ON-LINE APPENDIX
Multilevel Analysis with Few Clusters: Improving Likelihood-based Methods to Provide Unbiased Estimates and Accurate Inference

Martin Elff
Zeppelin University

Jan Paul Heisig
WZB Berlin Social Science Center

Merlin Schaeffer
University of Cologne

Susumu Shikano
University of Konstanz

Contents

A Details for the theoretical section of the paper

A.1 Notation of linear multilevel models

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 $\boldsymbol{\alpha}$ and $\boldsymbol{b}$, as well as the matrices $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{\Phi}$ and $\boldsymbol{V}$, are constructed in general and in a particular example.

A linear two-level model with k independent variables, n observations in total that are nested in m upper-level units can generally written in the following form:

$$y_{ij} = \alpha_0 + \alpha_1 x_{1ij} + \cdots + \alpha_k x_{kij} + b_{0j} + b_{1j} x_{h_1 jk} + \cdots + b_{qj} x_{h_q jk} + \epsilon_{ij} \tag{1}$$

where y_{ij} denotes the value of the dependent variable of the i-th individual-level observation nested in the j-th cluster or upper-level units and $x_{1ij}, \ldots, x_{kij}$ are the corresponding values of the independent variables (and $x_{h_1 ij}, \ldots, x_{h_q ij}$ is a subset of these). The constant α_0 and the coefficients $\alpha_1, \ldots, \alpha_k$ of the independent variables are referred to as *fixed effects* or *fixed effects coefficients*, since they are considered as (fixed, non-random) model parameters that are usually to be estimated from the data (i.e. the observed values of the dependent and the independent variables). b_{0j} is the *random intercept* for the j-th cluster or upper-level unit, $b_{1j}, \ldots, b_{qj}$ are the *random slopes* of a subset of the independent variables.

Hierarchical linear models are just an interpretation of certain multilevel models and can be brought in the form of equation (1). As an example, consider the hierarchical model

$$y_{ij} = a_{0j} + a_{1j} u_{ij} + \epsilon_{ij} \tag{2}$$

$$a_{0j} = \gamma_{00} + \gamma_{01} w_j + b_{0j} \tag{3}$$

$$a_{1j} = \gamma_{10} + \gamma_{11} w_j + b_{1j} \tag{4}$$

where u_{ij} is the value of an individual-level independent variable and ϵ_{ij} is an individual-level error, while a_{0j} and a_{1j} are a group-level intercept and slope. By substituting the right-hand sides of equations (3) and (4) for a_{0j} and a_{1j} in equation (2) and rewriting $x_{1ij} = w_j$, $x_{2ij} = u_{ij}$, $x_{3ij} = u_{ij} w_j$, $\alpha_0 = \gamma_{00}$, $\alpha_1 = \gamma_{01}$, $\alpha_2 = \gamma_{10}$, and $\alpha_3 = \gamma_{11}$, this model can be rewritten as a special case of equation (1).

By arranging the fixed effect coefficients in equation (1) into the vector $\boldsymbol{\alpha}$, the values of the independent variables $x_{1ij}, \ldots, x_{kij}$ into the vector $\boldsymbol{x}_i j$, the random intercepts and random slopes $b_{01}, \ldots, b_{qm}$ into the random vector $\boldsymbol{b}$ and by composing a vector $\boldsymbol{z}_i j$ of zeros, ones and the values $x_{h_1 ij}, \ldots, x_{h_q ij}$ at the appropriate places, equation (1) can be written in vector form:

$$y_{ij} = \boldsymbol{x}'_{ij} \boldsymbol{\alpha} + \boldsymbol{z}'_{ij} \boldsymbol{b} + \epsilon_{ij} \tag{5}$$

In general, any linear multilevel model can be written in this form. In matrix form, such a model can be written in matrix form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon} \tag{6}$$

where the *response vector* $\boldsymbol{y}$ has elements y_{ij}, the *predictor matrix* $\boldsymbol{X}$ has rows $\boldsymbol{x}'_{ij}$, the matrix $\boldsymbol{Z}$ has rows $\boldsymbol{z}'_{ij}$, and $\boldsymbol{\epsilon}$ has elements ϵ_{ij}.

A.2 Aspects of Multilevel Models: Distributional Assumptions and Log-likelihood Function

In the construction of multilevel models of the form (6) it is generally assumed that the random effects vector $\boldsymbol{b}$ and the residual error vector $\boldsymbol{\epsilon}$ are uncorrelated with one another and with $\boldsymbol{X}$ and have a normal distribution with zero expectation and covariance matrices $\boldsymbol{\Phi}$ and $\sigma^2\boldsymbol{I}$, respectively. Usually $\boldsymbol{\Phi}$ has a block-diagonal structure. For example, in a two-level model with one set of random intercepts and one set of random slopes (for a single independent variable), $\boldsymbol{\Phi}$ would be composed of m identical 2×2-matrices along the diagonal and zeroes everywhere else. The expectation and variance of $\boldsymbol{y}$ (conditional on or with fixed $\boldsymbol{X}$) are then:

$$\mathrm{E}(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\alpha} \tag{7}$$

and

$$\mathrm{Var}(\boldsymbol{y}|\boldsymbol{X}) = \mathrm{E}([\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}][\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}]') = \mathrm{E}(\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}][\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}]') = \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}' + \sigma^2\boldsymbol{I}. \tag{8}$$

In the following we consider the lower-level variance σ^2 and the random effects covariance matrix $\boldsymbol{\Phi}$ as depending on a vector $\boldsymbol{\theta}$ of *variance parameters*. In case of a two-level model with one set of random intercepts and one set of random slopes (for a single independent variable) $\boldsymbol{\theta}$ has four elements, one element for the variance of the random intercepts, one element of the variance of the random slopes, one element for the covariance of the random intercepts and random slopes, and one element for σ^2 the variance of the residual errors—even though $\boldsymbol{\Phi}$ would be a $2m \times 2m$-matrix (where m as before refers to the number of upper-level units). To emphasise that the covariance matrix of $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha} = \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$ depends, via $\boldsymbol{\Phi}$ and σ^2, on $\boldsymbol{\theta}$, we use the notation $\boldsymbol{V}(\boldsymbol{\theta})$ for this matrix.

If the random effects $\boldsymbol{b}$ were observed, one could estimate the model parameters in $\boldsymbol{\alpha}$, σ^2, and $\boldsymbol{\Phi}$ by maximising the "complete-data" log-likelihood:

$$\begin{aligned}\ell(\boldsymbol{\alpha}, \boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{b}) = &-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2}\ln\det(\boldsymbol{\Phi}) \\ &- \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{Z}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{Z}\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}'\boldsymbol{\Phi}^{-1}\boldsymbol{b}\end{aligned} \tag{9}$$

However, this complete-data log-likelihood cannot be used to estimate model parameters, because it depends on the unobserved random effects $\boldsymbol{b}$. This dependence can be eliminated by integrating out the random effects, to arrive at a marginal log-likelihood:

$$\begin{aligned}
\ell(\boldsymbol{\alpha}, \boldsymbol{\theta}; \boldsymbol{y}) &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}) \\
&\quad - \frac{1}{2}\ln\det(\boldsymbol{\Phi}) + \ln\int \exp\left[\frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})'\boldsymbol{Z}\boldsymbol{b} - \frac{1}{2}\boldsymbol{b}'\left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Phi}^{-1}\right)\boldsymbol{b}\right] \mathrm{d}\,\boldsymbol{b} \\
&= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}) \\
&\quad - \frac{1}{2}\ln\det(\boldsymbol{\Phi}) - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln\det\left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Phi}^{-1}\right) \\
&\quad - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})'\boldsymbol{Z}\left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Phi}^{-1}\right)^{-1}\frac{1}{\sigma^2}\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}) \\
&= -\frac{n+p}{2}\ln(2\pi) - \frac{1}{2}\ln\det\left(\sigma^2\boldsymbol{I} + \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}'\right) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})'\left(\sigma^2\boldsymbol{I} + \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}'\right)^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}) \\
&= -\frac{n+p}{2}\ln(2\pi) - \frac{1}{2}\log\det(\boldsymbol{V}(\boldsymbol{\theta})) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})'\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})
\end{aligned} \tag{10}$$

where $\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}' + \sigma^2\boldsymbol{I}$, n is the number of observations, and p is the number of columns in $\boldsymbol{Z}$. The relevant integration formula can be found as Theorem 15.12.1 in Harville (1997, 322). This derivation also relies on the Sherman-Morrison-Woodbury formula which implies

$$\left(\sigma^2\boldsymbol{I} + \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}'\right)^{-1} = \frac{1}{\sigma^2}\boldsymbol{I} - \frac{1}{\sigma^2}\boldsymbol{Z}\left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Phi}^{-1}\right)^{-1}\frac{1}{\sigma^2}\boldsymbol{Z}'.$$

Taking the first and second derivatives of the log-likelihood function (10) for the coefficient vector $\boldsymbol{\alpha}$ leads to

$$\frac{\partial\ell(\boldsymbol{\alpha}, \boldsymbol{\theta}; \boldsymbol{y})}{\partial\boldsymbol{\alpha}} = \boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}) \tag{11}$$

and

$$-\frac{\partial^2\ell(\boldsymbol{\alpha}, \boldsymbol{\theta}; \boldsymbol{y})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'} = \boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}. \tag{12}$$

A.3 Conditions for the Unbiasedness of Coefficient Estimators for Multilevel Models

Stegmueller (2013) suggests that frequentist estimators for coefficients in multilevel models are or may be biased if the number of clusters is small, while Bayesian MCMC techniques have no or at least a smaller bias. However, already a relatively simple frequentist estimator coefficients in a linear multilevel model, such as ordinary least squares (OLS), is unbiased, while on the other hand Bayes estimators are *never* unbiased unless under unusual circumstances (see Casella and Berger 2002, 368 or Lehmann and Casella 2011,

234). In the following we briefly discuss the properties of OLS and GLS as estimators of coefficients in linear multilevel models and present a proof for the unbiasedness of maximum likelihood estimators of these model parameters.

Ordinary least squared is the BLUE estimator for the coefficients of linear regression model with uncorrelated disturbances. While conditions of the Gauss-Markov theorem generally do not apply to multilevel models (Greene 2012, 100), an OLS estimator can still be computed. The OLS estimator for the fixed-effects vector of the multilevel model given by equation (6) is

$$\hat{\boldsymbol{\alpha}}_{\text{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{13}$$

The expected value (given $\boldsymbol{X}$) of the OLS estimator is

$$\mathrm{E}\,(\hat{\boldsymbol{\alpha}}_{\text{OLS}}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{E}(\boldsymbol{y}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\alpha} = \boldsymbol{\alpha}$$

and the variance is

$$\mathrm{Var}\,(\hat{\boldsymbol{\alpha}}_{\text{OLS}}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{Var}(\boldsymbol{y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

with $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}' + \sigma^2\boldsymbol{I}$.

In contrast, the variance computed under the conditions of the Gauss-Markov theorem (that are usually reported by statistical software packages along with OLS estimates) is

$$\widehat{\mathrm{Var}}_{\text{OLS}}\,(\hat{\boldsymbol{\alpha}}_{\text{OLS}}|\boldsymbol{X}) = \hat{\sigma}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

which is obviously incorrect if $\boldsymbol{\Phi}$ does not vanish, i.e. the random effects a non-zero variances. That is, while the OLS estimator retains its unbiasedness even if the (correct) model contains random effects, standard errors computed under the assumption of the Gauss Markov theorem will be too small. Statistical software packages such as *Stata* provide so-called "robust" or "cluster robust" standard errors motivated by the work of Huber (1967) and White (1982). These robust standard errors are based on the attempt to reconstruct the matrix $\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}$ without the need to estimate $\boldsymbol{\Phi}$ and σ^2.

If the covariance matrix of $\boldsymbol{b}$ and the variance of the elements of $\boldsymbol{\epsilon}$ ($\boldsymbol{\Phi}$ and σ^2, respectively) are known (which is generally not the case in practice), the variance of $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha} = \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$ is also known as $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}' + \sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. In this case (provided that $\boldsymbol{V}$ is non-singular) one can use the generalized least squares (GLS) estimator

$$\hat{\boldsymbol{\alpha}}_{\text{GLS}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{14}$$

to obtain estimates for the model coefficients. It is easy to see that the GLS estimator is unbiased:

$$\mathrm{E}\,(\hat{\boldsymbol{\alpha}}_{\text{GLS}}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\,\mathrm{E}(\boldsymbol{y}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\alpha} = \boldsymbol{\alpha}$$

and has variance:

$$\begin{aligned}\mathrm{Var}\,(\hat{\boldsymbol{\alpha}}_{\text{GLS}}|\boldsymbol{X}) &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\,\mathrm{Var}(\boldsymbol{y})\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\\ &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}.\end{aligned}$$

Aitken's theorem (Aitken 1936) states that the GLS estimator is not only unbiased, but a best linear unbiased estimator (BLUE). The proof for the BLUE property is relatively straightforward: with the modified data

$$\tilde{\boldsymbol{y}} = \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{y} \text{ and } \tilde{\boldsymbol{X}} = \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{X}$$

where $\boldsymbol{V}^{-\frac{1}{2}}$ is a matrix such that $\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}} = \boldsymbol{V}^{-1}$, e.g. a Cholesky decomposition of $\boldsymbol{V}^{-1}$ the GLS estimator is a variant of OLS:

$$\hat{\boldsymbol{\alpha}}_{\text{GLS}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} \tag{15}$$

therefore GLS is BLUE because this instance of an OLS estimator is BLUE. Conversely, insofar as the conventional OLS estimator in equation (13) differs from GLS as in equation (15), the latter cannot be BLUE and hence is inefficient (because less efficient than GLS).

