No Need to Turn Bayesian in Multilevel Analysis with Few Clusters
Elff, Martin; Heisig, Jan Paul; Schaeffer, Merlin; Shikano, Susumu

DOI:
10.17605/OSF.IO/Z65S4

Publication date:
2016

Document version
Early version, also known as pre-print

Citation for published version (APA):
No Need to Turn Bayesian in Multilevel Analysis with Few Clusters: How Frequentist Methods Provide Unbiased Estimates and Accurate Inference

Martin Elff  
Zeppelin University

Jan Paul Heisig  
WZB – Social Science Center Berlin

Merlin Schaeffer  
University of Cologne

Susumu Shikano  
University of Konstanz

Abstract

Comparative political science has long worried about the performance of multilevel models when the number of upper-level units is small. Exacerbating these concerns, an influential Monte Carlo study by Stegmueller (2013) suggests that frequentist methods yield biased estimates and severely anti-conservative inference with small upper-level samples. Stegmueller recommends Bayesian techniques, which he claims to be superior in terms of both bias and inferential accuracy. In this paper, we reassess and refute these results. First, we formally prove that frequentist maximum likelihood estimators of coefficients are unbiased. The apparent bias found by Stegmueller is simply a manifestation of Monte Carlo Error. Second, we show how inferential problems can be overcome by using restricted maximum likelihood estimators for variance parameters and a t-distribution with appropriate degrees of freedom for statistical inference. Thus, accurate multilevel analysis is possible without turning to Bayesian methods, even if the number of upper-level units is small.

Acknowledgements:
Earlier versions of this paper were presented at the annual meetings of the European Political Science Association, of the Midwest Political Science Association, the German Statistical Society (DStatG), and the meeting of the Political Methodology Specialist Group of the Political Studies 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 feedbacks. The authors are listed in alphatical order; they all contributed equally to the paper.
1 Introduction

Multilevel modeling has emerged as a standard tool for the comparative analysis of political attitudes and behavior (Steenbergen and Jones, 2002). However, small samples of upper-level units are a frequent concern of comparative political scientists, particularly among those who are engaged in cross-national research and who are interested in contextual effects of country characteristics. In an influential paper published in the American Journal of Political Science, Stegmueller (2013) correspondingly asks the important question: “How many countries for multilevel modeling?”1 He presents simulation evidence suggesting that mixed effects multilevel models estimated by frequentist methods yield biased point estimates and lead to dramatically anti-conservative inference for contextual effects when the number of countries (or, more generally, of clusters or upper-level units) is small. That is, when the number of clusters is below 20, point estimates of context effects seem to systematically misrepresent true effect sizes, and actual coverage rates of confidence intervals are far below the nominal level, implying downward biased $p$-values and overrejection of the null hypothesis of no effect.

With regard to inferential deficiencies, a more recent simulation study by Bryan and Jenkins (2016) reaches similar conclusions, as do simulation studies focusing on multilevel structural equation models (Meuleman and Billiet, 2009; Hox, van de Schoot, Matthijssse et al., 2012). A decade ago, Maas and Hox’s (2005; 2004) and Bell, Ferron and Kromrey’s (2008) simulation studies also cautioned that mixed effects models require a reasonable number of clusters for accurate inference. Thus, the question whether quantitative multilevel analysis requires a certain minimum number of upper-level units has long haunted (comparative) political science and the social sciences more generally.

Stegmueller (2013) goes beyond previous studies by also suggesting a solution to the problems bedeviling multilevel analysis. He finds that Bayesian point estimates of contextual effects tend to be much closer to the true parameter values. Moreover, Bayesian credible intervals appear to provide relatively accurate inference even when the number of clusters is small. Stegmueller (2013) concludes that “for researchers in comparative politics (and adjacent ‘comparative’ fields) who are interested in multilevel analyses, turning to a Bayesian approach might be a fruitful choice” (p. 759).

We refute these strong claims about the superiority of Bayesian over frequentist techniques of estimation. We show that even if based on as few as five clusters, frequentist methods

---

1As of December 10, 2016, Google Scholar recorded 218 citations of Stegmueller’s article, making it the most-cited item to have appeared in AJPS in 2013.
provide unbiased estimates of contextual effects and that accurate statistical inference for these effects is possible. Thus, we not only counter Stegmueller’s (2013) influential claims, but resolve a long-lasting concern of quantitative comparative political research.

In the first part of the paper, we address the alleged bias of frequentist parameter estimates. We summarize statistical theory to prove that unbiased point estimates of coefficient estimates in mixed effects models can be obtained using maximum likelihood techniques. We further show that the apparent bias of frequentist parameter estimates that shows up in Stegmueller’s (2013) simulation results is not due to inherent bias, but instead is attributable to Monte Carlo sampling error. The basic idea of a Monte Carlo simulation is to create a large number of artificial data sets by sampling repeatedly from a known data-generating process. By applying estimators of interest to the many simulated data sets and observing their performance, one can learn a lot about their properties even when these cannot be assessed analytically. Because Monte Carlo simulations involve sampling from the true data-generating process, Monte Carlo results are subject to error, just like conventional parameter estimates. They should therefore be accompanied by estimates of their reliability and the number of simulated data sets should be large enough to minimize the probability that findings are due to chance. We provide evidence that Stegmueller’s finding of parameter bias is due to an insufficient number of simulated data sets, in combination with the accidental choice of a random number seed that produces extreme results.²

In the second part of the paper, we address the alleged inferential problems diagnosed by Stegmueller (2013) and others, and show that they arise from two mistakes. First, Stegmueller (2013) considers only the maximum likelihood (ML) estimator, rather than the more appropriate restricted maximum likelihood (REML) estimator. It is well-known that ML estimates of variance components of mixed effects models (i.e., of random intercept and slopes variances) are severely biased for small samples of upper-level units. This is exactly why Patterson and Thompson (1971) developed REML as a remedy. Second, Stegmueller (2013) uses the standard normal (Gaussian) distribution to compute confidence limits and p-values. But inference about contextual effects is effectively based on a small (upper-level) sample even when lower-level samples are large. As with conventional linear regression,

²Put simply, a random number seed is an arbitrary number that defines a sequence of pseudo-random draws generated by a computer program, thereby ensuring the reproducibility of results that depend on a random component. Stegmueller (2013) repeatedly used the seed 12345 to start Monte Carlo sampling for each experimental condition studied in his paper. While the choice of this random number seed was clearly not intentional, our analysis suggests that it happens to result in a sequence of simulated data sets that create an exceptionally strong impression of bias of frequentist multilevel estimators. Moreover, repeatedly using the same seed throughout the paper leads to a lack of independence among experimental conditions that creates a misleading pattern of seemingly systematic bias.
for small samples one needs to construct $p$-values and confidence intervals from a (heavier-tailed) $t$-distribution with the appropriate degrees of freedom. Small-sample inference for mixed effects models is complicated by the fact that the true degrees of freedom are unknown (Schaalje, McBride, and Fellingham, 2001), but promising approximations are available. We consider two approaches, both of which we find to perform well: the $m - l - 1$ rule and the Satterthwaite approximation.

To assess the improvement in inferential accuracy from using both types of small-sample adjustments (REML and using the $t$-distribution with approximately correct degrees of freedom), we replicate Stegmueller’s (2013) Monte-Carlo simulations. We find that frequentist inference for both direct context effects and cross-level interactions is perfectly accurate if estimated by restricted maximum likelihood and based on the $t$-distribution with degrees of freedom approximated by the simple $m - l - 1$ rule or the more computationally demanding Satterthwaite method. Even for samples of only five clusters, frequentist inference is accurate in the set-up considered by Stegmueller (2013). Thus, while there may be other compelling reasons to favor Bayesian estimation, it is not necessary for achieving accurate inference for multilevel analyses.

2 Maximum Likelihood Estimates of Coefficients in Multilevel Models are Unbiased

Stegmueller (2013) claims that maximum likelihood (ML) point estimates of context effects in linear mixed effects multilevel models are biased. In this section we reassess this conclusion. We first discuss statistical theory that proves the unbiasedness of ML point estimates. The full analytical proof is quite technical and requires complex mathematical notation. We therefore present it in the supporting material. Here, we give a more intuitive and accessible summary of the main ideas. In a second step, we reconcile the analytical argument with Stegmueller’s (2013) results by demonstrating that the apparent bias of frequentist point estimates in his Monte Carlo simulations is merely a consequence of Monte Carlo error.

2.1 Why Maximum Likelihood Estimates of Coefficients in Multilevel Models are Unbiased

Quantitative scholars are often dealing with multilevel data that contain large samples at the lower level, but only a limited number of cases at the upper level. Such data can be de-
scribed as “large \(n\), small \(m\) data”, with \(n\) indicating the number of lower-level observations and \(m\) the number of clusters. A typical example is a cross-national survey of individuals nested in countries. Often, the primary goal in analyzing such data is to identify context effects, that is, to investigate how contextual variables that vary only at the cluster level affect lower-level outcomes or how they moderate relations between lower-level variables (Bryan and Jenkins, 2016; Stegmueller, 2013). GDP per capita, electoral system, or trade openness are typical examples of upper-level variables.

Several distinct approaches exist for explicitly analyzing multilevel data structures (for an overview see Kedar, 2005). In the social sciences, the most widely-used approach is to use mixed effects multilevel models (Steenbergen and Jones, 2002), sometimes also referred to as hierarchical (linear) models (Gelman and Hill, 2006).

Let us fix ideas by introducing the multiple-equations formulation of a simple multilevel model that is similar to the ones considered by Stegmueller (2013). Given a continuous outcome variable \(y\) that varies across individuals \(i\) \((i = 1, \ldots, n)\) and clusters \(j\) \((j = 1, \ldots, m)\) and depends on lower-level variable \(x\), context-level characteristic \(w\), and a “cross-level” interaction between \(w\) and \(x\), the multilevel model can be formalized as follows:

\[
\begin{align*}
    y_{ij} &= a_{0j} + a_{1j}x_{ij} + \epsilon_{ij} \quad (1) \\
    a_{0j} &= \alpha_{00} + \alpha_{01}w_{j} + b_{0j} \quad (2) \\
    a_{1j} &= \alpha_{10} + \alpha_{11}w_{j} + b_{1j} \quad (3)
\end{align*}
\]

\(\epsilon_{ij}\) is a lower-level error term, whereas \(b_{0j}\) and \(b_{1j}\) are upper-level errors for \(a_{0j}\) (the intercept) and \(a_{1j}\) (the coefficient of the lower-level predictor \(x\)), respectively. In other words, the model assumes that the intercept \(a_{0j}\) and the coefficient of lower-level predictor \(a_{1j}\) vary over the \(m\) clusters and that the cluster characteristic \(w_{j}\) accounts for some, but not all, of this variation (i.e., that there are upper-level errors).

