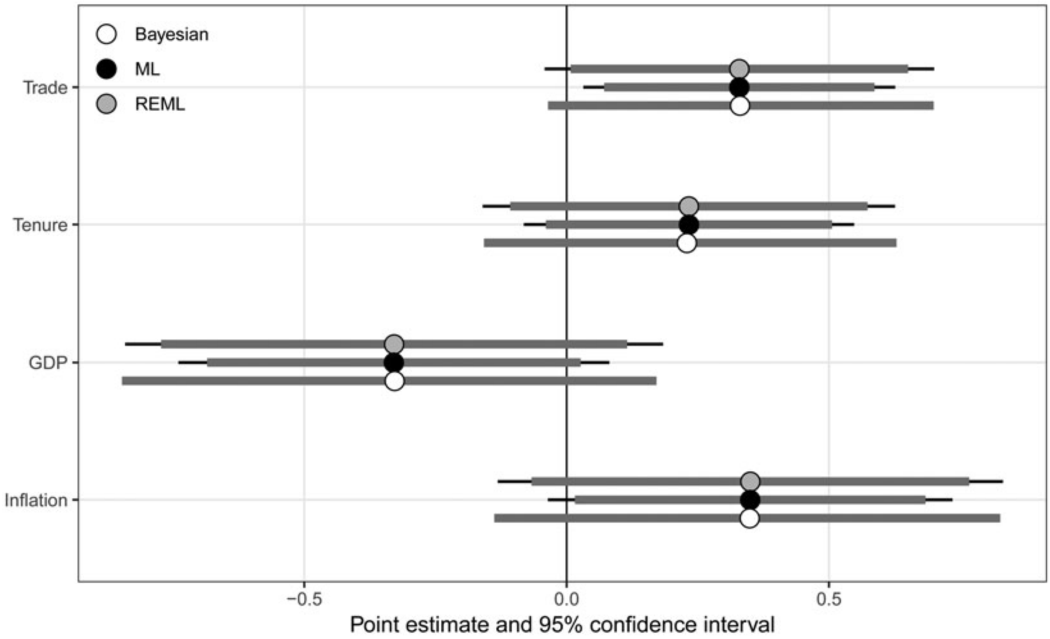

Figure 5. Country-level determinants of support for the European Union

Note: The figure displays point estimates and their 95 per cent confidence or credible intervals. For (restricted) ML estimates, thick lines represent 95 per cent confidence intervals based on the normal distribution, and thin antlers represent intervals based on the t-distribution with $m - l - 1$ degrees of freedom. Sample size: 10,777 individuals, 14 countries. This figure corresponds to Figure 8 on page 758 in Stegmueller (2013).

null hypothesis of no effect, whereas credible intervals based on a more demanding Bayesian Markov Chain Monte Carlo estimator include zero. The interpretation of this result in the context of Stegmueller's simulation evidence is that standard ML-based inference is anti-conservative and misleading. It leads to the rejection of null hypotheses that should have been retained and ultimately to 'different theoretical conclusions' (Stegmueller 2013, 757) than a Bayesian approach.

Figure 5 replicates Stegmueller's results and adds REML estimates as well as 95 per cent confidence intervals based on the normal distribution (thick lines) and on the t-distribution with $m - l - 1$ degrees of freedom (thin antlers). This practical example yields two main results. First, point estimates based on ML, REML and Bayesian MCMC are almost identical and only differ due to the randomness of MCMC. Secondly, our preferred method of assessing statistical uncertainty in the frequentist framework – using the REML estimator and the t-distribution with $m - l - 1$ (in this case: $14 - 4 - 1 = 9$) degrees of freedom – yields confidence intervals that are hardly distinguishable from the corresponding Bayesian credible intervals.

We take the results of this 'real-life example' as evidence that our recommendations will considerably improve the accuracy of frequentist inference in actual applications, which admittedly tend to be more complex than the stylized Monte Carlo DGPs that we borrowed from Stegmueller. Unlike in the Monte Carlo analysis, we cannot know if frequentist inference is accurate in the present setting, but the similarity to the Bayesian results is striking. Nevertheless, future research should investigate the performance of our recommendations using further Monte Carlo simulations with more complex setups.

Conclusions

A widely read and cited article by Stegmueller has raised serious concerns about the performance of standard likelihood-based methods of estimating multilevel models when the number of clusters is small. Stegmueller claims that these methods produce biased estimates of the coefficients of contextual variables, and that inferences about contextual effects may be strongly anti-conservative, potentially leading to an unjustified rejection of the null hypothesis of no effect. Especially with respect to statistical inference, several other studies have drawn similar conclusions (for example, Bryan and Jenkins 2016; Maas and Hox 2005).

In this article we have demonstrated that this pessimistic assessment of likelihood-based estimators of coefficients in multilevel models cannot be upheld. First, analytical results from the statistical literature indicate that ML estimates of context effects in linear multilevel models are unbiased – irrespective of the number of clusters and irrespective of whether ML or REML estimation is used. For generalized linear multilevel models such as multilevel probit, biases are possible when the *sizes* of the clusters are small, resulting in small lower-level samples. However, small lower-level sample sizes are rare in the country-comparative setting that motivated Stegmueller's analysis. Consistent with these assertions, our re-analysis of his Monte Carlo experiments provides no evidence of biased parameter estimates for either linear or generalized linear multilevel models.