Aitken's theorem rests on the assumption that the matrix $\boldsymbol{V}$ is fixed and thus does not depend on the values of the dependent variable $\boldsymbol{y}$. If σ^2 and $\boldsymbol{\Phi}$ are unknown, as generally is the case in practice, one has to use an estimate of $\boldsymbol{V}$, say $\hat{\boldsymbol{V}}$, which is based on the empirical data and thus no longer independent from $\boldsymbol{y}$. Thus, a crucial condition for Aitken's theorem no longer applies. Nevertheless, Kackar and Harville (1981) explicate conditions under which coefficient estimates remain unbiased. Consider the expected value of a feasible generalized least squares (FGLS) estimator with estimated $\hat{\boldsymbol{V}}$:[1]

$$\mathrm{E}\left(\hat{\boldsymbol{\alpha}}|\boldsymbol{X}\right) = \boldsymbol{\alpha} + \mathrm{E}\left[(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon})\right]. \tag{16}$$

If the expected value in the final term of Equation 16 is zero, the estimator is unbiased. This is the case if $\hat{\boldsymbol{V}}$ is constant or depends only on $\boldsymbol{X}$. As Kackar and Harville show, the expected value of this term is also zero if $\hat{\boldsymbol{V}}$ is a symmetric and translation-invariant function of $\boldsymbol{y}$. They further show that these requirements are satisfied if $\hat{\boldsymbol{V}}$ is based on MLEs of σ^2 and $\boldsymbol{\Phi}$. The symmetry and translation invariance of $\hat{\boldsymbol{V}}$ depends only on how it is constructed. In particular, it does *not* depend on the unbiasedness of the estimates of the variance parameters σ^2 and $\boldsymbol{\Phi}$. Therefore, the maximum likelihood estimates $\hat{\boldsymbol{\alpha}}$ of the fixed effect coefficients of a linear multilevel model are *unbiased*—even if σ^2 and $\boldsymbol{\Phi}$ are estimated with bias.[2]

While it is straightforward to demonstrate the unbiasedness of OLS and GLS estimators of fixed-effects coefficients of linear multilevel models, showing the unbiasedness of the ML estimator is more complicated. For the proof we adapt a more general theorem and proof given in Kackar and Harville (1981). But before that, we discuss the following lemma on which this proof depends:

Lemma 1. *Let $\boldsymbol{u}$ be a random vector with components U_i. If $\boldsymbol{u}$ 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., $g(-\boldsymbol{x}) = -g(\boldsymbol{x})$ for all $\boldsymbol{x}$), then the random variable $G = g(\boldsymbol{u})$ has a symmetric distribution around zero. Further, if its expectation exists, it is equal to zero.*

1. This type of estimators includes feasible generalized least squares estimators known from the econometrics of panel data (Baltagi 2008; Greene 2012) but notably also the maximum likelihood estimator for the fixed effects-coefficients of multilevel models.

2. An explicit proof is given in Appendix A.4.

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 $g(\boldsymbol{u})$ and $g(-\boldsymbol{u})$ have the same distribution. Let $\mathscr{A}$ refer to a subset of the range of $\boldsymbol{u}$ and $g(\mathscr{A})$ to the image of the set $\mathscr{A}$ with respect to the function g. Then, because the distribution of $\boldsymbol{u}$ is symmetric

$$\Pr(g(-\boldsymbol{u}) \in g(\mathscr{A})) = \Pr(-\boldsymbol{u} \in \mathscr{A}) = \Pr(\boldsymbol{u} \in \mathscr{A}) = \Pr(g(\boldsymbol{u}) \in g(\mathscr{A}))$$

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

$$\Pr(G \leq x) = \Pr(g(\boldsymbol{u}) \leq x) = \Pr(-g(\boldsymbol{u}) \geq -x) = \Pr(g(-\boldsymbol{u}) \geq -x). \tag{17}$$

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

$$\Pr(g(-\boldsymbol{u}) \geq -x) = \Pr(g(\boldsymbol{u}) \geq -x) = \Pr(-g(\boldsymbol{u}) \leq x) = \Pr(-G \leq x). \tag{18}$$

From equations (17) and (18) 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 $G = g(\boldsymbol{u})$ (i.e., $\Pr(G \leq x) = \int_{-\infty}^{x} p(x)\,\mathrm{d}\,x$). If the integral $\int_{-\infty}^{0} xp(x)\,\mathrm{d}\,x$ is finite,[3] then

$$\begin{aligned}
\mathrm{E}(g(\boldsymbol{u})) = \mathrm{E}(G) &= \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}\,x = \int_{-\infty}^{0} xp(x)\,\mathrm{d}\,x + \int_{0}^{\infty} xp(x)\,\mathrm{d}\,x \\
&= \int_{0}^{\infty} -xp(-x)\,\mathrm{d}\,x + \int_{0}^{\infty} xp(x)\,\mathrm{d}\,x \\
&= \int_{0}^{\infty} -xp(x)\,\mathrm{d}\,x + \int_{0}^{\infty} xp(x)\,\mathrm{d}\,x \\
&= \int_{0}^{\infty} (x - x)p(x)\,\mathrm{d}\,x = 0,
\end{aligned}$$

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 $\boldsymbol{\theta}$ of variance parameters, then log-likelihood can (as derived above) be written as:

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\theta}; \boldsymbol{y}) = c - \frac{1}{2}\log\det(\boldsymbol{V}(\boldsymbol{\theta})) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})'\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}) \tag{19}$$

3. A well-known counter-example is the Cauchy distribution.

where c is a summary that is computed only based on $\boldsymbol{y}$ and that does not depend on any parameter in $\boldsymbol{\alpha}$ of $\boldsymbol{\theta}$.

Taking the derivative for $\boldsymbol{\alpha}$ and setting it to zero leads to the following equation for the ML estimate

$$\hat{\boldsymbol{\alpha}} = [\boldsymbol{X}'\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{y}. \tag{20}$$

That is, $\boldsymbol{V}(\boldsymbol{\theta})$ is the covariance matrix of $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha} = \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$ (we use the modified notation $\boldsymbol{V}(\boldsymbol{\theta})$ to emphasise the dependence of the covariance matrix on the parameter vector $\boldsymbol{\theta}$).

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* (20) *exists, then* $\hat{\boldsymbol{\alpha}}$ *is unbiased. That is,*

$$\mathrm{E}(\hat{\boldsymbol{\alpha}}|\boldsymbol{X}) = \boldsymbol{\alpha} \tag{21}$$

Proof. The argument stated in the main text implies that the difference between both sides of equation (21) is

$$\mathrm{E}(\hat{\boldsymbol{\alpha}}|\boldsymbol{X}) - \boldsymbol{\alpha} = \mathrm{E}\left([\boldsymbol{X}'\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1}[\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}]\right) \tag{22}$$

The ML estimate $\hat{\boldsymbol{\alpha}}$ 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 $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$ (and thus $[\boldsymbol{X}'\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1}$) were constant, then this would simply follow from the fact that the expectation of $\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$ is zero by assumption. However $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$ depends on $\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$.

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

$$\boldsymbol{\psi}(\boldsymbol{u}) = [\boldsymbol{X}'\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{K}\boldsymbol{u}, \tag{23}$$

with $\boldsymbol{K} = (\boldsymbol{Z}, \boldsymbol{I})$ and $\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon} = \boldsymbol{K}\boldsymbol{u}$, so that right-hand side of equation (22) is identical to the expectation $\mathrm{E}(\boldsymbol{\psi}(\boldsymbol{u}))$. By assumption, $\boldsymbol{u}$ has a symmetric distribution around zero (viz., a multivariate normal distribution). Therefore, by Lemma 1, $\boldsymbol{\psi}(\boldsymbol{u})$ has zero expectation if each of its elements is an odd function of $\boldsymbol{u}$ (i.e., if $\boldsymbol{\psi}(-\boldsymbol{u}) = -\boldsymbol{\psi}(\boldsymbol{u})$). That is, what remains to be shown is that $\boldsymbol{\psi}(\boldsymbol{u})$ is indeed an odd function. This is demonstrated in two steps: First it is shown that ML estimator of the variance parameters $\boldsymbol{\theta}$ is translation-invariant and even, and second it is shown that this implies that $\boldsymbol{\psi}(\boldsymbol{u})$ is an odd function.

The first step of the proof is to show that the ML estimator of $\boldsymbol{\theta}$ 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.

$$\hat{\boldsymbol{\theta}}(\cdot) = \operatorname{argmax} \ell(\cdot; \boldsymbol{\theta}).$$

The ML estimator is called *translation invariant* if for any constant vector $\boldsymbol{k}$

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{k}) = \hat{\boldsymbol{\theta}}(\boldsymbol{y}).$$

Further, it is called *even*, if

$$\hat{\boldsymbol{\theta}}(-\boldsymbol{y}) = \hat{\boldsymbol{\theta}}(\boldsymbol{y}).$$

To demonstrate these properties of the ML estimator of $\boldsymbol{\theta}$, we utilise equation (20) to construct the concentrated log-likelihood or profile log-likelihood, in which the dependence on the fixed-effects coefficients $\boldsymbol{\alpha}$ is eliminated:

$$\begin{aligned}\ell_{\mathrm{p}}(\boldsymbol{\theta};\boldsymbol{y}) &= \ell(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\theta}},\boldsymbol{\theta};\boldsymbol{y}) \\ &= c - \frac{1}{2}\log\det(\boldsymbol{V}(\boldsymbol{\theta})) - \frac{1}{2}[\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\boldsymbol{\theta}}]'\boldsymbol{V}(\boldsymbol{\theta})^{-1}[\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\boldsymbol{\theta}}] \\ &= c - \frac{1}{2}\log\det(\boldsymbol{V}(\boldsymbol{\theta})) - \frac{1}{2}\boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta})]'\boldsymbol{V}(\boldsymbol{\theta})^{-1}[\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta})]\boldsymbol{y}\end{aligned} \tag{24}$$

with

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

and

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

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

$$[\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta})]\boldsymbol{X}\boldsymbol{k} = (\boldsymbol{X} - \boldsymbol{X})\boldsymbol{k} = 0.$$

hence

$$[\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta})](\boldsymbol{y} - \boldsymbol{X}\boldsymbol{k}) = [\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta})]\boldsymbol{y}$$

and therefore

$$\ell^*(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{k};\boldsymbol{\theta}) = \ell^*(\boldsymbol{y};\boldsymbol{\theta}).$$

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

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

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

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{k}) = \operatorname{argmax}\ell^*(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{k};\boldsymbol{\theta}) = \operatorname{argmax}\ell^*(\boldsymbol{y};\boldsymbol{\theta}) = \hat{\boldsymbol{\theta}}(\boldsymbol{y}).$$

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

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

Now that we have demonstrated that the ML estimator of $\boldsymbol{\theta}$ is translation-invariant and even, we show in the second step that this implies that the function defined in equation (23) is an odd function of the error components $\boldsymbol{u} = (\boldsymbol{b}', \boldsymbol{\epsilon})'$. To this purpose it is convenient to define

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

and

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

so that we can write

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

and

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

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

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \hat{\boldsymbol{\theta}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}) = \hat{\boldsymbol{\theta}}(\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}) = \hat{\boldsymbol{\theta}}(\boldsymbol{K}\boldsymbol{u}) = \hat{\boldsymbol{\theta}}(-\boldsymbol{K}\boldsymbol{u}).$$

hence

$$\boldsymbol{M}(\hat{\boldsymbol{\theta}}(\boldsymbol{y})) = \boldsymbol{M}(\hat{\boldsymbol{\theta}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})) = \boldsymbol{M}(\hat{\boldsymbol{\theta}}(\boldsymbol{K}\boldsymbol{u})) = \boldsymbol{M}(\hat{\boldsymbol{\theta}}(-\boldsymbol{K}\boldsymbol{u})).$$

and

$$\hat{\boldsymbol{M}}(\boldsymbol{y}) = \hat{\boldsymbol{M}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}) = \hat{\boldsymbol{M}}(\boldsymbol{K}\boldsymbol{u}) = \hat{\boldsymbol{M}}(-\boldsymbol{K}\boldsymbol{u})$$

so that equation (23) becomes

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

We can therefore see that

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

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

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

which concludes the proof of the theorem. □

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).

The above result that coefficient estimates are unbiased only applies to *linear* multilevel models. It is well-known that multilevel logistic regression and other generalised linear mixed models (GLMMs) such as multilevel probit *are* subject to small-sample biases.[4] Estimation of

4. In fact, maximum likelihood coefficient estimates exhibit a small-sample bias even in the case of "conventional" logistic or probit regression without random effects. That is, coefficient estimates tend to be systematically and substantially larger (in absolute size) than the corresponding true vales when the sample size is small. Even worse, it might occur that no MLE for a particular sample exists, because of the problem of *separation* (Zorn 2005).

GLMMs is either computationally highly demanding (involving numerical integration using methods such as multidimensional quadrature) or relies on approximations (most notably the Laplace approximation, see Breslow and Clayton 1993; Pinheiro and Bates 1995; McCulloch 1997; Booth and Hobert 1999; Caffo, Jank, and Jones 2005). Approximation methods in particular require large sample sizes to achieve (approximate) unbiasedness, but crucially the relevant sample size here is the one at the *lower level*. In most political science applications—and especially in the case of comparative cross-national analysis that motivated Stegmueller's study—the requirement of a large lower-level sample will typically be met.