While this hierarchical multiple-equations form is very intelligible from a substantive perspective, the expanded form is better suited to explain estimation and inference. Substituting equations (2) and (3) into equation (1) leads to the following expanded form of the model:

\[
y_{ij} = \alpha_{00} + \alpha_{10}x_{ij} + \alpha_{01}w_{j} + \alpha_{11}x_{ij}w_{j} + b_{0j} + x_{ij}b_{1j} + \epsilon_{ij}. \quad (4)
\]

In the terminology of multilevel modeling one often calls \(\alpha_{00}\), \(\alpha_{10}\), \(\alpha_{01}\), and \(\alpha_{11}\) “fixed
effects” and \( b_{0j} \) and \( b_{1j} \) “random effects”. In addition, one can refer to \( \alpha_{11} \) as a “cross-level interaction” between \( x_{ij} \) and \( w_j \), to \( b_{0j} \) as a “random intercept”, and to \( b_{1j} \) as a “random slope” term. Parallel to simple linear regression, the expected value of all error terms (i.e., of the \( b \)’s and \( \epsilon_{ij} \)) is zero. The name “mixed effects” (multilevel) model stems from the fact that these models combine fixed and random effects.

It is also common to express linear mixed effects multilevel models using matrix algebra, which is particularly useful for analyzing estimation and inference. A general notation is given by:

\[
y = X\alpha + Zb + \epsilon
\]

where \( X \) is a matrix containing the values of lower- and upper-level predictors as well as their interactions and \( \alpha \) is a vector of associated fixed effects. That is, in the translation of equation (4) into equation (5) a row of \( X \) consists of 1, \( x_{ij} \), \( w_j \) and \( x_{ij}w_j \), while \( \alpha \) consists of \( \alpha_{00} \), \( \alpha_{01} \), \( \alpha_{10} \), and \( \alpha_{00} \). \( Z \) is a matrix that assigns the random (intercept and slope) effects \( b \) of the clusters to the respective observations nested within them. Thus if individual \( i \) is in cluster \( j \), the corresponding row of \( Z \) consists of zeros, except a one at the \( j \)-th place and \( x_{ij} \) at the \( m + j \)-th place. Because the similarities of observations from the same cluster are explicitly modeled as random effects, the assumption that \( \epsilon \) is a vector of classical independent and identically distributed (i.i.d.) lower-level errors is plausible.

If \( \Phi \), the covariance matrix of \( b \), and \( \sigma^2 \), the variance of the elements of \( \epsilon \), are known, the variance of \( y - X\alpha = Zb + \epsilon \) is also known as \( V = \sigma^2I + Z\Phi Z' \). Aitken’s Theorem (Aitken, 1936; Greene, 2012) states that in this case the GLS estimator

\[
\hat{\alpha}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}y
\]

is unbiased and is even the best linear unbiased estimator for \( \alpha \).

Yet \( \sigma^2 \) and \( \Phi \) are generally unknown and have to be estimated. Nevertheless if \( \sigma^2 \) and \( \Phi \) are estimated by maximum likelihood, then a maximum likelihood estimate of \( \alpha \) is obtained by substituting \( \hat{V} \) for \( V \) in equation 6, where \( \hat{V} \) is computed from the ML estimates \( \hat{\sigma}^2 \) and \( \hat{\Phi} \). While some of the conditions of Aitken’s theorem do no longer hold, the ML estimate \( \hat{\alpha}_{ML} \) is still unbiased, that is, it equals the true population parameter in expectation (Kackar and Harville, 1981).
To show this, we substitute equation 5 into $y$ in the expected value of equation 6:

$$
E[\hat{\alpha}] = E[(X'V^{-1}X)^{-1}X'V^{-1}(X\alpha + Zb + \epsilon)]
$$

$$
= E[(X'V^{-1}X)^{-1}X'V^{-1}X\alpha] + E[(X'V^{-1}X)^{-1}X'V^{-1}(Zb + \epsilon)]
$$

$$
= \alpha + E[(X'V^{-1}X)^{-1}X'V^{-1}(Zb + \epsilon)]
$$

(7)

For a fixed matrix $V$, the expected value of the remainder term in the last line is obviously zero, because the expected values of both the cluster and the lower-level error terms are zero by assumption. As we show in Appendix B, the remainder term has zero expectation also when $V$ is not fixed but a symmetric function of $b$ and $\epsilon$, which is the case if it is based on a ML estimate of $\Phi$ (Kackar and Harville, 1981). This means that the expected value of our maximum likelihood estimate $\hat{\alpha}$ is the true population parameter vector, that is, it is unbiased, irrespective of whether the ML estimates of variance parameters are biased or not.

Contrary to Stegmueller’s (2013) interpretation, an inherent bias of frequentist (ML) point estimates therefore cannot be the reason why the average estimated context effects differ from the true parameter values in his simulation study. And, indeed, it is not. As we now demonstrate, Stegmueller’s seemingly alarming findings are simply due to Monte Carlo error.

2.2 Pitfalls of Monte Carlo Studies

We now show that Stegmueller’s (2013) Monte Carlo results do not contradict the unbiasedness of ML point estimates in linear mixed effects multilevel models. To follow our argument, it is essential to have a basic understanding of Monte Carlo simulations (for a more extensive introduction, see Carsey and Harden, 2013). The central idea of Monte Carlo methods is to simulate the hypothesized repeated sampling of statistical theory. More specifically, one repeatedly generates artificial data by sampling from a known “data-generating process” (DGP), creating a large number of simulated data sets. To each of these data sets one then applies an estimator (or a set of estimators) of interest.

The performance of an estimator over a large number of data sets (or “replications”) allows to assess its statistical properties under controlled conditions. Important statistical properties such as unbiasedness or efficiency have natural Monte Carlo analogues. For example, an unbiased estimator yields point estimates whose average value approaches the true parameter value as the number of replications gets large (the true parameter value is
known because the DGP is specified by the researcher). By examining different “experimental conditions”, that is, by varying certain aspects of the DGP, one can better understand how key features of actual applications (e.g., the number of clusters) affect the performance of an estimator.

Simulated data sets are created by sampling from a known data-generating process with random components. For this reason, Monte Carlo results are subject to sampling error, much like conventional parameter estimates. It is easy to see that it would be premature to categorize an estimator as biased just because the average parameter estimate across a small number of replications (say, 10 or 20) happens to differ from the true value. By the same token, one should be a lot more confident about this conclusion if a substantial deviation remained after simulating several thousands of data sets. More generally, it is possible and important to quantify the statistical uncertainty of Monte Carlo estimates, or to ensure that the number of replications is large enough to render this issue irrelevant.

The fact that Monte Carlo simulations necessarily have a random component implies that the exact results depend to some extent on the sequence of random—or, more precisely, pseudo-random—draws made by the computer program used to implement the simulation. It is therefore good practice to report the so-called “random number seed”, an arbitrary number ensuring that repeated runs of the same simulation exercise are based on the same sequence of pseudo-random draws and thus ultimately yield identical results. Different random number seeds will lead to (slightly) different results, but if the total number of replications is large enough this issue should be negligible. In a sense, measures of uncertainty for Monte Carlo estimates quantify the extent to which results are contingent on the arbitrary choice of a particular random number seed.

These considerations suggest that Stegmueller’s (2013) Monte Carlo analysis suffers from at least three flaws. First, his results are based on a relatively small number of 1,000 Monte Carlo trials, which leaves ample room for findings being simply due to chance. Second, and related, this possibility is exacerbated by the fact that Stegmueller does not quantify the uncertainty of his simulations, for example, by reporting Monte Carlo confidence intervals. Third, Stegmueller (2013) repeatedly uses the same random number seed (the five-digit number 12345) for each experimental condition. Our reanalysis suggests that this seed happens to produce a sequence of simulated data sets that create a rather strong impression of bias of frequentist estimators of multilevel models. It also leads to a lack of

---

3The only reference to Monte Carlo error in his article occurs in footnote 14 on page 752, which states the size of the simulation error for the coverage rates of confidence intervals. Monte Carlo error is not mentioned in relation to parameter bias.
independence among experimental conditions that creates a misleading pattern of seemingly systematic bias.

### 2.3 Monte Carlo Evidence on Bias of Coefficient Estimates

To illuminate the issues discussed in the previous subsection, we use Stegmueller’s (2013) publicly available replication files which allow everyone to exactly replicate his results using a free trial version of MPlus. Like him, we focus on the case where the intraclass correlation (i.e., the portion of the overall variance in the outcome variable that is between clusters) equals 0.10. In addition to exactly replicating his results, we also replicate them under two modifications that should not qualitatively change the results if they were as systematic as he claims, but alter them in foreseeable ways if the original findings were just a manifestation of Monte Carlo error. Our first modification increases the number of Monte Carlo replications by a factor of ten, from 1,000 to 10,000 replications, leaving everything else as it is. Thus, we only reduce the magnitude of Monte Carlo error, which we report using 95% confidence intervals. If, as we claim, frequentist ML point estimates are unbiased, we would expect the increase in the number of replications to shrink the estimated bias toward zero (= no bias) relative to Stegmueller’s results. Our second modification implements different (five-digit) random number seeds for each experimental condition, and again leaves everything else as it is, including the original number of 1,000 Monte Carlo trials. With this modification, the results should no longer suggest systematic bias across the different experimental conditions.

Figure 1 visualizes the results in the same fashion as Stegmueller (2013), apart from additionally including 95% confidence intervals that indicate Monte Carlo sampling variability. The black line with triangles shows the exact replication of Stegmueller (2013). The 95% confidence intervals show that the original estimates are subject to a lot of uncertainty. That said, they also imply that the result of positively biased point estimates of models based on cluster-samples smaller than 20 (or 10 in case of mixed effects probit models) is significant at the 5% level. This stands in some tension to our claim that ML point estimates are unbiased and that Stegmueller’s (2013) results are simply a consequence of Monte Carlo error. A look at Figure A1 in the Appendix shows, however, that Stegmueller’s (2013) original results were never significant for any of the three constitutive terms of the cross-level interaction (Figures 3, 4 and 5 of his article). This includes those estimates that he concluded to be the most biased.

---

4We used the replication files available on the AJPS dataverse (Stegmueller, 2012).
Figure 1: Performance of Point Estimates of Upper-Level Covariate Effect in Hierarchical Linear and Probit Models

<table>
<thead>
<tr>
<th>Linear Model</th>
<th>Probit Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
</tr>
</tbody>
</table>

Note: The figure displays relative biases of frequentist maximum likelihood point estimates (in % of the true effect size). Vertical lines depict 95% Monte Carlo confidence intervals for these results. The horizontal zero line denotes the reference of no bias. Black triangles replicate the results presented 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 increased from 1,000 to 10,000, leaving everything else as is. The second (gray squares) follows Stegmueller in using only 1,000 replications, but specifies different random number seeds for the different experimental conditions, whereas Stegmueller used the seed 12345 throughout.
This reminds us that every statistically significant finding is potentially subject to a type-I error, in our case, the error of falsely reject the null hypothesis of no bias in point estimates. Our two modifications of Stegmueller’s simulations (2013) provide strong evidence that this is indeed what is going on here. The dark-gray circled line shows results based on 10,000 instead of only 1,000 replications (but leaving everything else, including the random number seed, as it is). In line with our claims, these results are not only much more precise, as indicated by the much narrower confidence bands; the estimated magnitude of the parameter bias is also consistently shrunk towards zero. There remains no statistically significant evidence of bias for mixed effects probit models. For linear mixed effects models, results for 5 and 15 (but not 10) clusters remain statistically significant. The estimated size of relative parameter bias declines considerably, however, from roughly 10.25 and 5.7 to 4.15 and 1.85 percent of the true effect size, respectively.