Secondly, we have demonstrated that even with very few clusters, accurate inference for contextual effects is possible within the standard estimation framework, provided that two recommendations are followed. The first is to use REML estimation or a suitable extension for generalized linear multilevel models. The second is to approximate the distribution of the Wald test statistic for contextual effects using a *t*-distribution with the appropriate degrees of freedom rather than the standard normal distribution. Importantly, our results suggest that the appropriate degrees of freedom can be approximated quite easily, at least for simple hierarchical data structures. In our simulations, the $m-l-1$ rule – where m is the number of clusters and l is the number of predictors in the (implicit) upper-level regression – performed very well in both the linear and probit cases. In practice, validating the $m-l-1$ heuristic against the more computationally demanding Satterthwaite and Kenward-Roger approximations certainly will not hurt; we strongly recommend it for cross-classified and other non-hierarchical data structures. Yet if this is not feasible, our results suggest that the $m-l-1$ rule – combined with REML estimation and its extensions to generalized linear multilevel models – will go a long way towards achieving accurate inference. Taken together, these insights resolve lasting concerns of quantitative comparativists in political science and adjacent fields and have important implications for research practice.

In his study, Stegmueller recommends Bayesian MCMC estimation as a superior alternative to (frequentist) likelihood-based methods. Our results indicate that this claim of superiority is overstated. If applied in the right way, likelihood-based estimation performs on par with MCMC estimation in terms of both bias and statistical inference. Many applied researchers will likely consider this good news, as they have been trained primarily in the traditional paradigm and find MCMC procedures computationally costly and difficult to interpret. That likelihood-based and Bayesian methods, when implemented appropriately, yield very similar results for multilevel models with few but large clusters should not come as a surprise. While the two approaches have different underlying philosophies, they often lead to similar results if they are both applicable to the problem at hand. Both ML estimators and Bayes estimators are consistent for correctly specified models – that is, they approach the true parameter values as the sample size gets large.[16] For small samples, ML estimators are often biased,[17] but so are Bayes estimators.[18] Finally, likelihood-based methods can often be (re)interpreted in Bayesian terms.[19]

[16] For more on the consistency of ML estimators, see Casella and Berger (2002, 467ff). For details on the consistency of Bayes estimators, see Gelman et al. (2003, 106ff).

[17] See, e.g., Firth (1993) and Kosmidis and Firth (2009).

[18] See Casella and Berger (2002, 368) or Lehmann and Casella (2011, 234).

[19] For example, ML estimators can be seen as posterior modes with flat prior distributions. Bias reduction techniques, such as Firth's (1993) penalized ML estimator, can be seen as Bayes estimators with Jeffreys priors. An anonymous reviewer also suggested a Bayesian interpretation of restricted ML estimators. We discuss this interpretation in Appendix A.5.

While we have shown that certain readily available choices can considerably improve the accuracy of likelihood-based inferences, we nevertheless would like to emphasize that the problems identified by Stegmueller and others must be taken seriously. Many popular statistics packages default to the estimation approaches and statistical assumptions that these simulation studies (and ours) find to result in anti-conservative inference. For example, both *Stata* and *Mplus* rely on ML estimation and the normality assumption by default. We therefore suspect that many published multilevel analyses do indeed suffer from inaccurate statistical inference, although we cannot know for sure without detailed inspection of the individual studies.

Our analysis also carries three important lessons for the design of Monte Carlo studies. First, producers and readers should pay attention to Monte Carlo sampling error and communicate the uncertainty of their simulation results. Secondly, one should be wary of Monte Carlo studies with a small number of replications. Finally, repeatedly starting Monte Carlo simulations of different experimental conditions from the same random number seed may introduce interdependencies among experimental conditions and create seemingly systematic patterns, as it did in Stegmueller's finding of apparently biased coefficient estimates. We suspect that many other studies suffer from similar issues and hope that our analysis sensitizes both the producers and readers of Monte Carlo simulations to the importance of such technicalities.

An important remaining task is to study data-generating processes that are more complex and demanding than those investigated here. The results of Bryan and Jenkins (2016) suggest that REML falls short under more complex conditions, at least when it comes to the estimation of variance components, and one may also wonder how well degrees of freedom can be approximated in such situations. Again, it is encouraging that likelihood-based estimation, if conducted in accordance with our recommendations, performs similarly to a Bayesian MCMC estimator in the replication of Steenbergen and Jones' (2002) model of citizen support for the EU. Yet there is no guarantee that this good performance carries over to all of the diverse settings that are studied by social scientists.[20] Future simulation studies should therefore investigate more complicated data-generating processes.