A.4 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 modelling 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

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\theta}; \boldsymbol{y}) = c - \frac{1}{2}\log\det(\boldsymbol{V}(\boldsymbol{\theta})) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})'\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}) \tag{25}$$

where

$$\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}' + \sigma^2\boldsymbol{I}$$

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 $\boldsymbol{\alpha}$ can be obtained by the single GLS-step:

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

where

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

The concentrated log-likelihood or profile log-likelihood can be constructed thus:

$$\begin{aligned}
\ell_{\mathrm{p}}(\boldsymbol{\theta}; \boldsymbol{y}) &= \ell(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\theta}}, \boldsymbol{\theta}; \boldsymbol{y}) \\
&= c - \frac{1}{2}\log\det(\boldsymbol{V}(\boldsymbol{\theta})) - \frac{1}{2}[\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\boldsymbol{\theta}}]'\boldsymbol{V}(\boldsymbol{\theta})^{-1}[\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\boldsymbol{\theta}}] \\
&= c - \frac{1}{2}\log\det(\boldsymbol{V}(\boldsymbol{\theta})) - \frac{1}{2}\boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta})]'\boldsymbol{V}(\boldsymbol{\theta})^{-1}[\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta})]\boldsymbol{y} \\
&= c - \frac{1}{2}\log\det(\boldsymbol{V}(\boldsymbol{\theta})) - \frac{1}{2}\boldsymbol{y}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}[\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta})]\boldsymbol{y}.
\end{aligned}$$

with

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

The last step here is based on

$$\boldsymbol{P}(\boldsymbol{\theta})'\boldsymbol{V}(\boldsymbol{\theta})^{-1} = \boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{P}(\boldsymbol{\theta})$$

and

$$\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{P}(\boldsymbol{\theta})\boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{P}(\boldsymbol{\theta}).$$

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

$$\ell_{\text{REML}}(\boldsymbol{\theta}; \boldsymbol{y}) = \ell_{\text{p}}(\boldsymbol{\theta}; \boldsymbol{y}) - \frac{1}{2} \ln \det \left(\boldsymbol{X}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X} \right). \tag{26}$$

The REML estimator can be derived by integrating out the coefficient vector $\boldsymbol{\alpha}$ from the full likelihood function:

$$\begin{aligned}
\int \exp(\ell(\boldsymbol{\alpha}, \boldsymbol{\theta}; \boldsymbol{y}))\, \mathrm{d}\, \boldsymbol{\alpha} &= \exp(c) \det(\boldsymbol{V}(\boldsymbol{\theta}))^{-\frac{1}{2}} \int \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})' \boldsymbol{V}(\boldsymbol{\theta})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}) \right) \mathrm{d}\, \boldsymbol{\alpha} \\
&= \exp(c) \det(\boldsymbol{V}(\boldsymbol{\theta}))^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \boldsymbol{y}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{y} \right) \\
&\quad \cdot \int \exp\left(\frac{1}{2} \boldsymbol{y}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X}\boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}' \boldsymbol{X}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X}\boldsymbol{\alpha} \right) \mathrm{d}\, \boldsymbol{\alpha} \\
&= \exp(c) \det(\boldsymbol{V}(\boldsymbol{\theta}))^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \boldsymbol{y}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{y} \right) \\
&\quad \cdot \det\left(\boldsymbol{X}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X} \right)^{-\frac{1}{2}} \exp\left(\frac{1}{2} \boldsymbol{y}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X} \left(\boldsymbol{X}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X} \right)^{-\frac{1}{2}} \boldsymbol{X}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{y} \right) \\
&= \exp(c) \det(\boldsymbol{V}(\boldsymbol{\theta}))^{-\frac{1}{2}} \det\left(\boldsymbol{X}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X} \right)^{-\frac{1}{2}} \\
&\quad \cdot \exp\left(-\frac{1}{2} \boldsymbol{y}' \left[\boldsymbol{V}(\boldsymbol{\theta})^{-1} - \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X} \left(\boldsymbol{X}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X} \right)^{-\frac{1}{2}} \boldsymbol{X}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \right] \boldsymbol{y} \right) \\
&= \exp(c) \det(\boldsymbol{V}(\boldsymbol{\theta}))^{-\frac{1}{2}} \det\left(\boldsymbol{X}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X} \right)^{-\frac{1}{2}} \\
&\quad \cdot \exp\left(-\frac{1}{2} \boldsymbol{y}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} [\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta})] \boldsymbol{y} \right) \\
&= \det\left(\boldsymbol{X}' \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{X} \right)^{-\frac{1}{2}} \exp(\ell_{\text{p}}(\boldsymbol{\theta}; \boldsymbol{y})) \\
&= \exp(\ell_{\text{REML}}(\boldsymbol{\theta}; \boldsymbol{y}))
\end{aligned}$$

The relevant integration formula can be found as Theorem 15.12.1 in Harville (1997, 322).

A linear regression model with i.i.d. disturbances can be considered as a special case of a linear mixed model with $\boldsymbol{V} = \sigma^2 \boldsymbol{I}$. We now show that the unbiased variance estimator of the residual variance σ^2 is a special case of REML. First note that log-likelihood function with respect to the disturbance variance σ^2 is

$$\ell_{\text{p}}(\sigma^2; \boldsymbol{y}) = c - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\text{OLS}})' (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\text{OLS}}) = c - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \boldsymbol{y}' (\boldsymbol{I} - \boldsymbol{P}_X) \boldsymbol{y}$$

where n is the number of observations, c a data-dependent constant (i.e., not dependent on the parameters $\boldsymbol{\alpha}$ and σ^2) and $\boldsymbol{P}_X = \boldsymbol{X}[\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'$ (Harville 1997, 166ff). The corresponding REML objective function is

$$\begin{aligned}\ell_{\text{REML}}(\sigma^2; \boldsymbol{y}) &= c - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y} - \frac{1}{2}\ln\det\left(\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X}\right) \\ &= c - \frac{n-k-1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y} - \frac{1}{2}\ln\det(\boldsymbol{X}'\boldsymbol{X})\end{aligned}$$

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

$$\begin{aligned}-\frac{n-k-1}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y} &= 0 \\ \sigma^2(n-k-1) &= \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}\end{aligned}$$

the solution of which is

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

A.5 REML, Firth's Penalised Likelihood, and Bayes Estimators

It is noteworthy that the penalty term in the REML objective function is related to the information matrix with respect to $\boldsymbol{\alpha}$:

$$\frac{1}{2}\ln\det\left(\boldsymbol{X}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\right) = \frac{1}{2}\ln\det\left(-\frac{\partial\ell(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}\right) = \frac{1}{2}\ln\det\left(-\operatorname{E}\left[\frac{\partial\ell(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}\right]\right)$$

and is equal to the logarithm of Jeffreys' uniform prior (Jeffreys 1946) with respect to the coefficient vector $\boldsymbol{\alpha}$ for given $\boldsymbol{\theta}$:

$$p_{\text{Jf}}(\boldsymbol{\alpha}|\boldsymbol{\theta}) = \det\left(-\operatorname{E}\left[\frac{\partial\ell(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}\right]\right)^{\frac{1}{2}}$$

(where $\operatorname{E}(\boldsymbol{A})$ refers to the expected value of a random matrix $\boldsymbol{A}$). Note that in the linear-normal case

$$-\operatorname{E}\left(\frac{\partial\ell(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}\right) = -\frac{\partial\ell(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'} = \boldsymbol{X}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}.$$

However, this should not lead to the misunderstanding that REML is based on a Bayesian maximum a posteriori value of the coefficient vector $\boldsymbol{\alpha}$ (given the variance parameters). A maximum a posteriori (MAP) value of $\boldsymbol{\alpha}$ with a Jeffreys prior would be obtained by maximising the posterior

$$\begin{aligned}p(\boldsymbol{\alpha}|\boldsymbol{\theta},\boldsymbol{y}) &= p(\boldsymbol{\alpha}|\boldsymbol{\theta},\boldsymbol{y})p_{\text{Jf}}(\boldsymbol{\alpha}|\boldsymbol{\theta}) \\ &= \exp\left[\ell(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{y}) + \frac{1}{2}\ln\det\left(-\operatorname{E}\frac{\partial\ell(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}\right)\right].\end{aligned} \tag{27}$$

In the log-posterior, the modification of the log-likelihood thus has the *opposite* sign of the penalty term in the REML objective function. Further, in the normal-linear case, $p_{\mathrm{Jf}}(\boldsymbol{\alpha}|\boldsymbol{\theta})$ varies only with the variance parameters in $\boldsymbol{\theta}$, but not with the coefficients vector $\boldsymbol{\alpha}$. That is, for given variance parameters $\boldsymbol{\theta}$, the MAP value of $\boldsymbol{\alpha}$ is identical to the (conditional) MLE.

It is nevertheless tempting to attribute the bias reduction achieved by REML to the involvement of Jeffreys' prior. Firth (1993) proposes a method to correct or at least to reduce the finite sample bias of maximum likelihood estimators and draws a connection between his bias-corrected MLEs and MAP estimators with Jeffreys' prior. Firth's method applies to models where the log-likelihood function is non-linear in the parameters. For example, in case of logistic regression, his method involves maximising

$$\ell^*(\boldsymbol{\alpha}; \boldsymbol{y}) = \ell(\boldsymbol{\alpha}; \boldsymbol{y}) + \frac{1}{2} \ln \det \left(- \mathrm{E} \frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \right)$$

instead of the log-likelihood function $\ell(\boldsymbol{\alpha}; \boldsymbol{y})$. Note that the value of $\boldsymbol{\alpha}$ that maximises $\ell^*(\boldsymbol{\alpha}; \boldsymbol{y})$ (the penalised likelihood estimate) will be different from the one that maximises $\ell(\boldsymbol{\alpha}; \boldsymbol{y})$ (the MLE) in most generalised linear models with coefficient vector $\boldsymbol{\alpha}$ and will usually have a smaller bias than the MLE. An exception are linear regression models with normal distributed errors. In this case the MLE of the coefficient vector (which is identical to the OLS estimate) is already unbiased, while "penalty term" is constant with respect to the coefficient vector:

$$\frac{1}{2} \ln \det \left(- \mathrm{E} \frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \right) = \frac{1}{2} \ln \det \left(\boldsymbol{X}' \boldsymbol{X} \right).$$

Modifying the log-likelihood of a linear regression model with normal errors in the manner proposed by Firth (1993) will therefore not serve as a correction of any bias in the MLE of the regression coefficients, however the MLE of the coefficients is unbiased in this setup anyway.

Adding Jeffrey's prior with *respect to the error variance* in a normal-linear regression leads to results different from REML or bias-corrected estimator of the error variance. The second derivative of the log-likelihood for σ^2 in this case is:

$$\frac{\partial^2 \ell(\boldsymbol{\alpha}, \sigma^2; \boldsymbol{y})}{(\partial \sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}) = \frac{n}{2(\sigma^2)^2} - \frac{S^2}{(\sigma^2)^3}$$

(with $S^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})$).

Therefore, if Firth's penalised log-likelihood with respect to σ^2 takes the form

$$\begin{aligned}
\ell^*(\boldsymbol{\alpha}, \sigma^2; \boldsymbol{y}) &= \ell(\boldsymbol{\alpha}, \sigma^2; \boldsymbol{y}) + \ln \left(- \frac{\partial^2 \ell(\boldsymbol{\alpha}, \sigma^2; \boldsymbol{y})}{(\partial \sigma^2)^2} \right) \\
&= - \frac{n}{2} \ln(\sigma^2) - \frac{S^2}{2\sigma^2} + \ln \left(- \frac{n}{2(\sigma^2)^2} + \frac{S^2}{(\sigma^2)^3} \right) \\
&= - \frac{n}{2} \ln(\sigma^2) - \frac{S^2}{2\sigma^2} + \ln \left(- \frac{n\sigma^2}{2(\sigma^2)^3} + \frac{S^2}{(\sigma^2)^3} \right) \\
&= - \frac{n}{2} \ln(\sigma^2) - \frac{S^2}{2\sigma^2} + \ln \left(- \frac{n}{2} \sigma^2 + S^2 \right) - 3 \ln(\sigma^2)
\end{aligned}$$

$$= -\frac{n+6}{2}\ln(\sigma^2) - \frac{S^2}{2\sigma^2} + \ln\left(S^2 - \frac{n}{2}\sigma^2\right).$$

The derivative for σ^2 is

$$\frac{\partial \ell^*(\boldsymbol{\alpha}, \sigma^2; \boldsymbol{y})}{\partial \sigma^2} = -\frac{n+6}{2}\frac{1}{\sigma^2} + \frac{S^2}{2(\sigma^2)^2} - \frac{1}{\frac{2}{n}S^2 - \sigma^2}$$

That is, a necessary condition for a value of σ^2 that maximises the penalised log-likelihood is that this derivative is zero.

An unbiased estimator is

$$\hat{\sigma}^2 = \frac{S^2}{n-k-1},$$

substituting it into the formula for the derivative gives:

$$\begin{aligned}
\left.\frac{\partial \ell^*(\boldsymbol{\alpha}, \sigma^2; \boldsymbol{y})}{\partial \sigma^2}\right|_{\sigma^2=\hat{\sigma}^2} &= -\frac{n+6}{2}\frac{1}{\hat{\sigma}^2} + \frac{S^2}{2(\hat{\sigma}^2)^2} - \frac{1}{\frac{2}{n}S^2 - \hat{\sigma}^2} \\
&= -\frac{n+6}{2}\frac{n-k-1}{S^2} + \frac{S^2}{2}\left(\frac{n-k-1}{S^2}\right)^2 - \frac{1}{\frac{2}{n}S^2 - \frac{S^2}{n-k-1}} \\
&= -\frac{(n+6)(n-k-1)}{2S^2} + \frac{(n-k-1)^2}{2S^2} - \frac{1}{S^2}\frac{n(n-k-1)}{2(n-k-1)-n}.
\end{aligned}$$

This expression is generally not equal to zero, which means that the unbiased estimator of the error variance in a linear regression with normal distributed errors does not maximise Firth's penalised log-likelihood. As argued earlier, the unbiased estimator of the error variance is a special case of a REML estimator. We can therefore conclude from this counterexample that the REML estimator is not identical to Firth's bias reduction technique and also is not a Bayes estimator with a Jeffreys prior.

A.6 REML for Generalized Linear Mixed Models

Generalised 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 models, on the one hand, and generalised 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 generalised linear models, see Firth 1993). So the relatively reassuring result about the unbiasedness of estimators for parameters in normal-linear mixed model proved in section A.3, does not necessary carry over to generalised linear mixed models.[5]

5. 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.