What remains is the result that the simulated bias of frequentist ML point estimates is consistently positive, even though it declines as the number of clusters increases. As expected, our second modification swipes this pattern away. The gray squared line shows that when we use a different random number seed\(^5\) for each experimental condition, any pattern of systematic positive bias disappears. The estimated biases are all statistically insignificant, more moderately sized than in Stegmueller’s (2013) original analysis, and inconsistently signed. This is just what one would expect to find if our claim that ML point estimates are unbiased is correct.

In conclusion, this reanalysis and extension of Stegmueller’s (2013) simulation analysis corroborates our analytical argument that concerns about the biasedness of frequentist ML point estimates are unfounded. It also highlights three crucial pitfalls that have led to misleading conclusions: neglect of Monte Carlo error, a too small number of Monte Carlo replications, and the repeated use of the same random seed across all experimental conditions. Thus, our reanalysis also conveys important general advice for producers and readers of Monte Carlo simulation studies.

3 Accurate Inference for Mixed Effects Models

So far, we have demonstrated that maximum likelihood point estimates of linear and probit mixed effects models are unbiased. But Stegmueller’s (2013) claims also concern inference. In fact, several studies warn that frequentist estimation results in downward biased \(p\)-values.

\(^{5}\)We generated these random number seeds using the website www.random.org.
and overrejection of the null hypothesis of no effect when upper-level samples are small (Bryan and Jenkins, 2016; Meuleman and Billiet, 2009; Hox, van de Schoot, Matthijsse et al., 2012; Maas and Hox, 2005, 2004; Bell, Ferron, and Kromrey, 2008).

In this part of the paper, we first show that inference based on maximum likelihood (ML) is indeed inaccurate because the variance components of mixed effects models (i.e., the variances of random intercept and slopes) are underestimated for small upper-level samples. But we also elaborate that restricted maximum likelihood (REML) estimators provide a remedy and illustrate how it strongly tempers Stegmueller’s (2013) alarming findings.

In a further step, we identify a second reason why Stegmueller finds frequentist techniques to yield anti-conservative inferences: the use of the standard normal (Gaussian) distribution to derive confidence limits and $p$-values. If the number of upper-level units is small, estimates of context effects are effectively based on a small sample even when lower-level samples are large. Approximating their sampling distribution by an asymptotic normal distribution is inadequate, even though it is the default or even the only readily available option in some widely-used statistics packages such as Stata (until version 13) and MPlus (the software used in Stegmueller’s 2013 simulations). We show that inference can be dramatically improved by using a (heavier-tailed) $t$-distribution with the appropriate degrees of freedom.

### 3.1 Bias-Corrected Variance Estimates by Restricted Maximum Likelihood

As already noted, ML estimates of fixed effect coefficients in multilevel models are unbiased, even if the covariance matrix of the random effects is unknown and is estimated by ML as well. By contrast, the ML estimate of the covariance matrix of the random effects covariance matrix is biased. This is relevant for inference about fixed effects coefficients since this covariance matrix forms the base of constructing standard errors, and thus also of confidence intervals and $p$-values. Usually standard errors are obtained from the square roots of the diagonal of the asymptotic covariance matrix of the coefficient ML estimates:

$$\widehat{\text{Var}}(\hat{\alpha}_{\text{ML}}) = (X'\hat{V}^{-1}X)^{-1} = \left(X' \left[ \hat{\sigma}^2 I + Z\hat{\Phi}Z' \right]^{-1} X \right)^{-1}$$  \hspace{1cm} (8)

As a consequence, if there is a downward bias in the ML estimates $\hat{\sigma}^2$ and $\hat{\Phi}$, then the estimated standard errors of $\hat{\alpha}_{\text{ML}}$ are also biased downward.

A bias in variance estimators is not peculiar to multilevel modeling. Consider a sample
of normally distributed random variables. The ML estimate of the common variance is

\[ \hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2, \]

which is well-known to have the following expected value and bias

\[ E(\hat{\sigma}_{ML}^2) = \frac{n - 1}{n} \sigma^2, \quad \text{Bias}(\hat{\sigma}_{ML}^2) = -\frac{1}{n} \sigma^2. \]

It is obvious that this bias tends to zero as \( n \) grows large. Nevertheless, for small samples one may want to obtain a bias-corrected estimator of the variance, which is simply:

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

For ML estimates of variance parameters of multilevel models, there is also a bias-corrected version available, which was introduced by Patterson and Thompson (1971). In contrast to the case just discussed, this bias correction is not available in closed form, but is based on maximizing a modified likelihood function. Patterson and Johnson call the bias corrected estimates thus restricted maximum likelihood (REML) estimator. Since the mathematical details of REML estimators are demanding, we elaborate them in Appendix C, but we note here, that the bias-corrected sample variance as well as the bias-corrected estimator of the residual variance in linear regression can be considered as special cases of REML. REML estimators differ from ML estimators only in the way to compute variance parameters. They can be viewed as a generalization of bias-corrected estimator of the error variance in linear regression. Thus, it seems reasonable to expect that if variance parameters of multilevel models are estimated by REML instead of ML, the coverage error of interval estimates of fixed-effects coefficients found by Stegmueller (2013) can be considerably reduced.

### 3.2 Monte Carlo Evidence on Bias Correction by REML Estimators

To illustrate the performance of REML estimators, we again replicate Stegmueller’s (2013) simulations. However, because MPlus does not provide for the improved techniques dis-

\[ ^6 \text{With “sample” we refer to the usual concept of mathematical statistics of a sequence of identically and independently distributed random variables.} \]
cussed in this paper, we conduct this part of the re-analysis using R. We describe the details of our Monte Carlo design in Appendix D. It is basically identical to Stegmueller’s (2013) set up, with two slight modifications. First, we use R’s built-in random number generator and different random number seeds for different experimental conditions. Therefore we cannot exactly replicate Stegmueller’s findings, although we do arrive at qualitatively similar results. Second, we use 5,000 instead of 1,000 Monte Carlo replications to minimize the impact of Monte Carlo error.

Figure 2 shows bias in the estimated variances of random intercept terms for both ML and REML estimators. These results have no correspondence in Stegmueller (2013) where bias in variance estimates was not examined. We show results for the linear (left column) and probit (right column) cases. We also vary the size of the intraclass correlation from 0.05 (top row), to 0.10 (middle row), to 0.15 (bottom row).

For the ML-based random effects we see exactly the downward bias discussed above. The bias is largest in the case of only five clusters where the estimated variance of the random intercept is approximately 40% too small. As the number of clusters increases, the bias decreases steadily, but remains substantial even in the 30 cluster case. By contrast, the REML estimator produces unbiased variance estimates, even for the case of only five clusters. Figure A2 in the appendix shows similar results for the two random effects (i.e., random intercept and slope) of a model containing a cross-level interaction. In the following section, we illustrate how much of the inferential bias reported in Stegmueller (2013) can be attributed to the difference between the ML and REML estimators. We also explain and demonstrate how the remaining bias can be avoided by constructing p-values and confidence intervals from the t-distribution with approximately correct degrees of freedom. Since Figure 2 does not suggest that results depend on the size of the intraclass correlation, we focus on the case where the latter equals 0.10.

3.3 t-Distribution and Degrees of Freedom Approximation

No matter how large the number of lower-level observations is, data with small m (i.e., few clusters) provide only a limited number of cases for identifying the effects of upper-level variables. One well-known consequence is that the effects of upper-level variables will often be estimated with considerable statistical uncertainty, that is, with large standard errors. A second implication, which seems to be less widely recognized among practitioners, is that it

\footnote{For example, the support for REML is limited and a t-distribution cannot be used for the construction of confidence intervals.}
Figure 2: Performance of Variance Estimates of Random Intercepts in Hierarchical Linear and Probit Models

Note: The figure displays relative bias (in % of true size) in variance estimates for the random intercept, estimated using maximum likelihood (ML; gray squares) and restricted maximum likelihood (REML; black circles). Vertical lines depict 95% confidence intervals. The horizontal zero line denotes the reference of no bias. This figure has no correspondence in Stegmueller (2013) where bias in variance estimates is not examined.
is problematic to base statistical inference about the effects of upper-level variables on large-sample theory. In particular, it is very misleading to construct \( p \)-values, hypothesis tests, and confidence intervals for the coefficients of upper-level variables using the assumption that the corresponding \( t \)-statistic (i.e., the coefficient estimate divided by its standard error) follows the standard normal distribution. Unfortunately, this is the default and sometimes even the only readily available option in many statistics packages, among them \textit{Stata} (until version 13) and \textit{MPlus}. However, with small upper-level samples, inference for context effects should be based on the heavier-tailed \( t \)-distribution (Student, 1908) with appropriate degrees of freedom. This is not an arcane statistical point: as we demonstrate below, using a more appropriate \( t \)-distribution can essentially resolve the inferential problems reported by Stegmueller (2013)—and we suspect that it would also do so for other studies that report similar problems.

Using the \( t \)-distribution for inference about upper-level effects is not straightforward due to the fact that the appropriate degrees of freedom are unknown in the context of mixed effects models. Statisticians have proposed several approximations to the effective degrees of freedom over the past decades (c.f., Schaalje, McBride, and Fellingham, 2002; Manor and Zucker, 2004; Li and Redden, 2015). Some of these are clearly inadequate for inference about direct context effects and cross-level interactions, because they will typically result in degrees of freedom that far exceed the number of clusters. We therefore only consider the two most promising approaches for approximating the degrees of freedom for \( t \)-tests of context effects in mixed effects multilevel models: the \( m - l - 1 \) rule and the Satterthwaite approximation which estimates the degrees of freedom from the data.

Like Stegmueller (2013), we focus on inference for single parameters, that is, on inference about the coefficients of (upper-level) variables. We therefore do not include the so-called Kenward-Roger approximation (Kenward and Roger, 1997) in our simulations, a computationally more intensive alternative that is closely related to the Satterthwaite approximation and yields almost identical results in the single-parameter (or, more generally, single-constraint) case (Li and Redden, 2015). Practitioners interested in jointly testing multiple constraints (e.g., the hypothesis that all coefficients except for the intercept are zero) should consider the Kenward-Roger approximation and we hope that our contribution stimulates research on their relative performance in the kinds of “large \( n \), small \( m \)” settings that are typical of comparative politics and adjacent disciplines.
3.3.1 The $m - l - 1$ Rule

At least three textbooks on multilevel mixed effects models mention the $m - l - 1$ approximation to the degrees of freedom for testing context effects (Pinheiro and Bates 2000, 91-92; Raudenbush and Bryk 2002, 57-58; Snijders and Bosker 2012, 94-95). As before, $m$ denotes the number of clusters, $l$ the number of contextual effects, and 1 is added for the intercept. Apart from being implemented as default in the software package HLM (Raudenbush, Bryk, and Congdon, 2004), the approximation seems to be rarely used in practice. The rationale behind the rule is easiest to grasp in the multiple equations formulation of a two-level model as shown in equations 2 and 3 above.