Supplementary material. Data replication sets are available at https://doi.org/10.7910/DVN/CMMQRK and online appendices are available at https://doi.org/10.1017/S0007123419000097

Author ORCIDs. Martin Elff, 0000-0001-9032-9739; Jan Paul Heisig, 0000-0001-8228-1907; Merlin Schaeffer, 0000-0003-1969-8974

Acknowledgements. Earlier versions of this article were presented at the 2014 Conference of the European Political Science Association, the 2016 Conference of the Midwest Political Science Association, the 2016 Conference of the German Statistical Society (DStatG), the 2016 meeting of the Political Methodology Specialist Group of the Political Studies Association, the International Sociological Association RC 28 Spring Meeting 2017 and the 2017 Conference of the European Survey Research Association. We are grateful to Michael Peress, Sir David Hendry, and the members of the audience of the panels at the various conferences and seminars for their constructive comments and valuable feedback. We also thank the editor and two anonymous reviewers for their helpful comments. The authors are listed in alphabetical order; they all contributed equally to the article.

References

Breslow NE and Clayton DG (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88** (421), 9–25.

Breslow NE and Lin X (1995) Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82** (1), 81–91.

Bryan ML and Jenkins SP (2016) Multilevel modelling of country effects: a cautionary tale. *European Sociological Review* **32**(1), 3–22.

Casella G and Berger RL (2002) *Statistical Inference*, 2nd edn. Pacific Grove, CA: Duxbury.

Drum ML and McCullagh P (1993) REML estimation with exact covariance in the logistic mixed model. *Biometrics* **49** (3), 677–689.

[20] A minimum requirement of course is that the number of contextual variables is not too large, but satisfies $m - l - 1 > 1$.

Elff M et al. (2019) Replication data for: multilevel analysis with few clusters: improving likelihood-based methods to provide unbiased estimates and accurate inference. doi: 10.7910/DVN/CMMQRK, Harvard Dataverse, V1.

Fai HTA and Cornelius PL (1996) Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation* **54** (4), 363–378.

Firth D (1993) Bias reduction of maximum likelihood estimates. *Biometrika* **80** (1), 27–38.

Gelman A et al. (2003) *Bayesian Data Analysis*, 2nd edn. London: Chapman & Hall.

Gelman A and Hill J (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Giesbrecht FG and Burns JC (1985) Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results. *Biometrics* **41** (2), 477–486.

Heisig JP, Schaeffer M and Giesecke J (2017) The costs of simplicity: why multilevel models may benefit from accounting for cross-cluster differences in the effects of controls. *American Sociological Review* **82** (4), 796–827.

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** (13), 1249–1261.

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

Kosmidis I and Firth D (2009) Bias reduction in exponential family nonlinear models. *Biometrika* **96** (4), 793–804.

Kuznetsova A, Brockhoff PB and Christensen RHB (2017) lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* **82** (13), 1–26.

Lee W and Lee Y (2012) Modifications of REML algorithm for HGLMs. *Statistics and Computing* **22** (4), 959–966.

Lehmann EL and Casella G (2011) *Theory of Point Estimation*, 2nd edn. New York: Springer.

Lewis JB and Linzer DA (2005) Estimating regression models in which the dependent variable is based on estimates. *Political Analysis* **13** (4), 345–364.

Li P and Redden DT (2015) Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology* **15** (1), 281–296.

Liao JG and Lipsitz SR (2002) A type of restricted maximum likelihood estimator of variance components in generalised linear mixed models. *Biometrika* **89** (2), 401–409.

Lin X and Breslow NE (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* **91** (435), 1007–1016.

Maas CJ and Hox JJ (2005) Sufficient sample sizes for multilevel modeling. *Methodology* **1** (3), 86–92.

Meuleman B and Billiet J (2009) A Monte Carlo sample size study: how many countries are needed for accurate multilevel SEM? *Survey Research Methods* **3** (1), 45–58.

Moineddin R, Matheson FI and Glazier RH (2007) A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology* **7** (1), 34.

Noh M and Lee Y (2007) REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis* **98** (5), 896–915.

Patterson HD and Thompson R (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* **58** (3), 545–554.

Raudenbush SW and Bryk AS (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd edn. Thousand Oaks: Sage Pubn Inc.

Rönnegård L, Alam M and Shen X (2015) The HGLM package (version 2.0). Available from https://cran.r-project.org/package=hglm

Satterthwaite FE (1946) An approximate distribution of estimates of variance components. *Biometrics Bulletin* **2** (6), 110–114.

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), 512–524.

Snijders TAB and Bosker RJ (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*. London: Sage.

Steenbergen MR and Jones BS (2002) Modeling multilevel data structures. *American Journal of Political Science* **46** (1), 218–237.

Stegmueller D (2012) Replication data for: How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. hdl:1902.1/18628, Harvard Dataverse, V2.

Stegmueller D (2013) How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science* **57** (3), 748–761.

Zorn C (2005) A solution to separation in binary response models. *Political Analysis* **13** (2), 157–170.

Cite this article: Elff M, Heisig JP, Schaeffer M, Shikano S (2020). Multilevel Analysis with Few Clusters: Improving Likelihood-Based Methods to Provide Unbiased Estimates and Accurate Inference. *British Journal of Political Science* 1–15. https://doi.org/10.1017/S0007123419000097