In the following we first present the structure of generalised linear mixed models and discuss how ML and REML estimators would work in this type of models. If $f(\boldsymbol{y}|\boldsymbol{b};\boldsymbol{\alpha},\sigma^2)$ 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 generalised linear mixed model takes the form of the integral

$$\ell(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{y}) = c - \frac{1}{2}\ln\det(\boldsymbol{\Phi}) + \ln\int f(\boldsymbol{y}|\boldsymbol{b};\boldsymbol{\alpha},\sigma^2)\exp\left[-\frac{1}{2}\boldsymbol{b}'\boldsymbol{\Phi}^{-1}\boldsymbol{b}\right]\mathrm{d}\,\boldsymbol{b} \tag{28}$$

for which a solution formula exists only if the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{b}$ is normal. In the absence of a solution formula, the integral involved in the log-likelihood function of generalised linear mixed models can only approximately be computed. 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 penalised quasi-likelihood (PQL), is that it makes it easy to translate the concept of restricted maximum likelihood to generalised linear mixed models beyond the normal-linear case. The Laplace approximation is given by

$$\ell(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{y}) \approx\ c - \frac{1}{2}\ln\det(\boldsymbol{\Phi}) + \ln f(\boldsymbol{y}|\tilde{\boldsymbol{b}};\boldsymbol{\alpha},\sigma^2) - \frac{1}{2}\ln\det\left(\tilde{\boldsymbol{K}} + \boldsymbol{\Phi}^{-1}\right) - \frac{1}{2}\tilde{\boldsymbol{b}}'\boldsymbol{\Phi}^{-1}\tilde{\boldsymbol{b}}, \tag{29}$$

where $\tilde{\boldsymbol{b}}$ maximises the integrand in equation (28) or, equivalently, its logarithm

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

and

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

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

If the conditional distribution of the response is in a exponential family, a generalised linear mixed model is characterised by the *linear predictor*

$$\eta_i = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{b},$$

the conditional mean

$$\mu_i = \mathrm{E}\left(y_i|\tilde{\boldsymbol{b}}\right),$$

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

$$\eta_i = g(\mu_i),$$

and conditional variance

$$\mathrm{Var}(y_i|\tilde{\boldsymbol{b}}) = \sigma^2 a_i v(\mu_i)$$

(where σ^2 is the *dispersion parameter*, $v(\cdot)$ is the *variance function*, and a_i are some pre-determined 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 $\boldsymbol{\Phi}$ can be estimated by maximising the Laplace approximated marginal likelihood (29) by iteratively solving the GLS equation

$$\boldsymbol{X}'\tilde{\boldsymbol{V}}^{-1}\boldsymbol{X}\hat{\boldsymbol{\alpha}} = \boldsymbol{X}'\tilde{\boldsymbol{V}}^{-1}\boldsymbol{y}^* \tag{30}$$

where $\boldsymbol{y}^*$ 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

$$\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{W}}^{-1} + \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}', \tag{31}$$

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

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

Linear mixed models are a special case of generalised 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

$$q(\boldsymbol{\theta}; \boldsymbol{y}) = -\frac{1}{2}\det(\tilde{\boldsymbol{V}}) - \frac{1}{2}(\boldsymbol{y}^* - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\boldsymbol{\theta}})\tilde{\boldsymbol{V}}^{-1}(\boldsymbol{y}^* - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\boldsymbol{\theta}}). \tag{32}$$

as an objective function, which is analogous to the log-likelihood in the normal-linear case, yet only with the linearized dependent variable $\boldsymbol{y}^*$ instead of $\boldsymbol{y}$ and $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\theta}}$ is the solution to equation (30). For a "REML-like" variant of PQL, Breslow and Clayton (1993) propose to use instead the modified objective function

$$q^*(\boldsymbol{\theta}; \boldsymbol{y}) = q(\boldsymbol{\theta}; \boldsymbol{y}) - \frac{1}{2}\det\left(\boldsymbol{X}\tilde{\boldsymbol{V}}^{-1}\boldsymbol{X}\right). \tag{33}$$

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 $\boldsymbol{b}$, Lee and Nelder (1996) developed the so-called h-likelihood technique. They introduced this technique for the estimation of the parameters of *hierarchical generalised linear models* (HGLMs), which also allow the random effect vector $\boldsymbol{b}$ to have 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 (29), the main difference being that the term $-\frac{1}{2}\ln\det\left(\tilde{\boldsymbol{K}}+\boldsymbol{\Phi}^{-1}\right)$ 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 (Rönnegård, 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).

A.7 Improved approximations of the distribution of test statistics

In the following we contrast the conventional approach to the construction of test statistics for coefficients in linear multilevel models and assumptions about their distribution based on classical maximum likelihood theory with the more accurate approaches based on a t-distribution as discussed in the main text of the article.

Standard results of maximum likelihood theory (e.g. Gourieroux and Monfort 1995a, 180ff) imply that the ML estimates $\boldsymbol{\alpha}_{\text{ML}}$ of the fixed effects coefficients in a linear multilevel model have an asymptotic normal distribution with mean $\boldsymbol{\alpha}$ and variance

$$\operatorname{AVar}(\hat{\boldsymbol{\alpha}}) = \left(-\operatorname{E}\left[\frac{\partial^2 \ell}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}\right]\right)^{-1} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1},$$

if the model satisfies certain regularity conditions and is well specified, i.e. applies to the data being analysed. (Again, $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}' + \sigma^2$ is the variance matrix of $\boldsymbol{y}$.) That is, the larger the sample, size the better can the distribution of the fixed effects coefficients be approximated by such a normal distribution. Assuming that the dimension of $\boldsymbol{\alpha}$ $k+1$, for any $k+1$-dimensional constant vector $\boldsymbol{c}$

$$z = \frac{\boldsymbol{c}'(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})}{\operatorname{SE}(\boldsymbol{c}'\hat{\boldsymbol{\alpha}})} = \frac{\boldsymbol{c}'(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})}{\sqrt{\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}}} \tag{34}$$

then has an asymptotic standard normal distribution. Further, for any constant $k+1\times h$ matrix $\boldsymbol{C}$ with full rank $h \leq k+1$ the quadratic form

$$\begin{aligned} U &= (\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})'\boldsymbol{C}(\operatorname{AVar}(\boldsymbol{C}'\hat{\boldsymbol{\alpha}}))^{-1}\boldsymbol{C}'(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}) \\ &= (\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})'\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{C})^{-1}\boldsymbol{C}'(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}) \end{aligned} \tag{35}$$

has an asymptotic χ^2 distribution with h degrees of freedom.

Assuming these results, linear hypotheses of the form

$$\boldsymbol{c}'\boldsymbol{\alpha} = d$$

are typically tested using the Wald t-test statistic

$$t = \frac{\boldsymbol{c}'\hat{\boldsymbol{\alpha}} - d}{\widehat{\mathrm{SE}}(\boldsymbol{c}'\hat{\boldsymbol{\alpha}})} = \frac{\boldsymbol{c}'\hat{\boldsymbol{\alpha}} - d}{\sqrt{\boldsymbol{c}'(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}}} \tag{36}$$

where $\hat{\boldsymbol{V}} = \boldsymbol{Z}\boldsymbol{\Phi}(\hat{\boldsymbol{\theta}})\boldsymbol{Z}' + \hat{\sigma}^2$ is an ML or REML estimate of $\boldsymbol{V}$.

The consistency of ML and REML estimators implies that $\hat{\boldsymbol{V}}$ converges in probability to $\boldsymbol{V}$, so multilevel software often ignores the variation in *estimated* standard errors and computes p-values on the assumption that t also has an (asymptotic) standard normal distribution. Similarly, linear hypotheses of the form

$$\boldsymbol{C}'\boldsymbol{\alpha} = \boldsymbol{d}$$

are typically tested using the quadratic Wald test statistic

$$W = (\boldsymbol{C}'\hat{\boldsymbol{\alpha}} - \boldsymbol{d})'(\boldsymbol{C}'\boldsymbol{X}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\boldsymbol{C})^{-1}(\boldsymbol{C}'\hat{\boldsymbol{\alpha}} - \boldsymbol{d}) \tag{37}$$

for which often an asymptotic χ^2 distribution is assumed (see e.g. Gourieroux and Monfort 1995b, 81ff).

The test statistic in equation (36) is of particular practical importance since it used as "test of statistical significance" of individual coefficients in $\hat{\boldsymbol{\alpha}}$, in which case $\boldsymbol{c}$ is a vector with one element equal to unity and the other elements equal to zero and d being equal to zero. Confidence intervals are typically constructed using t as a pivotal quantity.

When is the sample size large enough so that asymptotic normality can be safely assumed for the test statistic t? Our simulation results reported in the main text indicate that for two-level models with small numbers m of upper-level units, Wald t-tests based on the assumption of asymptotic normality may lead to anti-conservative results, i.e. too narrow confidence intervals based on the t-statistic or, equivalently, to small p-values and thus potentially incorrect results of test of statistical significance. Apparently, it is not the total sample size n that is relevant for the distribution of the test statistics of certain coefficients, but the number m of upper-level units.

Our results also show that one can do better, by assuming a t-distribution with the appropriate degrees of freedom instead. In the following we discuss two cases where the distribution of Wald t-test statistics can be derived exactly for finite samples. In the first case, the correct degrees of freedom of the t-distribution increase with the total sample size n, but in the second case, the correct degrees of freedom do not increase with n, but with the number of upper-level units m. We use this second case as a motivation of the so-called $m - l - 1$ rule, for which we discuss more general conditions for applicability. Finally, we discuss more general approaches to determining the appropriate degrees of freedom. We provide a brief

summary of the approach of Giesbrecht and Burns (1985) to the single-constraint case. (See Schaalje, McBride, and Fellingham 2002, for a more extensive discussion).

To illustrate the role of Student's t-distribution for inference in (especially linear) multilevel models, we first consider the simple case of a linear regression with normal errors. Linear regression models with normal errors are one of the rare instances where the distribution of Wald test statistics (under the null hypothesis) and the distribution of pivotal quantities for the construction of confidence intervals can be derived exactly. Such a model can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ has a normal distribution with zero mean and variance $\sigma^2\boldsymbol{I}_n$ or equivalently the elements of $\boldsymbol{\epsilon}$ are i.i.d. normal distributed each with zero mean and variance σ^2.

When this model applies, the ML estimator is the OLS estimator, which can be expressed as

$$\hat{\boldsymbol{\alpha}}_{\text{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y} = \boldsymbol{\alpha} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{\epsilon}$$

If $\boldsymbol{X}$ has full column rank $k+1$ then $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{\epsilon}$ has a normal distribution with expectation $\boldsymbol{0}$ and variance matrix $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. $\hat{\boldsymbol{\alpha}}_{\text{OLS}}$ therefore also has a normal distribution with expectation $\boldsymbol{\alpha}$ and the same variance matrix. The unbiased estimator of the error variance is

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\text{OLS}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\text{OLS}})}{n-k-1} = \frac{S^2}{n-k-1}.$$

The standardised sum of squares S^2/σ^2

$$\begin{aligned}
\frac{S^2}{\sigma^2} &= \frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\text{OLS}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\text{OLS}}) \\
&= \frac{1}{\sigma^2}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} \\
&= \frac{1}{\sigma^2}\boldsymbol{\epsilon}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{\epsilon}. \\
&= \frac{\boldsymbol{\epsilon}'}{\sigma}(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\frac{\boldsymbol{\epsilon}}{\sigma}.
\end{aligned}$$

has a χ^2 distribution with $n-k-1$ degrees of freedom, because $\boldsymbol{\epsilon}/\sigma$ has a standard normal distribution and $\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is an idempotent matrix of rank $n-k-1$ (Greene 2012, 1084).

Consider now the quantity

$$z = \frac{\boldsymbol{c}'(\hat{\boldsymbol{\alpha}}_{\text{OLS}} - \boldsymbol{\alpha})}{\text{SE}(\boldsymbol{c}'\hat{\boldsymbol{\alpha}}_{\text{OLS}})} = \frac{\boldsymbol{c}'(\hat{\boldsymbol{\alpha}}_{\text{OLS}} - \boldsymbol{\alpha})}{\sqrt{\boldsymbol{c}'\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} = \frac{\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{\epsilon}}{\sqrt{\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}}$$

The numerator has a normal distribution with zero expectation and variance $\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}$. The denominator is the square root of this variance, hence z has a normal distribution with

zero mean and unit variance. Since the distribution of z is functionally independent from the model parameters $\boldsymbol{\alpha}$ and σ^2, z is a pivotal quantity. For the usual Wald t-test statistic we get

$$t = \frac{\boldsymbol{c}'(\hat{\boldsymbol{\alpha}}_{\text{OLS}} - \boldsymbol{\alpha})}{\widehat{\text{SE}}(\boldsymbol{c}'\hat{\boldsymbol{\alpha}}_{\text{OLS}})} = \frac{\boldsymbol{c}'(\hat{\boldsymbol{\alpha}}_{\text{OLS}} - \boldsymbol{\alpha})}{\sqrt{\boldsymbol{c}'\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} = \frac{\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{\epsilon}}{\sqrt{\hat{\sigma}^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} = \frac{z}{\sqrt{\frac{S^2/\sigma^2}{n-k-1}}}.$$

The numerator is a standard normal distributed random variable, the denominator is the square root of a χ^2-distributed random variable divided by its $n-k-1$ degrees of freedom. Such a ratio is known to have a Student's t-distribution with $n-k-1$ degrees of freedom (Greene 2012, 1062). If $n-k-1$ is small, either because n is small or because k is large, the distribution of t will markedly differ from a standard normal: Its variance will be larger, i.e. $\text{Var}(t) = 1 + 2/(n-k-3)$ and it will have a higher kurtosis. One of the consequences will be that acknowledging such a t-distribution will lead to more conservative tests and p-values than the assumption of a standard normal distribution.

To illustrate the consequences of clustering for the distribution of test statistics, we now consider the extreme case of a "second-level only" model with random intercepts on the second level and no individual-level errors. Such a model corresponds to the two-level case with an intraclass-correlation of $\rho = 1$. Such a model can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{b},$$

where $\boldsymbol{X}$ is a $n \times (k+1)$ matrix, $\boldsymbol{Z}$ is an $n \times m$ matrix composed of zeroes and ones and $\boldsymbol{b}$ is an m-dimensional random vector with multivariate normal distribution with i.i.d. elements each with variance ϕ. The between- and within-variation of $\boldsymbol{y}$ can be separated using the matrices $\boldsymbol{M}_{\boldsymbol{Z}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$: and $\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' = \boldsymbol{Z}\boldsymbol{M}_{\boldsymbol{Z}}$:

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{y} \\
\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y} &= \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{b} \\
(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{y} &= (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{X}\boldsymbol{\alpha}
\end{aligned}$$

because $\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Z} = \boldsymbol{Z}$. This implies that $(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{y}$ is not a random variable but a constant.