A straightforward justification of the $m - l - 1$ rule is that $m - l - 1$ would be the correct degrees of freedom for the single-parameter tests of $H_0 : \alpha_{01} = 0$ vs. $H_A : \alpha_{01} \neq 0$ and of $H_0 : \alpha_{11} = 0$ vs. $H_A : \alpha_{11} \neq 0$, if equations 2 and 3 were fit separately using a conventional (upper-level) regression model, as is actually done in so-called two-step estimation (c.f., Lewis and Linzer, 2005). The same applies to $\alpha_{00}$ (the overall intercept) and $\alpha_{10}$ (the main effect of $x_{ij}$).

The implication to use $m - l - 1$ degrees of freedom may be less obvious for the parameters in equation 3, that is, for the main effect of $x_{ij}$ and the coefficient of the cross-level interaction $x_{ij}w_j$. Here it is important to note the following: equation 4 shows that $\alpha_{11}$ is the coefficient of a cross-level multiplicative interaction between $x_{ij}$ and $w_j$. Equations 2 and 3 show that, although both $x_{ij}$ and the interaction $x_{ij}w_j$ vary within clusters, their slopes are estimated as context-level parameters. In this respect, the $m - l - 1$ rule differs crucially from the so-called “between-within degrees of freedom” (Schluchter and Elashoff, 1990). The latter assumes $m - l - 1$ degrees of freedom when testing the coefficients of variables that vary only between clusters, but uses the (usually much larger) residual degrees of freedom for variables that vary within clusters—as is the case for both $x_{ij}$ and $x_{ij}w_j$.

3.3.2 Satterthwaite Degrees of Freedom

An alternative to the $m - l - 1$ rule is to approximate degrees of freedom using a generalization of Satterthwaite’s (1946) method, as developed for single-constraint tests by Giesbrecht and Burns (1985) and extended to multiple-constraints F-tests by Fai and Cornelius (1996). We provide a brief summary of the single-constraint case (See Schaalje, McBride, and Fellingham, 2002, for a more extensive discussion).
The $t$-statistic for testing the single linear constraint $H_0: c'\alpha = 0$ is:

$$t = \frac{c'\hat{\alpha}}{\sqrt{c'\hat{\Sigma}_\alpha c}}$$

where $\hat{\Sigma}_\alpha$ denotes the estimated covariance matrix of the vector of parameter estimates $\hat{\alpha}$ and $c$ is a pre-determined vector. If the significance of a single coefficient estimate is to be tested, exactly one element of $c$ equals unity and all other elements equal zero. The statistic follows an approximate $t$-distribution with unknown degrees of freedom. Building on Satterthwaite (1946), Giesbrecht and Burns (1985) propose to approximate these degrees of freedom by

$$df = \frac{2(c'\hat{\Sigma}_\alpha c)^2}{\text{Var}(c'\hat{\Sigma}_\alpha c)}$$

where $\text{Var}(c'\hat{\Sigma}_\alpha c)$, can be approximated using the multivariate delta method. In contrast to the $m - l - 1$ rule, the Satterthwaite approximation estimates the degrees of freedom from the data, and is more widely applicable. In particular, it can provide approximate degrees of freedom for complex multilevel designs, for which the $m - l - 1$ rule cannot be applied (e.g. cross-classified structures).

The $m - l - 1$ and Satterthwaite approximation are not currently available in all major statistics packages. To our knowledge, SAS is the only package that implements these method for generalized linear mixed models such as mixed effects probit or logit. Even for linear mixed models, they are often not readily available. For example, Stata has only recently (in version 14) introduced them in the new $dfmethod$ option to the $\text{mixed}$ command. Moreover, the advantages of using the Satterthwaite (or Kenward-Roger) methods are probably largest when working with complex, non-hierarchical structures (such as cross-classified models). When dealing with the simple, hierarchical structures that are common in comparative politics and studied by Stegmueller (2013), the simple $m - l - 1$ heuristic may perform quite well.

### 3.4 Monte Carlo Evidence on Accurate Frequentist Inference

We now show that the finding of anti-conservative inference in Stegmueller (2013) stems from two sources. First, ML systematically underestimates the size of the random effects variances when the number of clusters is small. Since the random effects variances are involved in the estimation of standard errors for the fixed effects, ML-based inference for these
fixed effects is anti-conservative. Second, drawing on the normal (Gaussian) distribution to construct confidence intervals and p-values for the coefficients of contextual variables ignores that the latter are effectively based on a small sample. Inference should be based on the much heavier-tailed $t$-distribution with approximately correct degrees of freedom.

To illustrate these points, we draw on the replication of Stegmueller’s (2013) simulations that underlies the analysis of bias in variance components in Figure 2. As noted above, we replicate the original simulations using R. This allows us to implement a REML-like estimator for the probit model. Moreover, we can also obtain Satterthwaite degrees of freedom for the linear model. To our knowledge, there is no R function implementing this approximation for the probit model, but we can still use the $m - l - 1$ rule in this case (for further details on the simulation set up and on the R functions used, see supporting material.

We begin by examining the accuracy of confidence intervals based on REML estimators and on the $t$-distribution with degrees of freedom approximated by the $m - l - 1$ rule, which we can construct for both linear and probit mixed effects models. In a second step, we take a closer look and the linear case and compare the performance of the $m - l - 1$ rule to that of the Satterthwaite approximation.

Figure 3 shows bias in coverage rates for the direct context effect. Figure A3 in the Appendix shows results that are similar in conclusion for all three constitutive terms of a cross-level interaction. Coverage is defined as the percentage of all estimated 95% confidence intervals (i.e., for all of the 5,000 Monte Carlo replications) that actually cover the true value of the direct context effect, which we happen to know because this is a simulation. Bias in coverage or “CI non-coverage”, as Stegmueller has it, is the percentage point deviation of the actual coverage rate from the nominal level of 95%, because an accurate 95% confidence interval covers the true value in 95% of the cases. A solid horizontal line indicates the optimal value zero non-coverage. The Monte Carlo error is small due to the large number of simulations. To not overload the figure, we do not visualize confidence intervals for the individual non-coverage estimates. Instead, we include dashed horizontal lines that signal the limits of a two-sided 95% confidence interval for a Bernoulli-distributed random variable with $p(1) = .95$ (an accurate confidence interval would be such a variable). Non-coverage estimates that fall between the two dashed lines are not significantly different from zero at the 5% level (i.e., we cannot reject the hypothesis that their actual coverage rate equals the nominal level). Results are almost identical for the linear and probit cases, so we do not differentiate between the two in the following discussion.

The dashed gray line with hollow squares shows results that nicely replicate Stegmueller’s
Figure 3: Performance of Interval Estimates of Upper-Level Covariate Effect in Hierarchical Linear and Probit Models ($\gamma_1$ in Stegmueller's notation)

Note: Displayed are relative bias (in %) of 95% confidence interval coverage rates. The horizontal zero line denotes the reference of no bias. The dashed horizontal lines denote 95% test intervals and thereby express the uncertainty of these simulation results. This figure corresponds to the right-hand panel 'CI non-coverage' of Figure 2 on page 754 in Stegmueller (2013).
(2013) finding of confidence intervals that are much too narrow, particularly for small upper-level samples. If Stegmueller (2013) had used REML instead of ML, the extent of the problem would have been much smaller. The dashed gray line with hollow circles shows a maximum bias of ca. −9 percentage points rather than the horrifying −18 percentage points found when using ML estimators with the normal distribution. Constructing confidence intervals from the appropriate $t$-distribution, however, seems even more important than using REML. The solid black line with filled squares shows that if we stick with biased ML, but construct the confidence intervals from the $t$-distribution with $m - l - 1$ degrees of freedom, the maximum bias declines to a mere −3.8 percentage points.

Yet, the most important result of our replication study is that if we combine the REML estimator with the $m - l - 1$ rule, statistical inference is perfectly accurate. Apart from some random deviation, the solid black line with filled squares basically mirrors the horizontal line of no bias. This result goes far beyond a critique of Stegmueller (2013). Instead it suggests that a long-lasting concern of quantitative comparative research can be addressed quite easily. Even if based on a sample of only five clusters, accurate multilevel inference is possible. Of course, the confidence intervals constructed from a $t$-distribution with at most 3 degrees of freedom ($5 - 1 - 1$) will typically be huge. But they will cover the true coefficient estimate with an accuracy of 95%. Moreover, Figure A3 in the Appendix shows that all these insights hold similarly for all three constitutive terms of a cross-level interaction.

Given these results, one might wonder why we should bother with the complex and computationally intensive Satterthwaite (or Kenward-Roger) degrees of freedom approximations; the simple $m - l - 1$ rule seems to perform perfectly fine. This simple heuristic does indeed seem to carry a long way towards accurate inference. But we should also recognize the simplicity of the above-discussed simulation design. In particular, the $m - l - 1$ rule may seem most appropriate in a set-up like the current one, where lower-level variables exclusively vary within clusters and contextual predictors only vary between clusters. This equals a situation in which there are no compositional differences between clusters with regard to the lower-level predictors. To the degree that there are compositional differences, lower-level predictors also have between-cluster variance. Perhaps the simple $m - l - 1$ rule (with $l$ referring only to the number of upper-level predictors) fails if a model contains lower-level predictors with considerable compositional differences across clusters, as the distinction between lower-level and upper-level predictors becomes blurry in this situation.

A full-blown study of this issue is beyond the scope of our contribution. But to probe into it a bit, Figure 4 focuses on the linear case and reports coverage rates based on the REML
Figure 4: Performance of Degrees of Freedom Approximations for Hierarchical Linear Models

<table>
<thead>
<tr>
<th>Lower–level covariate with no compositional differences</th>
<th>Lower–level covariate with 50% compositional differences</th>
</tr>
</thead>
</table>

Note: Displayed are relative bias (in %) of 95% confidence interval coverage rates. The horizontal zero line denotes the reference of no bias. The dashed horizontal lines denote 95% test intervals and thereby express the uncertainty of these simulation results. This figure has no correspondence in Stegmueller (2013).
estimator and a $t$-distribution, with degrees of freedom approximated by the $m - l - 1$ rule and the Satterthwaite approximation. As before, Figure A4 in the Appendix shows that all conclusions similarly hold for cross-level interactions. In comparison to Figure 3, we now strongly zoom in to display the differences between the two approximations. The bias scale now covers a negligible range, from ca. $-0.5$ to $0.3$ percentage points. The left panel shows results for the simulation design discussed above, that is, there are no compositional difference with respect to the lower-level variable. Under this scenario there are essentially no differences between the $m - l - 1$ rule and the Satterthwaite method.