We consider now the case that matrix $\boldsymbol{X}$ can be separated into $l+1$ columns collected into $\boldsymbol{X}_{\text{b}}$ and $k-l$ columns collected into $\boldsymbol{X}_{\text{w}}$ so that

$$\begin{aligned}
\boldsymbol{Z}'\boldsymbol{X}_{\text{w}} = 0 &\Rightarrow \boldsymbol{P}_{\boldsymbol{Z}}'\boldsymbol{X}_{\text{w}} = 0 \\
\boldsymbol{P}_{\boldsymbol{Z}}'\boldsymbol{X}_{\text{b}} = \boldsymbol{X}_{\text{b}} &\Rightarrow (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{X}_{\text{b}} = 0,
\end{aligned}$$

which means that the columns of $\boldsymbol{X}_{\text{w}}$ represent only those covariates that vary *only within* groups, while columns of $\boldsymbol{X}_{\text{b}}$ represent covariates that vary *only between* groups and the constant term of the linear model. We define the subvectors $\boldsymbol{\alpha}_{\text{b}}$ and $\boldsymbol{\alpha}_{\text{b}}$ of $\boldsymbol{\alpha}$ to be composed of the corresponding columns of $\boldsymbol{X}$ such that

$$\boldsymbol{X}\boldsymbol{\alpha} = \boldsymbol{X}_{\text{b}}\boldsymbol{\alpha}_{\text{b}} + \boldsymbol{X}_{\text{w}}\boldsymbol{\alpha}_{\text{w}}.$$

We thus get

$$\boldsymbol{P_Z y} = \boldsymbol{X}_\mathrm{b}\boldsymbol{\alpha}_\mathrm{b} + \boldsymbol{Z b} \tag{38}$$

$$(\boldsymbol{I} - \boldsymbol{P_Z})\boldsymbol{y} = \boldsymbol{X}_\mathrm{w}\boldsymbol{\alpha}_\mathrm{w} \tag{39}$$

Equation (39) does not contain a random element, hence an "estimate" of $\boldsymbol{\alpha}_\mathrm{w}$ can be computed by solving an over-determined linear system of equations, for which the solution exists by assumption (we are assuming here that the model is correct):

$$(\boldsymbol{X}'_\mathrm{w}\boldsymbol{X}_\mathrm{w})^{-1}\boldsymbol{X}'_\mathrm{w}(\boldsymbol{I} - \boldsymbol{P_Z})\boldsymbol{y} = \boldsymbol{\alpha}_\mathrm{w}$$

Because $\boldsymbol{M_Z P_Z} = \boldsymbol{M_Z}$ and $\boldsymbol{M_Z Z} = \boldsymbol{I}_m$, equation (38) is equivalent to a group-level regression model:

$$\begin{aligned} \boldsymbol{P_Z y} &= \boldsymbol{X}_\mathrm{b}\boldsymbol{\alpha}_\mathrm{b} + \boldsymbol{Z b} \Rightarrow \\ \boldsymbol{M_Z y} &= \boldsymbol{M_Z X}_\mathrm{b}\boldsymbol{\alpha}_\mathrm{b} + \boldsymbol{b} \quad \text{or} \\ \bar{\boldsymbol{y}} &= \bar{\boldsymbol{X}}_\mathrm{b}\boldsymbol{\alpha}_\mathrm{b} + \boldsymbol{b} \end{aligned}$$

where $\bar{\boldsymbol{y}} = \boldsymbol{M_Z y}$ and $\bar{\boldsymbol{X}}_\mathrm{b} = \boldsymbol{M_Z X}_\mathrm{b}$. The j-th element of the m-dimensional vector $\bar{\boldsymbol{y}}$ contains the mean of all elements of $\boldsymbol{y}$ for which the elements of the j-th column of $\boldsymbol{Z}$ equal unity and j-th row of $\bar{\boldsymbol{X}}_\mathrm{b}$ contains the means of the corresponding rows of $\bar{\boldsymbol{X}}_\mathrm{b}$. The ML estimator in this case is again an OLS estimator:

$$\hat{\boldsymbol{\alpha}}_\mathrm{b} = (\bar{\boldsymbol{X}}'_\mathrm{b}\bar{\boldsymbol{X}}_\mathrm{b})^{-1}\bar{\boldsymbol{X}}'_\mathrm{b}\bar{\boldsymbol{y}}$$

An unbiased estimator of the variance parameter ϕ is:

$$\hat{\phi} = \frac{(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}_\mathrm{b}\hat{\boldsymbol{\alpha}}_\mathrm{b})'(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}_\mathrm{b}\hat{\boldsymbol{\alpha}}_\mathrm{b})}{m - l - 1}$$

The Wald t-statistic for $\boldsymbol{\alpha}_\mathrm{b}$ becomes:

$$t = \frac{\boldsymbol{c}'(\hat{\boldsymbol{\alpha}}_\mathrm{b} - \boldsymbol{\alpha}_\mathrm{b})}{\widehat{\mathrm{SE}}(\boldsymbol{c}'\hat{\boldsymbol{\alpha}}_\mathrm{b})} = \frac{\boldsymbol{c}'(\bar{\boldsymbol{X}}'_\mathrm{b}\bar{\boldsymbol{X}}_\mathrm{b})^{-1}\bar{\boldsymbol{X}}'_\mathrm{b}\boldsymbol{b}}{\sqrt{\hat{\phi}\boldsymbol{c}'(\bar{\boldsymbol{X}}'_\mathrm{b}\bar{\boldsymbol{X}}_\mathrm{b})^{-1}\boldsymbol{c}}}.$$

For the same reasons as above, this statistic has a Student's t-distribution with $m - l - 1$ degrees of freedom. This is a direct example of the $m - l - 1$ rule. However, there may be other instances where this $m - l - 1$ may be used to at least approximate the distribution of a Wald t-test statistic. The rationale behind the rule is easiest to grasp in the multiple equations formulation of a two-level model with cross-level interactions as described above by equations (2), (3), and (4). If one interprets equations (3) and (4) as regressions with a_{0j} and a_{1j} as dependent variables then a t-distribution with $m - l - 1$ degrees of freedom (where $j = 1, \ldots, m$ and $l = 2$) would be appropriate for testing hypotheses about α_{00}, α_{01}, α_{10}, or α_{11}. Such an interpretation will be particularly suited for cases where the variance of ϵ_{ij} is small relative to the variances of b_{0j} and b_{1j}.

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.

Various other techniques to provide a more flexible approximations to the distribution of test statistics been developed in the literature. They are typically adaptations or generalisations of Satterthwaite's (1946) method. In contrast to the $m - l - 1$ rule, these approximations estimate the degrees of freedom from the data and are therefore far more widely applicable. In particular, such techniques can provide the approximate degrees of freedom for complex multilevel designs, for which the $m - l - 1$ rule is not applicable (e.g. cross-classified structures, structures with more than two levels, etc.). As an example we discuss the adaption of Satterthwaite's method to single-constraint tests by Giesbrecht and Burns (1985), which has later been extended to multiple-constraints F-tests by Fai and Cornelius (1996). (See Schaalje, McBride, and Fellingham 2002, for a more extensive discussion). To provide the appropriate degrees of freedom for the approximation the distribution of the test statistic

$$t = \frac{\boldsymbol{c}'(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})}{\widehat{\mathrm{SE}}(\boldsymbol{c}'\hat{\boldsymbol{\alpha}})} = \frac{\boldsymbol{c}'(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})}{\sqrt{\boldsymbol{c}'(\boldsymbol{X}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}}},$$

Giesbrecht and Burns (1985) propose as degrees of freedom of the approximating t-distribution

$$\mathrm{df} = \frac{2(\boldsymbol{c}'(\boldsymbol{X}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c})^2}{\mathrm{Var}(\boldsymbol{c}'(\boldsymbol{X}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c})}$$

where $\mathrm{Var}(\boldsymbol{c}'(\boldsymbol{X}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c})$ is approximated by the multivariate delta method:

$$\mathrm{Var}(\boldsymbol{c}'(\boldsymbol{X}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}) \approx \frac{\partial \boldsymbol{c}'(\boldsymbol{X}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}}{\partial \boldsymbol{\theta}'} \mathrm{AVar}(\hat{\boldsymbol{\theta}}) \frac{\partial \boldsymbol{c}'(\boldsymbol{X}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}}{\partial \boldsymbol{\theta}}.$$

Kenward and Roger (1997) provide a further improved, yet computationally more complex approximation to the distribution of t-test and F-test statistics.

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 generalised 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 *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.

B Further details on the simulation study

B.1 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 40 and 41 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 40) the context-level variable w_j has a simple additive effect on the (lower-level) outcome y_{ij}:

$$y_{ij} = 0 + 0.25x_{ij} + 0.2w_j + b_{1j} + \epsilon_{ij} \tag{40}$$

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 B4), 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 ϵ_{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 B2). 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

$$y_{ij} = 0 + 0.25x_{ij} + 0.2w_j + 0.1x_{ij}w_j + b_{1j} + b_{2j}x_{ij} + \epsilon_{ij} \tag{41}$$

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 40 and 41 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 ϵ_{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 optimiser *bobyqa*

(Bates 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 generalised 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.

B.2 Additional results

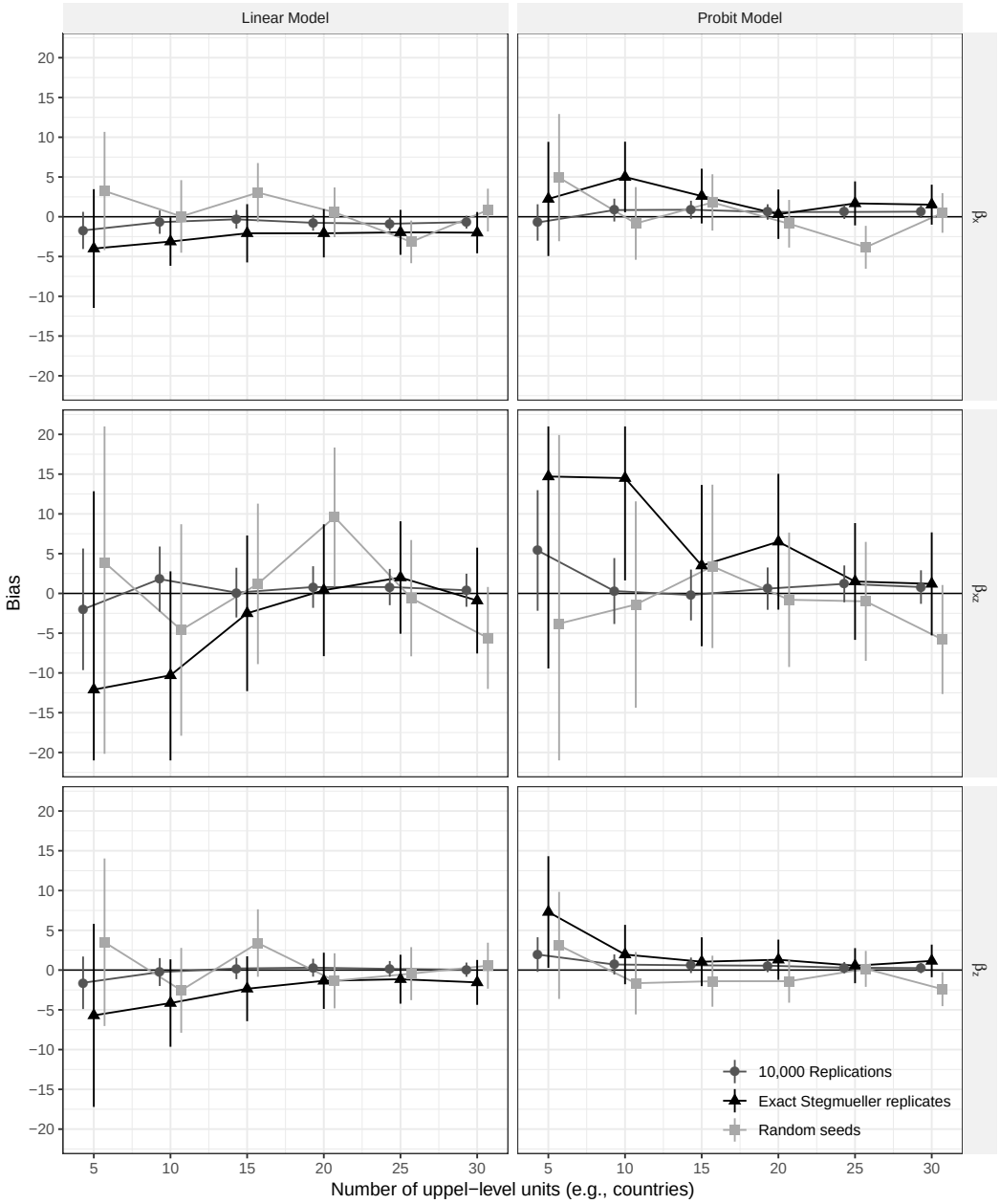
For reasons of space restrictions, only a subset of our Monte Carlo simulation results are reported in the main text of our article. Thus Figure 1 in the main text shows the simulated distribution of estimates of a context effect coefficient, with 1,000, and 10,000 replications using Stegmueller's settings for the random seed, as well as 1,000 replications using random seeds that vary across settings. Figure B1 shows the distribution of the estimates of all constitutive terms of models that extend those of Figure 1 by a cross-level interaction model: the main effect coefficient β_x of an individual-level covariate, the main effect coefficient β_z of a contextual covariate, and the coefficient β_{xz} of a cross-level interaction term. The results are essentially the same as in Figure 1: The apparent "bias" of coefficient estimates is closer to zero for larger Monte Carlo sample sizes (i.e. 10,000 replications) and with independent random seeds is as often larger than zero as it is smaller than zero.

Figure 2 in the main text shows simulation results about the performance of ML and REML estimators (or quasi-ML and quasi-REML estimators in case of probit models) of a random intercept variance parameter in multilevel models with a contextual effect but no cross-level interaction and an intra-class correlation (ICC) equal to 0.1. Figure B2 shows the results for models extended by a cross-level interaction and various settings of the ICC. Again, the depicted results corroborate those of the corresponding figure in the main text: Using REML instead of ML greatly reduces the bias in the estimates of the variances of random intercepts and random slopes, to a degree at which it is almost negligible.

In similar way as before, Figure B3 presents results from the extended simulation study with cross-level interactions. It corresponds to Figure 3 in the main text and shows how the choice of the estimator (ML or REML in case of linear models; quasi-ML or quasi-REML in case of probit models) for the variance parameters and the choice of the assumed sampling distribution of test statistics contribute to the accuracy of confidence intervals for coefficients. In contrast to Figure 3 it describes the performance of confidence intervals for the coefficients of main and interaction effects, but like Figure 3 it shows the results only for an ICC value of 0.1. It mirrors the results of that figure in that neither the choice of the estimator nor of the sampling distribution of test-statistics is sufficient for accurate inference and that both have to be chosen correctly.