For the right panel, we modified the data generating process to introduce substantial between-cluster variance in the lower level predictor. More specifically, we introduced random variation in the cluster means in order to generate simulated data sets where, on average, 50 percent of the total variance of the lower-level predictor are between clusters. The results give no indication that the $m - l - 1$ rule performs worse under these conditions.

Overall, these results send a much more optimistic message concerning the possibility of accurate frequentist inference in few-cluster settings than Stegmueller (2013) and several other previous studies. At the same time, it is clear that the stylized DGP's considered here are exceptionally simple and well-behaved. Future research needs to investigate how these methods perform under more complex and realistic conditions.

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

As a first step towards applying these methods under more complex and realistic conditions, we follow Stegmueller (2013) and replicate Steenbergen and Jones’ (2002) model of citizen support for the European Union. Support for the European Union is modeled as a function of a country’s trade balance and tenure in the union, with GDP and monetary inflation as crucial controls. Stegmueller finds that frequentist confidence intervals based on the ML estimator and the normal approximation are considerably smaller than Bayesian credible intervals. For two contextual variables, trade balance and inflation, this leads to a rejection of the null hypothesis of no effect, where Bayesian credible intervals would retain it.

Figure 5 replicates Stegmueller’s results and adds REML estimates. In addition, we present two types of 95% confidence intervals for the two frequentist (i.e., ML and REML) estimates. Thick black lines depict intervals based on the normal distribution, while thin red antlers indicate the wider intervals based on the $t$-distribution with $m - l - 1$ degrees
Figure 5: Country-Level Determinants of Support for the European Union

Note: Displayed are point estimates and their 95% confidence or credible intervals. The thin red antlers signify the difference between normal (thick black lines) and $t$-distribution with $m - 1 - 1$ degrees of freedom based confidence intervals for (restricted) maximum likelihood estimates. The vertical zero line denotes the reference of no effect. Sample size: 10,777 individuals, 14 countries. This figure corresponds to Figure 8 on page 758 in Stegmueller (2013).
of freedom.

The most important lessons from this practical example are as follows: First, point estimates based on ML, REML and Bayesian MCMC are almost identical. The miniscule difference between the Bayesian and the other estimates is due to the randomness of MCMC. Second, 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 very similar to the corresponding Bayesian credible intervals. In fact, the preferred frequentist intervals even tend to be slightly more conservative than their Bayesian counterparts. In combination with the simulation evidence above, our best guess is that these confidence intervals are accurate. Third, using the $t$-distribution with nine degrees of freedom also improves the interval estimates based on full maximum likelihood so that we cannot reject the null-hypothesis concerning the inflation variable. This is, however, not enough to make a correct decision for the trade variable. Only the combination of REML and the $t$-distribution with approximated degrees of freedom is sufficient.

Thus, frequentist REML and Bayesian multilevel analysis produce very similar estimates of statistical uncertainty, provided that the former uses a $t$-distribution with appropriate degrees of freedom. There may be other benefits that justify the computational cost of Bayesian MCMC estimation, but a gain in inferential accuracy does not seem to be one of them.

5 Conclusions

Stegmueller’s (2013) widely-read and cited paper sends a sobering message concerning the performance of frequentist multilevel models in few-cluster setting. Stegmueller concludes that frequentist coefficient estimates for contextual variables are biased and that statistical inference for these effects is severely anti-conservative when the number of clusters or upper-level units (e.g. countries in case of cross-national comparative studies) is small. He further touts Bayesian MCMC methods as a supposedly superior alternative.

In this article, we have demonstrated that Stegmueller’s assessment of frequentist multilevel models cannot be upheld. First, we have shown, both analytically and by replicating and extending Stegmueller’s Monte Carlo simulations, that frequentist estimators of coefficients of multilevel models are unbiased—irrespective of the number of clusters and irrespective of whether the maximum likelihood or the restricted maximum likelihood
estimator is used. Second, we have discussed the advantages of the REML estimator in few-cluster settings and demonstrated that it can yield accurate inference for contextual effects if based on the $t$-distribution with correct degrees of freedom. These insights resolve long-lasting concerns of quantitative comparative political science and adjacent fields and have important implications for practitioners.

Our re-analysis of Stegmueller’s (2013) simulation evidence also carries important lessons on the design of Monte Carlo studies. In particular, we have identified three issues that can at least partly account for the misleading conclusions drawn from the simulation evidence. First, Stegmueller paid little attention to Monte Carlo sampling error and did not communicate the uncertainty of his simulation results. For instance, his results pertaining to the apparent bias of point estimates for cross-level interactions were never statistically significant. Second, and related, Stegmueller chose a relatively small number of 1,000 Monte Carlo replications. As a consequence many of his Monte Carlo estimates suggested large biases, but the apparently meaningful deviations were simply due to chance—a fact that was difficult to recognize because his graphs did not communicate the extent of Monte Carlo error. Once we increased the number of replications for the exact same random process, biases consistently shrank towards zero. Finally, Stegmueller (2013) repeatedly used the same random number seed for all experimental conditions, thereby introducing interdependencies among experimental conditions that created an appearance of systematic bias. When we repeated the same simulations with a different random number seed for each experimental condition, the systematic pattern disappeared. We suspect that many other studies suffer from similar shortcomings and hope that our analysis sensitizes both the producers and readers of Monte Carlo simulations to the importance of such technicalities.

Our analysis shows that two choices are critical for achieving correct inference about contextual effects when using frequentist multilevel models: using REML rather than ML estimation and using the $t$-distribution with the approximately correct degrees of freedom rather than the standard normal (Gaussian) distribution to obtain $p$-values and confidence limits. Thus, it is possible to get it right, but it is also easy to get it wrong. While we have not conducted a systematic review of published research, our casual impression is that the use of REML estimation is quite common, whereas the use of a $t$-distribution is not, partly because the normal approximation is the default in many widely used statistics packages. Unfortunately, our findings also indicate that the choice of the reference distribution is the more consequential one when it comes to achieving correct inference. In one important sense, Stegmueller’s (2013) findings therefore continue to give reason for concern: many
published studies presumably report severely anti-conservative inference for the effects of contextual variables.

A crucial challenge in obtaining accurate inference for frequentist estimates of context effects is to identify the correct degrees of freedom for the \( t \)-distribution. We have examined two options in this paper, the simple \( m - l - 1 \) rule (where \( m \) refers to the number of clusters and \( l \) to the number of contextual predictors) and the Satterthwaite approximation that uses a data-driven approach. Reassuringly, both methods performed very well in our replication and extension of Stegmueller’s (2013) simulation design, yielding accurate inference even in the extreme case of only five clusters. We are optimistic about their potential and it is reassuring that the \( m - l - 1 \) rule produces very reasonable confidence intervals in our replication of Steenbergen and Jones’ (2002) model of citizen support for the European Union. However, there is no guarantee that the good performance of the degree-of-freedom approximations carries over to all of the diverse settings that are studied by social scientists.\(^8\) Future simulation studies should therefore investigate the performance of the \( m - l - 1 \) rule and of the data-driven Satterthwaite and Kenward-Roger approximations under more complex and demanding data generating processes.

While our study shows that accurate frequentist inference for context effects is possible, Bayesian methods can have their merits in certain situations. But Stegmueller’s punchline that “The clearest advantage of employing a Bayesian approach to multilevel modeling lies in its excellent confidence interval coverage” (Stegmueller, 2013, 759), cannot be upheld.

---

\(^8\)A minimum requirement of course is that the number number of contextual variables is not too large, but satisfies \( m - l - 1 > 1 \).
Appendix A  Details on the Mathematical Notation Used

In this section we explain the mathematical notation used in the main text and the following sections of the appendix. We describe how the vectors $\alpha$ and $b$, as well as the matrices $X$, $Z$, $\Phi$ and $V$, are constructed in general and in a particular example.

Equation (4), which describes a simple two-level model with random intercepts and one set of random slopes, can also be written in a form involving vectors:

$$y_{ij} = x'_{ij} \alpha + z'_{ij} b + \epsilon_{ij}$$

where $\alpha$ is the vector with elements $\alpha_{00}$, $\alpha_{10}$, $\alpha_{01}$, and $\alpha_{11}$; $b$ is the vector with elements $b_{11}, b_{21}, \ldots, b_{1m}, b_{2m}$ (with $m$ as the number of clusters); $x_{ij}$ is the vector with elements 1, $x_{ij}$, $w_j$, and $x_{ij}w_j$ (and $x'_{ij}$ its transpose); and finally $z_{ij}$ is a vector with all elements equal to zero except for those elements equal to 1 and $x_{ij}$, respectively, at the appropriate places such that $z'_{ij} b = b_{0j} + x_{ij} b_{lj}$. If we collect the response observations $y_{ij}$ into the vector $y$ and the errors $\epsilon_{ij}$ into the vector $\epsilon$ and further the matrix $X$ with the vectors $x_{ij}$ as rows and the matrix $Z$ with the vectors $z_{ij}$ then we arrive at a special case of the more general matrix form of multilevel models given by the following equation (5).

In case of a two-level random-intercept model with $n$ observations, $m$ clusters at the second level, and one independent variable with values $x_1, \ldots, x_n$, then $X$ is a $n \times 2$ matrix, $Z$ is a $n \times m$, $\Phi$ is a $m \times m$ diagonal matrix and $\sigma^2 I$ is a $n \times n$ diagonal matrix with

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad Z = \begin{bmatrix} 1 & 0 \\ \vdots & \cdot \cdot \cdot \\ 0 & 1 \end{bmatrix}, \quad \Phi = \begin{bmatrix} \phi & \cdot \cdot \cdot \\ \cdot \cdot \cdot & \phi \end{bmatrix}, \quad \text{and} \quad \sigma^2 I = \begin{bmatrix} \sigma^2 & \cdot \cdot \cdot \\ \cdot \cdot \cdot & \sigma^2 \end{bmatrix}.$$
The variable $Z$ is an $n \times 2m$ matrix and $\Phi$ is a $2m \times 2m$ matrix with

$$Z = \begin{bmatrix}
1 & x_1 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots \\
1 & x_{i_1} & 0 & 0 \\
0 & 0 & 1 & x_{i_{m+1}} \\
\vdots & \vdots & \vdots & \vdots \\
0 & 0 & 1 & x_n
\end{bmatrix}, \quad \Phi = \begin{bmatrix}
\phi_{11} & \phi_{12} \\
\phi_{12} & \phi_{22} \\
\phi_{11} & \phi_{12} \\
\phi_{12} & \phi_{22}
\end{bmatrix}$$

Of course, the preceding examples are relatively simple cases. More general multilevel models with several levels of nested or cross-classified random effects can still be expressed in such a matrix form. The complexity of the random-effects structure will typically become manifest in the structure of the matrix $Z$, while matrix $\Phi$ will retain its block-diagonal structure.