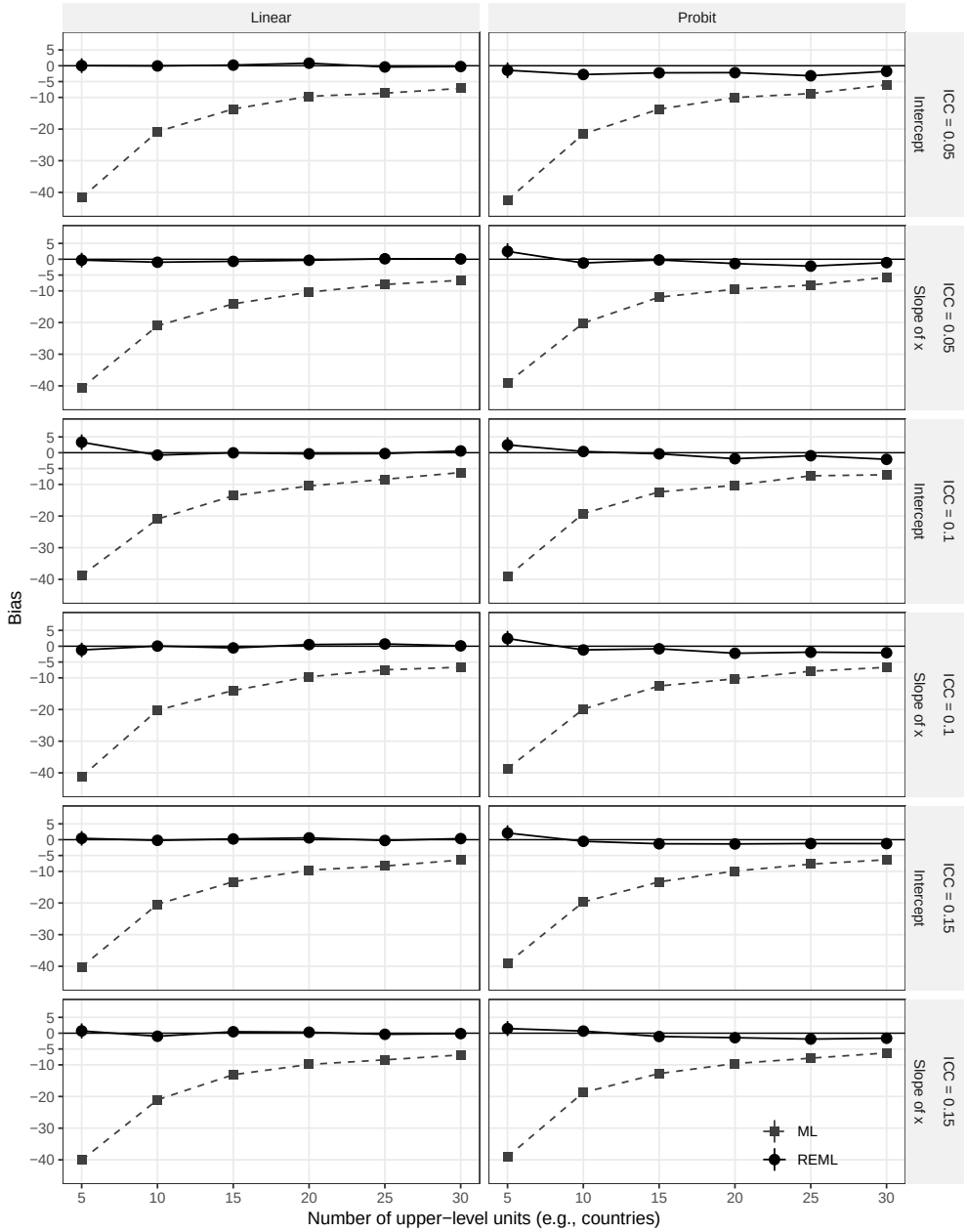
Finally, Figure B4 compares the $m - l - 1$ rule and the Satterthwaite approximation in terms of accuracy of the confidence intervals for the main and interaction effect coefficients in the cross-level interaction models that were already the topics of the previous four figures. However, it covers two variants of the simulation study the results of which were shown in the previous figure: In the variant the lower-level covariate has no compositional differences between upper-level units, in the second variant 50% of the variance of the covariate is between upper-level units. Like Figure 4 in the main text, it demonstrates that the $m - l - 1$ rule and the Satterthwaite approximation give results that are almost indistinguishable from each other.

Figure B1: Performance of Point Estimates of All Three Constitutive Terms of a Cross Level Interaction in Multilevel Linear and Probit Models



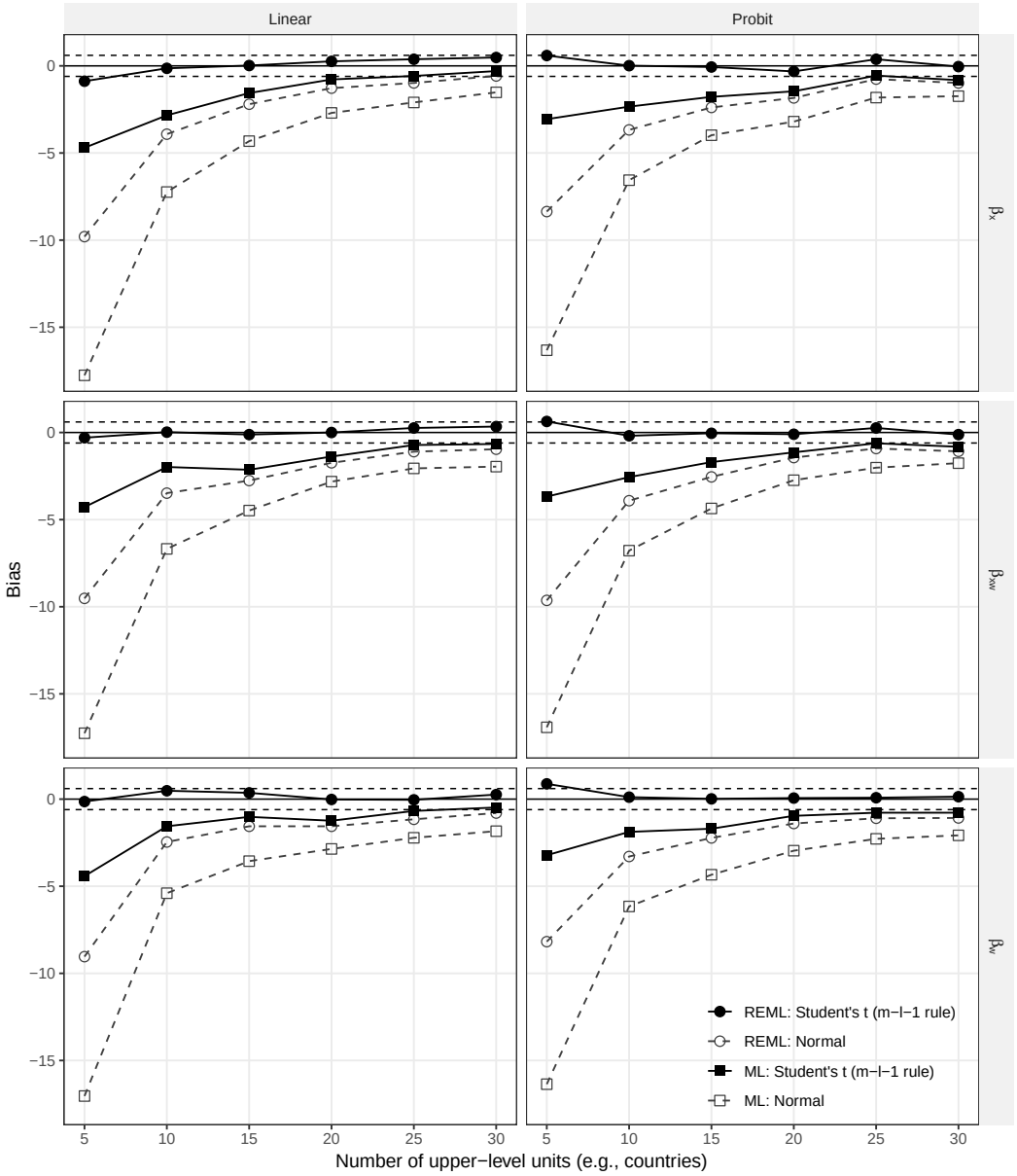
Note: The figure displays relative biases of maximum likelihood point estimates (in % of the true effect size). Vertical lines depict 95% Monte Carlo confidence intervals for these results. Because the results are displayed as bias in percent (i.e., average deviation from the true effect measured as percent of the true effect), we estimate the 95% confidence intervals as: $\text{CI}_{95} = 100 \cdot \left(\left(\bar{b} \pm 1.96\text{SD}(b) \Big/ \sqrt{M}\right) - \beta\right) \Big/ \beta$, where b is the coefficient estimate of interest, β is the true effect, and M is the number of Monte Carlo trials). The horizontal zero line denotes the reference of no bias. Black triangles replicate the results presented the left column ("Estimate") of Figure 5 on page 757 in Stegmueller (2013). We additionally present two modifications of Stegmueller's analysis. The first (black circles) increases the number of replications from 1,000 to 10,000, leaving everything else 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.

Figure B2: Performance of Estimators of Random Intercept Variances and Random Slopes in Multilevel Linear and Probit Models



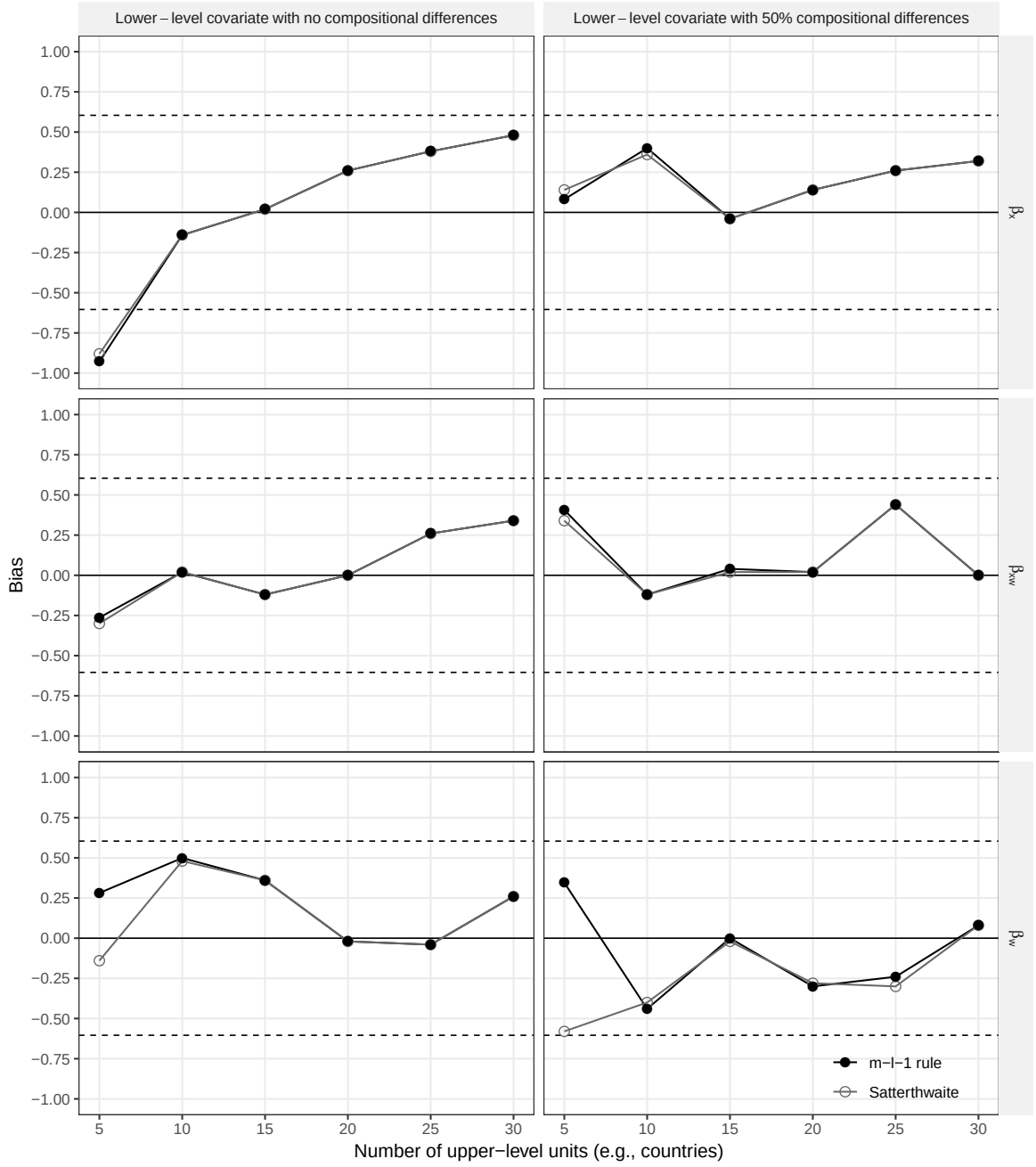
Note: The figure displays relative bias (in % of true size) in variance estimates for the random intercept. Vertical lines depict 95% confidence intervals. We estimate the 95% confidence intervals as: $\text{CI}_{95} = 100 \cdot \left(\left(\bar{\hat{\phi}} \pm 1.96\text{SD}(\hat{\phi}) \Big/ \sqrt{M} \right) - \phi \right) \Big/ \phi$, where $\bar{\hat{\phi}}$ is the average of the estimates of the random intercept variance, ϕ is the true random intercept variance, and M is the number of Monte Carlo trials). The horizontal zero line denotes the reference of no bias. This figure has no correspondence in Stegmueller 2013.

Figure B3: Performance of Confidence Intervals of All Three Constitutive Terms of a Cross Level Interaction in Multilevel Linear and Probit Models



Note: The figures shows percentage point deviations of actual coverage rates from the nominal value of 95%. The horizontal zero line denotes the reference of no bias (i.e., actual equals nominal coverage rate). The dashed horizontal lines indicate 95% test intervals and are constructed as follows: $\text{TI}_{95} = 0 \pm 100 \cdot 1.96\sqrt{0.95(1-0.95)/M}$, where M is the number of Monte Carlo trials. For an accurate estimator of the 95% confidence interval (i.e., one that has an actual coverage rate of 95%), the estimated actual coverage rate based should fall into TI_{95} in 95% of the time. It is in this sense that estimated coverage rates falling outside TI_{95} constitute statistically significant evidence against an accurate coverage rate. 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 B4: Performance of Degrees of Freedom Approximations for the Sampling distribution of Test Statistics for All Three Constitutive Terms of a Cross Level Interaction in Linear Multilevel 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. This figure has no correspondence in Stegmueller 2013.

C Improved Methods for Inference about Multilevel Models with Few Clusters in *R*

There are two major packages that can be used in multilevel analysis in *R*: package *nlme* (Pinheiro et al. 2018), which usually is part of a regular *R* installation due to its status as a "recommended" package, and package *lme4* (Bates et al. 2015), which needs to be separately installed, usually from the "Comprehensive *R* Archive Network" (CRAN). The following two subsections describe how the improved methods for inference about multilevel models–using the REML estimator and using a t-distribution for confidence intervals, hypothesis tests and p-values–with few clusters can be applied with each of these two packages. In both instances, the data and model of the empirical application from Steenbergen and Jones (2002) discussed at the end of the main part of the paper. We will not discuss the data preparation for this application, but an *R*-script for this is provided with the replication material.[6] The data is loaded into the *R* session using the code:

```
load("steenbergen-jones-data.RData")
```

This assumes that the data file `"steenbergen-jones-data.RData"` is located in the working directory of the current *R* session. The data file contains a single data frame, with variables:

Variable name	Description
`support`	Respondents' support for the EU
`tenurez`	A county's tenure within the EU (standardized) to zero mean and unit variance
`tradez`	A county's within-EU trade (standardized)
`gdpz`	A county's GDP per capita (standardized)
`inflz`	A county's Inflation (standardized)
`inclow`	Dummy variable indicating whether the respondent is in the lowest income quartile
`inchi`	Dummy variable indicating whether the respondent is in the the highest income quartile
`income`	A factor variable with four levels for respondents' membership in an income quartile (This variable is not in the *Stata*-version of the data set)
`lright`	A variable that contains respondents' left-right self-placement
`oplead`	An indicator of opinion leadership (centred)
`male`	A dummy variable for male gender

For origin and further details about the data, see the ReadMe.html of our replication package as well as Steenbergen and Jones (2002).

6. This *R*-script generates the data in the same way as the *Stata* "genData.do", which generates the data for our replication of Steenbergen and Jones discussed in the article and used in the *Stata* illustration in the following section.

C.1 Improved Inference Methods with the *nlme* Package

The following code is used to estimate Steenbergen and Jones' (2002) two-level model using the package *nlme*. It involves the function `lme()`, which is called with a specification of the dependent and independent variables as first argument, an argument tagged `method=` to select the estimator (in this case maximum likelihood), an argument tagged `random=` that specifies the random effects structure, and an argument tagged `data=` that specifies the data frame that contains the dependent and independent variables:

```
sj.ml <- lme(support ~ tenurez + tradez + gdpz + inflz
                       + inclow + inchi + lright + olead + male + age,
             method = "ML",
             random = ~1 | country,
             na.action = na.exclude, # Deal with missing values
             data   = steenbergen_jones_data)
```

The argument `random=~1|country` indicates that the model should include a random intercept that varies with the levels (values) of the factor variable `country`. To select the ML estimator we have to explicitly ask for it, because REML is the default. To get information about coefficient estimates, standard errors, test statistics and *p*-values, we (as usually) apply the `summary()` function:

```
summary(sj.ml)
```

This leads to the output:

```
Linear mixed-effects model fit by maximum likelihood
 Data: steenbergen_jones_data
       AIC      BIC    logLik
  43700.04 43794.74 -21837.02

Random effects:
 Formula: ~1 | country
        (Intercept) Residual
StdDev:    0.425438 1.831069

Fixed effects: support ~ tenurez + tradez + gdpz + inflz + inclow + inchi +
     lright + olead + male + age
                Value  Std.Error    DF   t-value p-value
(Intercept)  5.421258 0.14124679 10757  38.38146  0.0000
tenurez      0.235522 0.14068641     9   1.67409  0.1284
tradez       0.329053 0.13150993     9   2.50212  0.0337
gdpz        -0.498780 0.27512221     9  -1.81294  0.1033
inflz        0.338609 0.16509114     9   2.05104  0.0705
inclow      -0.130189 0.04889183 10757  -2.66279  0.0078
inchi        0.089671 0.04542176 10757   1.97419  0.0484
```

```
lright       0.034570 0.00883003 10757   3.91509  0.0001
olead        0.105053 0.02030881 10757   5.17280  0.0000
male1        0.025821 0.03568690 10757   0.72355  0.4694
age         -0.011110 0.00107387 10757 -10.34606  0.0000
 Correlation:
        (Intr) tenurz tradez gdpz   inflz  inclow inchi  lright olead  male1
tenurez -0.132
tradez   0.129 -0.280
gdpz     0.370 -0.503  0.365
inflz    0.271 -0.362  0.377  0.701
inclow  -0.044  0.001  0.005 -0.006  0.001
inchi   -0.092  0.003  0.012 -0.001  0.005  0.224
lright  -0.239  0.008 -0.002  0.001  0.002  0.015 -0.028
olead    0.019 -0.004 -0.001 -0.008 -0.011  0.057 -0.088  0.042
male1   -0.126 -0.002 -0.001  0.001  0.000  0.046 -0.028 -0.001 -0.118
age     -0.315  0.000  0.007 -0.002  0.003 -0.119  0.067 -0.059  0.031 -0.023

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-3.0467052 -0.6444306  0.1088729  0.7577502  2.4460038

Number of Observations: 10777
Number of Groups: 14
```

As can be seen, the `summary()` function from the *nlme* package employs the so-called within-between heuristic (cf., Schaalje, McBride, and Fellingham 2002), for the selection of the degrees of freedom for the test statistics of of the coefficients. The within-between heuristic equals the $m - l - 1$ heuristic for the case of direct context effects. Beware however, that the within-between heuristic differs fundamentally from the $m - l - 1$ heuristic for the case of cross-level interactions.

Overall then, to make use of the full set of methodological improvements suggested in our paper for the case of direct context effects, one only needs to use the optional argument `method=` appropriately to `"REML"` or simply to drop it, because `"REML"` is actually the default setting:

```
sj.reml <- lme(support ~ tenurez + tradez + gdpz + inflz
                       + inclow + inchi + lright + olead + male + age,
            random = ~1 | country,
            na.action = na.exclude, # Deal with missing values
            data   = steenbergen_jones_data)
summary(sj.reml)
```

The output is

```
Linear mixed-effects model fit by REML
 Data: steenbergen_jones_data
       AIC      BIC    logLik
  43748.38 43843.07 -21861.19

Random effects:
 Formula: ~1 | country
        (Intercept) Residual
StdDev:   0.5332954 1.831578

Fixed effects: support ~ tenurez + tradez + gdpz + inflz + inclow + inchi +
     lright + olead + male + age
                Value Std.Error    DF   t-value p-value
(Intercept)  5.420907 0.1699667 10757  31.89393  0.0000
tenurez      0.235291 0.1754975     9   1.34071  0.2129
tradez       0.329125 0.1640793     9   2.00589  0.0758
gdpz        -0.498034 0.3432000     9  -1.45115  0.1807
inflz        0.338932 0.2059839     9   1.64543  0.1343
inclow      -0.129702 0.0488821 10757  -2.65336  0.0080
inchi        0.090128 0.0454130 10757   1.98464  0.0472
lright       0.034653 0.0088285 10757   3.92509  0.0001
olead        0.105046 0.0203059 10757   5.17317  0.0000
male1        0.025872 0.0356789 10757   0.72513  0.4684
age         -0.011104 0.0010736 10757 -10.34241  0.0000
 Correlation:
        (Intr) tenurz tradez gdpz   inflz  inclow inchi  lright olead  male1
tenurez -0.135
tradez   0.135 -0.281
gdpz     0.383 -0.503  0.365
inflz    0.282 -0.363  0.378  0.701
inclow  -0.036  0.001  0.004 -0.005  0.001
inchi   -0.077  0.002  0.010 -0.001  0.004  0.224
lright  -0.199  0.007 -0.002  0.001  0.002  0.015 -0.028
olead    0.016 -0.003 -0.001 -0.006 -0.009  0.057 -0.088  0.042
male1   -0.105 -0.002 -0.001  0.001  0.000  0.046 -0.028 -0.001 -0.118
age     -0.262  0.000  0.005 -0.001  0.002 -0.119  0.067 -0.059  0.031 -0.023

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-3.0505486 -0.6445138  0.1093930  0.7574557  2.4478278

Number of Observations: 10777
Number of Groups: 14
```

The difference between ML and REML in this case is that with REML the standard errors are estimated slightly larger, so that the p-values have twice the size of their counterparts with ML.

As mentioned above, the within-between heuristic fails in the case of cross-level interactions. This is because the within-between rule equals the $m - l - 1$ rule only in the absence of within-cluster variation of a predictor variable. That is, contextual predictors, such as GDP, do not vary within upper-level units but only between them. However, in case of a cross-level interaction, i.e. in interaction between a lower-level predictor variable and an upper-level independent variable, the values of the corresponding product term vary within the upper-level units. Therefore, the within-between rule will treat the interaction term and the corresponding lower-level main effect as lower-level predictors. Our theoretical discussion clarifies, however, that both the cross-level interaction term and the lower-level main effect should be treated as contextual predictors, implying that the $m - l - 1$ rule applies. Our Monte Carlo simulation results corroborate this conclusion and vice versa question the accuracy of the within-between heuristic. The Kenward-Roger method might be of help in this case, but the package *pbkrtest* (Halekoh and Højsgaard 2014), which provides for the computation of the Kenward-Roger degrees of freedom is not applicable to models estimated with the *nlme* package. The situation is different for models estimated with the *lme4* package, which is discussed in the next subsection.

C.2 Improved Inference Methods with the *lme4* Package

The package *lme4* is not included in a standard installation of *R* and therefore must be installed as an add-on package. The "canonical" source for add-on packages is the "Comprehensive *R* Archive Network" (CRAN) which has the web address `https://cran.r-project.org`. Usually (in particular if one uses the *RStudio* `https://rstudio.com` interface for *R*), to install *lme4* it suffices to run

```
install.packages("lme4")
```

which causes this package to be installed, along with those packages on which *lme4* depends (i.e., the packages *Matrix*, *Rcpp*, and *RcppEigen*).

The code needed to get ML-estimates Steenbergen and Jones' (2002) multilevel model with the *lme4* package is slightly different from the code for estimating it with the *nlme* package:

```
sj4.ml <- lmer(support ~ tenurez + tradez + gdpz + inflz
                         + inclow + inchi + lright + olead + male + age
                         + (1 | country),
               REML = FALSE,
               data = steenbergen_jones_data)
```

Here, the random effects structure is specified differently from the call to `lme()` of the *nlme* package. Instead of a `random=` argument, the random effects specification is part of the

model formula, where the term `(1 | country)` indicates that the model should include random intercepts that vary across the levels (values) of the factor variable `country`. Again, REML is the default setting, so in order to get ML estimates, we have to explicitly add the optional argument `REML = FALSE`.

Again, with the function `summary()` we get model estimates, along with standard errors and test statistics. However, in contrast to *nlme*, the `summary()` method function for objects created with the *lme4* package does not report degrees of freedom and *p*-values:

```
summary(sj4.ml)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: support ~ tenurez + tradez + gdpz + inflz + inclow + inchi +
    lright + olead + male + age + (1 | country)
   Data: steenbergen_jones_data

     AIC      BIC   logLik deviance df.resid
 43700.0  43794.7 -21837.0  43674.0    10764

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0467 -0.6444  0.1089  0.7578  2.4460

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.181    0.4254
 Residual             3.353    1.8311
Number of obs: 10777, groups:  country, 14

Fixed effects:
             Estimate Std. Error t value
(Intercept)  5.421258   0.141175  38.401
tenurez      0.235522   0.140615   1.675
tradez       0.329053   0.131443   2.503
gdpz        -0.498780   0.274982  -1.814
inflz        0.338609   0.165007   2.052
inclow      -0.130189   0.048867  -2.664
inchi        0.089671   0.045399   1.975
lright       0.034570   0.008826   3.917
olead        0.105053   0.020298   5.175
male1        0.025821   0.035669   0.724
age         -0.011110   0.001073 -10.351

Correlation of Fixed Effects:
        (Intr) tenurz tradez gdpz   inflz  inclow inchi  lright olead  male1
tenurez -0.132
tradez   0.129 -0.280
```

```
gdpz     0.370 -0.503  0.365
inflz    0.271 -0.362  0.377  0.701
inclow  -0.044  0.001  0.005 -0.006  0.001
inchi   -0.092  0.003  0.012 -0.001  0.005  0.224
lright  -0.239  0.008 -0.002  0.001  0.002  0.015 -0.028
olead    0.019 -0.004 -0.001 -0.008 -0.011  0.057 -0.088  0.042
male1   -0.126 -0.002 -0.001  0.001  0.000  0.046 -0.028 -0.001 -0.118
age     -0.315  0.000  0.007 -0.002  0.003 -0.119  0.067 -0.059  0.031 -0.023
```

In order to get REML estimates, all one has to do when using the *lme4* package is to drop the argument `REML = FALSE` or explicitly change it to `REML = TRUE`. This leads to the following code

```
sj4.reml <- lmer(support ~ tenurez + tradez + gdpz + inflz
                         + inclow + inchi + lright + olead + male + age
                         + (1 | country),
             REML = TRUE,
             data = steenbergen_jones_data)
summary(sj4.reml)
```

or simply

```
sj4.reml <- lmer(support ~ tenurez + tradez + gdpz + inflz
                         + inclow + inchi + lright + olead + male + age
                         + (1 | country),
             data = steenbergen_jones_data)
summary(sj4.reml)
```

with the following result:

```
Linear mixed model fit by REML ['lmerMod']
Formula: support ~ tenurez + tradez + gdpz + inflz + inclow + inchi +
    lright + olead + male + age + (1 | country)
   Data: steenbergen_jones_data

REML criterion at convergence: 43722.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0505 -0.6445  0.1094  0.7575  2.4478

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.2844   0.5333
 Residual             3.3547   1.8316
Number of obs: 10777, groups:  country, 14
```

```
Fixed effects:
             Estimate Std. Error t value
(Intercept)  5.420907   0.169967  31.894
tenurez      0.235291   0.175498   1.341
tradez       0.329125   0.164079   2.006
gdpz        -0.498034   0.343200  -1.451
inflz        0.338932   0.205984   1.645
inclow      -0.129702   0.048882  -2.653
inchi        0.090128   0.045413   1.985
lright       0.034653   0.008829   3.925
olead        0.105046   0.020306   5.173
male1        0.025872   0.035679   0.725
age         -0.011104   0.001074 -10.342

Correlation of Fixed Effects:
        (Intr) tenurz tradez gdpz   inflz  inclow inchi  lright olead  male1
tenurez -0.135
tradez   0.135 -0.281
gdpz     0.383 -0.503  0.365
inflz    0.282 -0.363  0.378  0.701
inclow  -0.036  0.001  0.004 -0.005  0.001
inchi   -0.077  0.002  0.010 -0.001  0.004  0.224
lright  -0.199  0.007 -0.002  0.001  0.002  0.015 -0.028
olead    0.016 -0.003 -0.001 -0.006 -0.009  0.057 -0.088  0.042
male1   -0.105 -0.002 -0.001  0.001  0.000  0.046 -0.028 -0.001 -0.118
age     -0.262  0.000  0.005 -0.001  0.002 -0.119  0.067 -0.059  0.031 -0.023
```