If it is assumed, as usual, that $b$ and $\epsilon$ are orthogonal to each other, then (for given $\sigma^2$ and $\Phi$) the covariance matrix of $y - X\alpha$ and, equivalently, of $Zb + \epsilon$ can be expressed as

$$\text{Cov}(y - X\alpha) = E((y - X\alpha)(y - X\alpha)')$$
$$= E((Zb + \epsilon)(Zb + \epsilon)')$$
$$= ZE(bb')Z' + E(\epsilon\epsilon')$$
$$= Z\Phi Z' + \sigma^2 I$$

In the next section, we will consider the lower-level variance $\sigma^2$ and the random effects covariance matrix $\Phi$ as depending on a vector $\theta$ of variance parameters. In the case of a two-level model with random intercepts and a single set of random intercepts, we will have four variance parameters (e.g., enumerated as $\theta_1 = \sigma^2$, $\theta_2 = \phi_{11}$, $\theta_3 = \phi_{12}$, and $\theta_4 = \phi_{22}$). To emphasize that the covariance matrix of $y - X\alpha = Zb + \epsilon$ depends, via $\Phi$ and $\sigma^2$, on $\theta$, we use the notation $V(\theta)$ for this matrix.
Appendix B  ML Estimates of Fixed Effects in Linear Mixed Models Are Unbiased: A Proof

To show the unbiasedness of ML estimates of fixed effects coefficients, we adapt a proof of a more general proposition given in Kackar and Harville (1981). But before that, we discuss the following lemma on which this proof depends:

Lemma 1. Let \( \mathbf{u} \) be a random vector with components \( U_i \). If \( \mathbf{u} \) has a symmetric distribution around zero (i.e., \( \Pr(U_i \geq u^*) = \Pr(U_i \leq -u^*) = \Pr(-U_i \geq u^*) \) for each of its components and if \( g \) is an odd function (i.e., \( g(-x) = -g(x) \) for all \( x \)), then the random variable \( G = g(\mathbf{u}) \) has a symmetric distribution around zero. Further, if its expectation exists, it is equal to zero.

The proof of the first claim of this lemma closely follows the one given by Kackar and Harville (1981, 1257), but is a bit more explicit, while the second claim is based on a standard result of probability theory.

Proof. First we show that \( g(\mathbf{u}) \) and \( g(-\mathbf{u}) \) have the same distribution. Let \( \mathcal{A} \) refer to a subset of the range of \( \mathbf{u} \) and \( g(\mathcal{A}) \) to the image of the set \( \mathcal{A} \) with respect to the function \( g \). Then, because the distribution of \( \mathbf{u} \) is symmetric

\[
\Pr(g(-\mathbf{u}) \in g(\mathcal{A})) = \Pr(-\mathbf{u} \in \mathcal{A}) = \Pr(\mathbf{u} \in \mathcal{A}) = \Pr(g(\mathbf{u}) \in g(\mathcal{A}))
\]

We now turn to the two claims made in the lemma. Since \( g \) is an odd function, we have for any real number \( x \)

\[
\Pr(G \leq x) = \Pr(g(\mathbf{u}) \leq x) = \Pr(-g(\mathbf{u}) \geq -x) = \Pr(g(-\mathbf{u}) \geq -x) = \Pr(-G \leq x). \tag{10}
\]

Since \( \mathbf{u} \) and \( -\mathbf{u} \) have the same distribution, \( g(\mathbf{u}) \) and \( g(-\mathbf{u}) \) have the same distribution (i.e., \( \Pr(g(-\mathbf{u}) \leq x) = \Pr(g(\mathbf{u}) \leq x) \)). Together with the assumption that \( g \) is odd, this leads to

\[
\Pr(g(-\mathbf{u}) \geq -x) = \Pr(g(\mathbf{u}) \geq -x) = \Pr(-g(\mathbf{u}) \leq x) = \Pr(-G \leq x). \tag{11}
\]

From equations (10) and (11) it follows that \( \Pr(G \leq x) = \Pr(-G \leq x) \), which is the first claim of the lemma.

Now let \( p(x) \) denote the density of the distribution of \( G = g(\mathbf{u}) \) (i.e., \( \Pr(G \leq x) = \int_{-\infty}^{x} p(t) \, dt \)).
\[ \int_{-\infty}^{\infty} p(x) \, dx \]. If the integral \( \int_{-\infty}^{0} xp(x) \, dx \) is finite, then

\[
E(g(u)) = E(G) = \int_{-\infty}^{\infty} xp(x) \, dx = \int_{0}^{\infty} xp(x) \, dx + \int_{-\infty}^{0} xp(x) \, dx = \int_{0}^{\infty} -xp(-x) \, dx + \int_{0}^{\infty} xp(x) \, dx = \int_{0}^{\infty} (x - x)p(x) \, dx = 0,
\]

which proves the second claim of the lemma. □

After establishing the above lemma, we can now turn to discuss the unbiasedness of ML estimates of fixed-effects coefficients. If we assume that both the disturbances and the random effects have a (multivariate) normal distribution with zero mean and covariance matrices dependent on a vector \( \theta \) of variance parameters, then the log-likelihood function (after integrating out the random effects) is

\[
\ell(\alpha, \theta; y) = c - \frac{1}{2} \log \det(V(\theta)) - \frac{1}{2} (y - X\alpha)'V(\theta)^{-1}(y - X\alpha) \quad (12)
\]

(where \( c \) is a summary that is computed only based on \( y \) and that does not depend on any parameter in \( \alpha \) or \( \theta \)). Taking the derivative for \( \alpha \) and setting it to zero leads to the following equation for the ML estimate

\[
\hat{\alpha} = [X'V(\hat{\theta})^{-1}X]^{-1}X'V(\hat{\theta})^{-1}y \quad (13)
\]

where

\[
V(\theta) = \sigma^2 I + Z\Phi Z'.
\]

That is, \( V(\theta) \) is the covariance matrix of \( y - X\alpha = Zb + \epsilon \) (we use the modified notation \( V(\theta) \) to emphasize the dependence of the covariance matrix on the parameter vector \( \theta \)).

Based on this setup we prove the following theorem:

**Theorem 1.** If the assumptions of the normal linear mixed model are satisfied, and if the expectation of the ML estimate given by (13) exists, then \( \hat{\alpha} \) is unbiased. That is,

\[
E(\hat{\alpha}) = \alpha \quad (14)
\]

\[A \text{ well-known counter-example is the Cauchy distribution.}\]
Proof. The argument stated in the main text implies that the difference between both sides of equation (14) is

\[ E(\hat{\alpha}) - \alpha = E \left( [X'V(\hat{\theta})^{-1}X]^{-1}X'V(\hat{\theta})^{-1}[Zb + \epsilon] \right) \]  

(15)

The ML estimate \( \hat{\alpha} \) is unbiased if (and only if) this difference equals zero. What therefore remains to be shown, is that the expectation on the right-hand side of the equation is equal to zero. If \( V(\hat{\theta}) \) (and thus \( [X'V(\hat{\theta})^{-1}X]^{-1}X'V(\hat{\theta})^{-1} \)) were constant, then this would simply follow from the fact that the expectation of \( Zb + \epsilon \) is zero by assumption. However \( V(\hat{\theta}) \) depends on \( Zb + \epsilon \).

If one collects the error component vectors \( b \) and \( \epsilon \) into the vector \( u = (b', \epsilon)' \) then one can define a function \( \psi(u) \) as

\[ \psi(u) = [X'V(\hat{\theta})^{-1}X]^{-1}X'V(\hat{\theta})^{-1}Ku, \]

(16)

with \( K = (Z, I) \) and \( Zb + \epsilon = Ku \), so that equation (15) is the expectation \( E(\psi(u)) \). By assumption, \( u \) has a symmetric distribution around zero (viz., a multivariate normal distribution). Therefore, by Lemma 1, \( \psi(u) \) has zero expectation, if each of its elements is an odd function of \( u \) (i.e., if \( \psi(-u) = -\psi(u) \)). That is, what remains to be shown is that \( \psi(u) \) is indeed an odd function. This is demonstrated in two steps: First it is shown that ML estimator of the variance parameters \( \theta \) is translation-invariant and even, and second it is shown that this implies that \( \psi(u) \) is an odd function.

The first step of the proof is to show that the ML estimator of \( \theta \) is translation-invariant and even. These properties can be explained if we consider the ML estimator as a function of the observations of the dependent variable, i.e.

\[ \hat{\theta}(\cdot) = \text{argmax } \ell(\cdot; \theta). \]

The ML estimator is called translation invariant if for any constant vector \( k \)

\[ \hat{\theta}(y - Xk) = \hat{\theta}(y). \]

Further, it is called even, if

\[ \hat{\theta}(-y) = \hat{\theta}(y). \]

To demonstrate these properties of the ML estimator of \( \theta \), we utilize equation (13) to
construct the concentrated log-likelihood or profile log-likelihood, in which the dependence on the fixed-effects coefficients $\alpha$ is eliminated:

$$
\ell_p(\theta; y) = \ell(\hat{\alpha}_\theta, \theta; y) \\
= c - \frac{1}{2} \log \det(V(\theta)) - \frac{1}{2} [y - X\hat{\alpha}_\theta]' V(\theta)^{-1} [y - X\hat{\alpha}_\theta] \\
= c - \frac{1}{2} \log \det(V(\theta)) - \frac{1}{2} y'[I - P(\theta)]' V(\theta)^{-1} [I - P(\theta)] y
$$

with

$$
P(\theta) = X[X'V(\theta)^{-1}X]^{-1}X'V(\theta)^{-1}
$$

and

$$
X\hat{\alpha}_\theta = P(\theta)y
$$

Since $P(\theta)X = X$, for any vector $k$ we have

$$
[I - P(\theta)]Xk = (X - X)k = 0.
$$

hence

$$
[I - P(\theta)](y - Xk) = [I - P(\theta)]y
$$

and therefore

$$
\ell^*(y - Xk; \theta) = \ell^*(y; \theta).
$$

That is, the profile log-likelihood is a translation-invariant function of the observed values of the dependent variable. Further, because $\ell^*(y; \theta)$ is a quadratic form in $y$, we have

$$
\ell^*(-y; \theta) = \ell^*(y; \theta).
$$

Since the profile log-likelihood is translation invariant, the ML estimator of $\theta$ is also translation invariant:

$$
\hat{\theta}(y - Xk) = \text{argmax} \ell^*(y - Xk; \theta) = \text{argmax} \ell^*(y; \theta) = \hat{\theta}(y).
$$

Further its evenness implies the evenness of the ML estimator of $\theta$:

$$
\hat{\theta}(-y) = \text{argmax} \ell^*(-y; \theta) = \text{argmax} \ell^*(y; \theta) = \hat{\theta}(y).
$$

Now that we have demonstrated that the ML estimator of $\theta$ is translation-invariant and
even we show in the second step that this implies that the function defined in equation (16) is an odd function of the error components $u = (b', \epsilon)'$. To this purpose it is convenient to define

$$M(\theta) = [X'V(\theta)^{-1}X]^{-1}X'V(\theta)^{-1}$$

and

$$\hat{M}(y) = M(\hat{\theta}(y))$$

so that we can write

$$\hat{\alpha}_\theta = M(\theta)y$$

and

$$\hat{\alpha} = \hat{M}(y)y.$$

From the translation-invariance and the evenness of the ML estimate of $\theta$ follows

$$\hat{\theta}(y) = \hat{\theta}(y - X\alpha) = \hat{\theta}(Zb + \epsilon) = \hat{\theta}(Ku) = \hat{\theta}(-Ku).$$

hence

$$M(\hat{\theta}(y)) = M(\hat{\theta}(y - X\alpha)) = M(\hat{\theta}(Ku)) = M(\hat{\theta}(-Ku)).$$

and

$$\hat{M}(y) = \hat{M}(y - X\alpha) = \hat{M}(Ku) = \hat{M}(-Ku)$$

so that equation (16) becomes

$$\psi(u) = \hat{M}(y)Ku = \hat{M}(Ku)Ku.$$  

We can therefore see that

$$\psi(-u) = \hat{M}(K[-u])K[-u] = -\hat{M}(-Ku)Ku = -\hat{M}(Ku)Ku = -\psi(u),$$

that is, $\psi(u)$ is an odd function. From this and the symmetry of distribution of $u$, we can conclude with the help of Lemma 1 that, if the relevant expectation exists, it is

$$E(\hat{\alpha}(y; \hat{\theta}) - \alpha) = E(\psi(u)) = 0,$$

which concludes the proof of the theorem. $\square$

The theorem proved in this section of the appendix establishes the unbiasedness of an
ML estimator of fixed-effects coefficients if it exists. However, this is a pre-condition that does not need to be always satisfied. An obvious necessary condition is that there are sufficient data to identify the estimates. That is, the number of observations at the lower-level and at the upper-level should be larger than the number of parameters in the model. Some sufficient conditions for the existence of ML estimates are discussed by Jiang (1999).

**Appendix C  The Restricted Maximum Likelihood Estimator**

**Appendix C.1  REML for Linear Mixed Effects Models**

In the main text we discuss restricted maximum likelihood estimators as a remedy for the bias of ML estimates in variance components. Restricted maximum likelihood estimators were introduced by Patterson and Thompson (1971) and can be interpreted as special cases of modified profile likelihood estimators later introduced by Cox and Reid (1987). These modified profile likelihood estimators have their use beyond multilevel modeling and an application of the modified profile likelihood principle to some common example may help to understand how REML estimators are able to improve over ML estimators of variance parameters.

Recall that the log-likelihood function of a linear mixed effects model is given by

\[
\ell(\alpha, \theta; y) = c - \frac{1}{2} \log \det(V(\theta)) - \frac{1}{2}(y - X\alpha)'V(\theta)^{-1}(y - X\alpha)
\]  

(18)

where

\[
V(\theta) = \sigma^2 I + Z\Phi Z'
\]

In the previous section, we already employed the fact that, for given values of the variance parameters, ML estimates for the fixed effects coefficients in \( \alpha \) can be obtained by the single GLS-step:

\[
\hat{\alpha}_\theta = M(\theta)y
\]

where

\[
M(\theta) = [X'V(\theta)^{-1}X]^{-1}X'V(\theta)^{-1}.
\]
The concentrated log-likelihood or profile log-likelihood can be constructed thus:

\[ \ell_p(\theta; y) = \ell(\hat{\alpha}_\theta, \theta; y) \]

\[ = c - \frac{1}{2} \log \det(V(\theta)) - \frac{1}{2} [y - X\hat{\alpha}_\theta]'V(\theta)^{-1}[y - X\hat{\alpha}_\theta] \]

\[ = c - \frac{1}{2} \log \det(V(\theta)) - \frac{1}{2} y'[I - P(\theta)]'V(\theta)^{-1}[I - P(\theta)]y \]

\[ = c - \frac{1}{2} \log \det(V(\theta)) - \frac{1}{2} y'V(\theta)^{-1}[I - P(\theta)]y. \]

with

\[ P(\theta) = XM(\theta) = X[X'V(\theta)^{-1}X]^{-1}X'V(\theta)^{-1}. \]

The last step here is based on

\[ P(\theta)'V(\theta)^{-1} = V(\theta)^{-1}P(\theta) \]

and

\[ V(\theta)^{-1}P(\theta)P(\theta) = V(\theta)^{-1}P(\theta). \]

The REML estimator proposed by Patterson and Thompson (1971) then is the value of \( \theta \) that maximises

\[ \ell_{\text{REML}}(\theta; y) = \ell_p(\theta; y) - \frac{1}{2} \ln \det \left( XV^{-1}X \right). \] (19)

A REML estimator is obviously a special case of a modified profile likelihood estimator (Cox and Reid, 1987). To understand maximum modified profile likelihood estimators, one needs to understand profile likelihood functions. Suppose \( \psi \) is the parameter vector of a statistical model and suppose further that it can be split into two parts \( \lambda \) and \( \theta \), such that the log-likelihood function \( \ell(\psi; y) = \ell(\lambda, \theta; y) \) can easily be maximized for \( \lambda \) with \( \theta \) held fixed (as in the case of regression coefficients in a normal linear model, or the fixed-effects coefficients in a normal linear mixed effects model). If \( \hat{\lambda}_\theta \) denotes the value of \( \lambda \) that maximizes \( \ell(\lambda, \theta; y) \) for \( \theta \) held fixed, then the profile log-likelihood function is a function of \( \theta \) alone defined as:

\[ \ell_p(\theta; y) = \ell(\hat{\lambda}_\theta, \theta; y) \]

The modified profile likelihood approach suggested by Cox and Reid (1987) consists in maximizing instead of the profile log-likelihood just discussed, a modified version thereof,
The profile log-likelihood function with respect to the disturbance variance \(\sigma^2\) is

\[
\ell_p(\sigma^2; y) = c - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2}(y - X\hat{\alpha}_{OLS})'(y - X\hat{\alpha}_{OLS}) = c - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2}y'(I - P_X)y
\]

where \(n\) is the number of observations, \(c\) a data-dependent constant (i.e., not dependent on the parameters \(\alpha\) and \(\sigma^2\)) and \(P_X = X[X'X]^{-1}X'\) (Harville, 1997, 166ff). The corresponding modified profile log-likelihood function is

\[
\ell_{mp}(\sigma^2; y) = c - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2}y'(I - P_X)y - \frac{1}{2} \ln \det \left(1\sigma^2 X'X\right)
\]

since \(\det \left(\sigma^{-2}X'X\right) = (\sigma^{-2})^{k+1} \det (X'X)\), because \(X'X\) is a \((k + 1) \times (k + 1)\) matrix, and therefore \(\frac{1}{2} \ln \det \left(\frac{1}{\sigma^2}X'X\right) = \frac{1}{2} \ln \det (X'X) - \frac{k+1}{2} \ln \sigma^2\). Setting \(\partial\ell_{mp}(\sigma^2)/\partial \sigma^2\) to zero leads to

\[
-\frac{n - k - 1}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4}y'(I - P_X)y = 0
\]

\[
\sigma^2(n - k - 1) = y'(I - P_X)y
\]

the well-known unbiased estimator of \(\sigma^2\),

\[
\hat{\sigma}^2 = \frac{(y - X\hat{\alpha}_{OLS})'(y - X\hat{\alpha}_{OLS})}{n - k - 1}.
\]

It should be noted that if a linear regression model with i.i.d. disturbances is considered as a special case of a linear mixed model with \(V = \sigma^2 I\), then the REML estimator of \(\sigma^2\) is the same as the one just derived. Furthermore, the unbiased estimator of the variance discussed in the main text is just another special case, with \(V = \sigma^2 I, X = 1\) (a vector with all elements of which are equal to unity so that \(X\hat{\alpha}_{OLS} = \bar{y}1\)), and \(k = 0\).

**Appendix C.2 REML for Generalized Linear Mixed Models**

Generalized linear mixed models (GLMMs) that go beyond the normal-linear type pose particular challenges in addition to those involved in the estimation of linear mixed mod-
els, on the one hand, and generalized linear models (without random effects), on the other hand. First, they lead to likelihood functions that involve (sometimes high-dimensional) integrals that do not have a closed-form solution. Second, due to the non-linearity in the link between coefficients and the conditional expectation of the response variable, coefficient estimates inevitably are biased in small samples (McCullagh and Nelder, 1989) and it may be difficult to establish how quickly this bias vanishes as the sample size increases (for bias correction in generalized linear models, see Firth, 1993). So the relatively reassuring result about the unbiasedness of estimators for parameters in normal-linear mixed model proved in Appendix B, does not necessary carry over to generalized linear mixed models.\(^{10}\)

In the following we first present the structure of generalized linear mixed models and discuss how ML and REML estimators would work in this type of models. If \(f(y|b; \alpha, \sigma^2)\) is the density or probability mass function of the conditional distribution of the response for given values of the random effects vector, then the log-likelihood function for a generalized linear mixed model takes the form of the integral

\[
\ell(\alpha, \theta; y) = c - \frac{1}{2} \ln \det(\Phi) + \ln \int f(y|b; \alpha, \sigma^2) \exp \left[ -\frac{1}{2} b'\Phi^{-1}b \right] db
\]

for which a solution formula exists only, if the conditional distribution of \(y\) given \(b\) is normal. In the absence of a solution formula, the integral involved in the log-likelihood function of generalized linear mixed models cannot be computed exactly, but only be approximated. The chief analytical approximation in use is the Laplace approximation (Breslow and Clayton, 1993), whereas the most widely used numeric approximations are Gauss-Hermite quadrature and Monte Carlo integration (McCulloch, 1997; Booth and Hobert, 1999; Caffo, Jank, and Jones, 2005).

The crucial advantage of the Laplace approximation, introduced by Breslow and Clayton (1993) as penalized quasi-likelihood (PQL), is that it makes it easy to translate the concept of restricted maximum likelihood to generalized linear mixed models beyond the normal-linear case. The Laplace approximation is given by

\[
\ell(\alpha, \theta; y) \approx c - \frac{1}{2} \ln \det(\Phi) + \ln f(y|\tilde{b}; \alpha, \sigma^2) - \frac{1}{2} \ln \det \left( \tilde{K} + \Phi^{-1} \right) - \frac{1}{2} \tilde{b}'\Phi^{-1}\tilde{b},
\]

\(^{10}\)This makes it all the more surprising that the bias found by Stegmueller (2013) in coefficient estimates of a multilevel probit-model is smaller than the one he finds for a linear multilevel model.
where \( \tilde{b} \) maximizes the integrand in equation (21) or, equivalently, its logarithm

\[
\ln f(y|b; \alpha, \sigma^2) - \frac{1}{2} b' \Phi^{-1} b
\]

and

\[
\tilde{K} = - \left( \frac{\partial^2 \ln f(y|b; \alpha, \sigma^2)}{\partial b \partial b'} \right)_{b=\tilde{b}}
\]

is the Hessian of \( \ln f(y|b; \alpha, \sigma^2) \) evaluated at \( b = \tilde{b} \).

If the conditional distribution of the response is in a exponential family, a generalized linear mixed model is characterized by the linear predictor

\[
\eta_i = X\alpha + Zb,
\]

the conditional mean

\[
\mu_i = E(y_i|\tilde{b}),
\]

the link function \( g(\cdot) \) that gives

\[
\eta_i = g(\mu_i),
\]

and conditional variance

\[
\text{Var}(y_i|\tilde{b}) = \sigma^2 a_i \nu(\mu_i)
\]

(where \( \sigma^2 \) is the dispersion parameter, \( \nu(\cdot) \) is the variance function, and \( a_i \) are some predetermined weights, as they arise, for instance, because of the denominator of a binomial distribution). Breslow and Clayton (1993) point out that the fixed-effects coefficients for given \( \Phi \) can be estimated by maximizing the Laplace approximated marginal likelihood (22) by iteratively solving the GLS equation

\[
X'\tilde{V}^{-1}X\hat{\alpha} = X'\tilde{V}^{-1}y^*
\]

where \( y^* \) is the usual “working response” known from the GLM literature (McCullagh and Nelder, 1989) with components

\[
y^*_i = \tilde{\eta}_i + (y_i - \tilde{\mu}_i) \frac{\partial \eta_i}{\partial \mu_i}
\]

and GLS weighting matrix

\[
\tilde{V} = \tilde{W}^{-1} + Z\Phi Z',
\]

(24)
where $\tilde{W}^{-1}$ is the inverse of a diagonal matrix with diagonal elements

$$\tilde{w}_{ii} = [\sigma^2 a_i v(\tilde{\mu}_i)]^{-1} \left( \frac{\partial \tilde{\mu}_i}{\partial \eta_i} \right)^2.$$ 

Linear mixed models are a special case of generalized linear mixed models where $\mu_i = \eta_i$, $v(\mu_i) = 1$, and $a_i = 1$. Multilevel probit models (as considered in the main text of this paper) are characterised by the probit function (i.e., the inverse of the cumulative probability function of the standard normal distribution) as link function, $v(\mu_i) = \mu_i (1 - \mu_i)$, and for binary responses, $a_i = 1$. (For binomial counts $a_i$ will be data-dependent.)

For the estimation of the variance parameters, Breslow and Clayton (1993) propose to use

$$q(\theta; y) = -\frac{1}{2} \det(V) - \frac{1}{2} (y^* - X\hat{\alpha}_\theta) V^{-1} (y^* - X\hat{\alpha}_\theta). \quad (25)$$

as an objective function, which is analogous to the log-likelihood in the normal-linear case, yet only with the linearized dependent variable $y^*$ instead of $y$ and $\hat{\alpha}_\theta$ is the solution to equation (23). For a “REML-like” variant of PQL, Breslow and Clayton (1993) propose to use instead the modified objective function

$$q^*(\theta; y) = q(\theta; y) - \frac{1}{2} \det \left( X\tilde{\Sigma}^{-1} X \right). \quad (26)$$

It should be noted that the accuracy of the Laplace approximation depends on the size of the upper-level units (and not on their number). With smaller sizes of upper-level units, the Laplace approximation may lead to bias (usually a downward bias of the variance parameters). For dealing with such situations, bias-corrections based on a higher-order Laplace approximation have been proposed (Breslow and Lin, 1995; Lin and Breslow, 1996) as well as Monte-Carlo integration approaches, that allow to increase the accuracy of the approximation of the integrals involved in the likelihood to any desired degree by increasing the number of Monte Carlo replicates (algorithms for automatically increasing the Monte Carlo sample sizes have been proposed by Booth and Hobert 1999 and Caffo, Jank, and Jones 2005). How the “logic” of REML can be applied to these setups is much less straightforward, but see McCullagh and Tibshirani (1990).

As an alternative that is computationally less demanding than methods based on numerical or even analytical approximations and that at the same time is also more general by allowing for non-normal distributions of the random effects $b$, Lee and Nelder (1996) developed the so-called $h$-likelihood technique. They introduced this technique for the
estimation of the parameters of hierarchical generalized linear models (HGLMs), which allow the random effect vector $b$ to have not only a normal but also a non-normal distribution. That is GLMMs are a special case of HGMLs. A $h$-likelihood function for a HGLM with normal random-effects distribution is almost identical to the Laplace approximation of the marginal likelihood given by equation (22), the main difference being that the term $-\frac{1}{2} \ln \det (\tilde{K} + \Phi^{-1})$ is absent in the $h$-likelihood. In fact the REML-like modifications to the $h$-likelihood technique discussed by Lee and Lee (2012), which form the basis of the software used for our Monte Carlo study of mixed probit model estimation (Ronnegård, Alam, and Shen, 2015), lead to an objective function that is virtually identical to the REML-modification of PQL-objective function discussed by Breslow and Clayton (1993).

Appendix D Monte Carlo Simulation Design

To test our claim that the inferential problems of frequentist mixed effects models stem from using maximum likelihood (instead of restricted maximum likelihood) and from drawing on the normal instead of the $t$-distribution with approximated degrees of freedom, we replicate Stegmueller’s (2013) Monte Carlo simulation. The crucial dimension of comparison is the size of the upper-level sample. Like Stegmueller (2013), we vary the number of clusters $m$, which ranges from 5 to 30 in steps of 5, with each cluster having 500 lower-level observations. Equations 27 and 28 describe the two basic data generating processes (DGP) for the linear case. The DGPs are identical to those studied by Stegmueller (2013), but we use slightly different notation for consistency with the above discussion. In the first variant of the DGP (equation 27) the context-level variable $w_j$ has a simple additive effect on the (lower-level) outcome $y_{ij}$:

$$y_{ij} = 0 + 0.25x_{ij} + 0.2w_j + b_{ij} + \epsilon_{ij}$$ (27)

Generally, $x_{ij}$ and $w_j$ are independent and normally distributed with means of 0 and standard deviations of 1, just like in Stegmueller (2013). However, to test the robustness of the $m - l - 1$ rule (see Figures 4 and A4), we additionally consider cross-cluster compositional differences in $x_{ij}$ by randomly shifting cluster means away from 0. The random shifts have a mean of 0 and a standard deviation of 1. Accordingly, this introduces cross-cluster compositional differences in $x_{ij}$ of 50%. Because this operation introduces additional (cross-cluster or between) variation to $x_{ij}$, we rescale $x_{ij}$ afterwards, so that it again has a mean of
0 and standard deviations of 1. The residual/lower-level error $\epsilon_{ij}$ has a variance of 2 and the upper-level random effect $b_{1j}$ has a variance of 0.2222 in the baseline case with an intraclass correlation of 0.10. Like Stegmueller (2013), we also considered intraclass correlations of 0.05 and 0.15 (by appropriately modifying the variance of $b_{1j}$). As in Stegmueller’s analysis, this did not influence the results (see Figures 2 and A2). We therefore focuses on the case ICC = 0.10 in the main article.

The second variant of the DGP additionally includes a cross-level interaction between $x_{ij}$ and $w_j$

$$y_{ij} = 0 + 0.25x_{ij} + 0.2w_j + 0.1x_{ij}w_j + b_{1j} + b_{2j}x_{ij} + \epsilon_{ij}$$  \hspace{1cm} (28)

where $b_{2j}$ is an additional random effect on the slope of $x_{ij}$ with a variance of 0.3. The covariance between $b_{1j}$ and $b_{2j}$ is set to yield a correlation of approximately 0.39.

Again following Stegmueller (2013), we use the same basic DGPs for the generalized linear/probit case. However, the DGPs in equations 27 and 28 are now used to construct a latent continuous variable $y^*_ij$. The dichotomous outcome variable for the probit model is 1 if $y^*_ij > 0$ and 0 otherwise. Moreover, the variance of $\epsilon_{ij}$ is set to 1 in the probit case and the variances and covariances of the upper-level random effects are modified to yield the same correlations and intraclass correlations as in the linear case.

We conduct all simulations in R version 3.3.1, and estimate the linear mixed effects models using the lmer function from the lme4 package with the default optimizer bobyqa (Bates, Maechler, Bolker et al., 2015). The probit mixed effects models are estimated using the hglm2 function from the hglm package (Rönnegård, Alam, and Shen, 2015), which allows for the application of REML-type modifications to the PQL-technique, which is commonly used for parameter estimation of generalized linear mixed models. Rönnegård, Alam, and Shen (2015) call their estimation technique EQL1 and base it on Lee and Nelder (1996) and Lee and Lee (2012). Yet the REML-like modification from Lee and Lee (2012) that Rönnegård, Alam, and Shen (2015) use differs from Breslow and Clayton’s (1993) only in certain algorithmic details. We performed Satterthwaite approximations using the lmerTest (Kuznetsova, Brockhoff, and Bojesen Christensen, 2015) package. Currently, however, these approximations are not implemented for non-linear models.

Appendix E Additional Results
Figure A1: Performance of Point Estimates of All Three Consecutive Terms of a Cross Level Interaction in Hierarchical Linear and Probit Models

<table>
<thead>
<tr>
<th>Linear Model</th>
<th>Probit Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image-url" alt="Graph" /></td>
<td><img src="image-url" alt="Graph" /></td>
</tr>
</tbody>
</table>

Note: Displayed are relative biases (in %) of point estimates and 95% confidence intervals of these results. The 95% confidence intervals are cut at −21 and 21 percent bias respectively to advance visualization. Particularly the strongly asymmetric confidence intervals extend far beyond the cut-off values. The horizontal zero line denotes the reference of no bias. From the top to the bottom row, this figure corresponds to the left-hand panels ‘Estimate’ of Figure 3 (p. 755), Figure 4 (p. 756), and Figure 5 (p. 757) in Stegmueller (2013).
Figure A2: Performance of Variance Estimates of Random Intercepts and Random Slopes in Hierarchical Linear and Probit Models

Note: Displayed are relative biases (in %) variance estimates of random intercept and 95% confidence intervals of these results. The horizontal zero line denotes the reference of no bias. This figure has no correspondence in Stegmueller (2013).
Figure A3: Performance of Interval Estimates of All Three Consecutive Terms of a Cross Level Interaction in Hierarchical Linear and Probit Models

Note: Displayed are relative bias (in %) of 95% confidence interval coverage rates. The horizontal zero line denotes the reference of no bias. The dashed horizontal lines denote 95% test intervals and thereby express the uncertainty of these simulation results. From the top to the bottom row, this figure corresponds to the right-hand panels 'CI non-coverage' of Figures 3 (p. 755), Figures 4 (p. 756), and Figure 5 (p. 757) in Stegmueller (2013).
Figure A4: Performance of Degrees of Freedom Approximations for All Three Consecutive Terms of a Cross Level Interaction in Hierarchical Linear Models

**Legend:**
- **m−l−1 rule**
- **Satterthwaite**

**Note:** Displayed are relative bias (in %) of 95% confidence interval coverage rates. The horizontal zero line denotes the reference of no bias. The dashed horizontal lines denote 95% test intervals and thereby express the uncertainty of these simulation results. This figure has no correspondence in Stegmueller (2013).
References


Kuznetsova, Alexandra, Brockhoff, Per Bruun, and Bojesen Christensen, Rune Haubo, 2015. “lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package).”