While it is possible to get REML-based standard errors and thus test statistics for the coefficients, *lme4* explicitly does not provide the infrastructure to conduct inferences in terms of significance tests, confidence intervals or *p*-values, because there is no analytical solution on how degrees of freedom should be calculated. The degrees of freedom can only be approximated and there are competing heuristics to do so. Therefore, *lme4* refrains to implement a default. One way of getting *p*-values nevertheless is using the *lmerTest* package, which overrides the `lmer()` function of the *lme4* package with its own version. For the current example, the model is therefor re-fitted with the *lmerTest* variant of `lmer()` and the summary function is called again:

```
sj4LT.reml <- lmer(support ~ tenurez + tradez + gdpz + inflz
                         + inclow + inchi + lright + olead + male + age
                         + (1 | country),
             data = steenbergen_jones_data)
summary(sj4LT.reml)
```

While `summary()` now reports degrees of freedom and *p*-values along with estimates, standard errors, and test statistics, the output appears less "user-friendly":

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: support ~ tenurez + tradez + gdpz + inflz + inclow + inchi +
    lright + olead + male + age + (1 | country)
   Data: steenbergen_jones_data

REML criterion at convergence: 43722.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0505 -0.6445  0.1094  0.7575  2.4478

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.2844   0.5333
 Residual             3.3547   1.8316
Number of obs: 10777, groups:  country, 14

Fixed effects:
              Estimate Std. Error         df t value Pr(>|t|)
(Intercept)  5.421e+00  1.700e-01  1.197e+01  31.894 5.98e-13 ***
tenurez      2.353e-01  1.755e-01  8.978e+00   1.341  0.21295
tradez       3.291e-01  1.641e-01  8.966e+00   2.006  0.07595 .
gdpz        -4.980e-01  3.432e-01  8.977e+00  -1.451  0.18077
inflz        3.389e-01  2.060e-01  8.964e+00   1.645  0.13443
inclow      -1.297e-01  4.888e-02  1.076e+04  -2.653  0.00798 **
inchi        9.013e-02  4.541e-02  1.076e+04   1.985  0.04721 *
lright       3.465e-02  8.829e-03  1.076e+04   3.925 8.72e-05 ***
olead        1.050e-01  2.031e-02  1.076e+04   5.173 2.34e-07 ***
male1        2.587e-02  3.568e-02  1.076e+04   0.725  0.46839
age         -1.110e-02  1.074e-03  1.076e+04 -10.342  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
        (Intr) tenurz tradez gdpz   inflz  inclow inchi  lright olead  male1
tenurez -0.135
tradez   0.135 -0.281
gdpz     0.383 -0.503  0.365
inflz    0.282 -0.363  0.378  0.701
inclow  -0.036  0.001  0.004 -0.005  0.001
inchi   -0.077  0.002  0.010 -0.001  0.004  0.224
lright  -0.199  0.007 -0.002  0.001  0.002  0.015 -0.028
olead    0.016 -0.003 -0.001 -0.006 -0.009  0.057 -0.088  0.042
```

```
male1   -0.105 -0.002 -0.001  0.001  0.000  0.046 -0.028 -0.001 -0.118
age     -0.262  0.000  0.005 -0.001  0.002 -0.119  0.067 -0.059  0.031 -0.023
```

It appears that the package *lmerTest* does not do anything to provide improved confidence intervals, although some approximated degrees of freedom are clearly smaller than 10. That is, the confidence intervals are identical irrespective of whether they are based on *lme4* or *lmerTest* model objects, i.e. in our example

```
confint(sj4.reml)
```

and

```
confint(sj4LT.reml)
```

give identical results:

```
Computing profile confidence intervals ...
                    2.5 %       97.5 %
.sig01       0.3021227400  0.653900119
.sigma       1.8068776643  1.855804738
(Intercept)  5.1307495120  5.711355557
tenurez     -0.0603337230  0.531066669
tradez       0.0526560276  0.605546425
gdpz        -1.0765440368  0.079987155
inflz       -0.0082173849  0.685870132
inclow      -0.2259803086 -0.034397949
inchi        0.0006783272  0.178663444
lright       0.0172702375  0.051870432
olead        0.0652657041  0.144841245
male1       -0.0440943000  0.095736994
age         -0.0132141937 -0.009006406
```

To facilitate the use of inference techniques based on a t-distribution, one of the authors of the paper created an *R* package called *iimm* (for "improved inference for multilevel models"), which is currently available on Github (Elff 2018a). This package can be installed using:

```
devtools::install_github("melff/iimm")
```

If the "devtools" package is not installed on the system, it has to be first installed from CRAN with

```
install.packages("devtools")
```

The package *iimm* essentially provides a single function `lmer_t()`, which adds degrees of freedom, confidence intervals, and p-values to the results of `lmer()`. It allows for three options for the computation of the degrees of freedom: a simple $m - l - 1$ heuristic (the default setting), the Satterthwaite method from the *lmerTest* package (which are obtained

without overriding the function `lmer()`) and the Kenward-Roger method from the *pbkrtest* package. It should be noted that the three methods considerably differ in speed: the $m-l-1$ heuristic is the fastest and the Kenward-Roger method the slowest.

The following code calls `lmer_t()` with each of the three methods and records the time:

```
t0 <- Sys.time()
sj4.reml.heur.t <- lmer_t(sj4.reml)
t1 <- Sys.time()
sj4.reml.Satter.t <- lmer_t(sj4.reml,method="Satterthwaite")
t2 <- Sys.time()
sj4.reml.KR.t <- lmer_t(sj4.reml,method="Kenward-Roger")
t3 <- Sys.time()
```

The object `sj4.reml.heur.t` now contains the estimates of the multilevel model from `sj4.reml` and additionally degrees of freedom, confidence intervals and p-values based on the $m-l-1$ heuristic. The `sj4.reml.Satter.t` contains this additional information based on the Satterthwaite method while in `sj4.reml.KF.t` the additional information is based on the Kenward-Roger method. The variables `t0`, `t1`, `t2`, and `t3` contain the time points before and after the respective calls to `lmer_t()` and thus allow to measure how long they take. The time measurement indicates that the heuristic method takes about 0.02 seconds, the Satterthwaite method takes 0.35 seconds, while the Kenward-Roger method takes almost 2 minutes(!):

```
> t1 - t0
Time difference of 0.01696873 secs
> t2 - t1
Time difference of 0.3486767 secs
> t3 - t2
```

A comparison of the model summaries obtained with the three methods shows that the three methods lead to similar conclusions. The results of the Satterthwaite and Kenward-Roger methods are almost identical, and give slightly more conservative degrees of freedom than the heuristic method:

```
summary(sj4.reml.heur.t)
```

```
Linear mixed model fit by REML ['lmerMod']
   t-tests use the Heuristic method.
Formula: support ~ tenurez + tradez + gdpz + inflz + inclow + inchi +
    lright + olead + male + age + (1 | country)
   Data: steenbergen_jones_data

REML criterion at convergence: 43722.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0505 -0.6445  0.1094  0.7575  2.4478
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.2844   0.5333
 Residual             3.3547   1.8316
Number of obs: 10777, groups:  country, 14

Coefficients:
            Estimate Std.Err t value  country Lower   Upper   Pr(>|t|)
(Intercept)  5.4209  0.1700   31.8939    13    5.0537  5.7881 9.899e-14 ***
tenurez      0.2353  0.1755    1.3407     9   -0.1617  0.6323 0.212872
tradez       0.3291  0.1641    2.0059     9   -0.0420  0.7003 0.075829  .
gdpz        -0.4980  0.3432   -1.4511     9   -1.2744  0.2783 0.180685
inflz        0.3389  0.2060    1.6454     9   -0.1270  0.8049 0.134290
inclow      -0.1297  0.0489   -2.6534 10765   -0.2255 -0.0339 0.007981  **
inchi        0.0901  0.0454    1.9846 10765    0.0011  0.1791 0.047210  *
lright       0.0347  0.0088    3.9251 10765    0.0173  0.0520 8.724e-05 ***
olead        0.1050  0.0203    5.1732 10765    0.0652  0.1448 2.343e-07 ***
male1        0.0259  0.0357    0.7251 10765   -0.0441  0.0958 0.468387
age         -0.0111  0.0011  -10.3424 10765   -0.0132 -0.0090 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
summary(sj4.reml.Satter.t)
```

```
Linear mixed model fit by REML ['lmerMod']
   t-tests use the Satterthwaite method.
Formula: support ~ tenurez + tradez + gdpz + inflz + inclow + inchi +
    lright + olead + male + age + (1 | country)
   Data: steenbergen_jones_data

REML criterion at convergence: 43722.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0505 -0.6445  0.1094  0.7575  2.4478

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.2844   0.5333
 Residual             3.3547   1.8316
Number of obs: 10777, groups:  country, 14

Coefficients:
            Estimate Std.Err t value    Df    Lower   Upper   Pr(>|t|)
(Intercept)  5.4209  0.1700   31.8939    12.0  5.0505  5.7913 5.980e-13 ***
tenurez      0.2353  0.1755    1.3407     9.0 -0.1619  0.6324 0.212951
```

```
tradez       0.3291  0.1641    2.0059     9.0 -0.0423  0.7005 0.075948  .
gdpz        -0.4980  0.3432   -1.4511     9.0 -1.2747  0.2786 0.180768
inflz        0.3389  0.2060    1.6454     9.0 -0.1273  0.8052 0.134425
inclow      -0.1297  0.0489   -2.6534 10759.4 -0.2255 -0.0339 0.007981  **
inchi        0.0901  0.0454    1.9846 10759.8  0.0011  0.1791 0.047210  *
lright       0.0347  0.0088    3.9251 10761.2  0.0173  0.0520 8.724e-05 ***
olead        0.1050  0.0203    5.1732 10762.5  0.0652  0.1448 2.343e-07 ***
male1        0.0259  0.0357    0.7251 10757.5 -0.0441  0.0958 0.468387
age         -0.0111  0.0011  -10.3424 10758.3 -0.0132 -0.0090 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
summary(sj4.reml.KR.t)
```

```
Linear mixed model fit by REML ['lmerMod']
   t-tests use the Kenward-Roger method.
Formula: support ~ tenurez + tradez + gdpz + inflz + inclow + inchi +
    lright + olead + male + age + (1 | country)
   Data: steenbergen_jones_data

REML criterion at convergence: 43722.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0505 -0.6445  0.1094  0.7575  2.4478

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.2844   0.5333
 Residual             3.3547   1.8316
Number of obs: 10777, groups:  country, 14

Coefficients:
            Estimate Std.Err t value    Df    Lower   Upper   Pr(>|t|)
(Intercept)  5.4209  0.1700   31.8938    12.0  5.0505  5.7913 5.796e-13 ***
tenurez      0.2353  0.1755    1.3407     9.0 -0.1618  0.6324 0.212905
tradez       0.3291  0.1641    2.0059     9.0 -0.0422  0.7004 0.075902  .
gdpz        -0.4980  0.3432   -1.4511     9.0 -1.2745  0.2785 0.180720
inflz        0.3389  0.2060    1.6454     9.0 -0.1272  0.8051 0.134376
inclow      -0.1297  0.0489   -2.6533 10759.4 -0.2255 -0.0339 0.007983  **
inchi        0.0901  0.0454    1.9846 10759.8  0.0011  0.1791 0.047218  *
lright       0.0347  0.0088    3.9249 10761.2  0.0173  0.0520 8.731e-05 ***
olead        0.1050  0.0203    5.1728 10762.5  0.0652  0.1449 2.348e-07 ***
male1        0.0259  0.0357    0.7251 10757.5 -0.0441  0.0958 0.468389
age         -0.0111  0.0011  -10.3423 10758.3 -0.0132 -0.0090 < 2.2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Apparently, the additional computational burden of the Kenward-Roger method in comparison to the Satterthwaite method may not be worth its while.

The following code uses the function `mtable()` from the package *memisc* (Elff 2018b) to compare estimates, standard errors, and p-values obtained with ML and REML, with and without t-distribution based on the Satterthwaite method.

```
library(memisc)
# First we create a coefficient style
# that combines estimates and p-values
setCoefTemplate(pval=c(
                    est="($est:#)($p:*)",
                    p="(($p:#))"
                ))
# Second we construct ML and REML tables
mtML <-   mtable("Normal"      = sj4.ml,
                 coef.style="pval")
mtREML <- mtable(Normal        = sj4.reml,
                 Satterthwaite = sj4.reml.Satter.t,
                 coef.style="pval")
# ... and combine them
c(ML=mtML,
  REML=mtREML)
```

```
=============================================================
                        ML                   REML
                  ------------ --------------------------
                      Normal        Normal    Satterthwaite
-------------------------------------------------------------
  (Intercept)          5.421***      5.421***     5.421***
                      (0.000)       (0.000)      (0.000)
  tenurez              0.236         0.235        0.235
                      (0.094)       (0.180)      (0.213)
  tradez               0.329*        0.329*       0.329
                      (0.012)       (0.045)      (0.076)
  gdpz                -0.499        -0.498       -0.498
                      (0.070)       (0.147)      (0.181)
  inflz                0.339*        0.339        0.339
                      (0.040)       (0.100)      (0.134)
  inclow              -0.130**      -0.130**     -0.130**
                      (0.008)       (0.008)      (0.008)
  inchi                0.090*        0.090*       0.090*
                      (0.048)       (0.047)      (0.047)
  lright               0.035***      0.035***     0.035***
```

```
                      (0.000)       (0.000)      (0.000)
  olead                0.105***      0.105***     0.105***
                      (0.000)       (0.000)      (0.000)
  male1                0.026         0.026        0.026
                      (0.469)       (0.468)      (0.468)
  age                 -0.011***     -0.011***    -0.011***
                      (0.000)       (0.000)      (0.000)
-------------------------------------------------------------
  Var(residual)        3.353         3.355        3.355

-------------------------------------------------------------
  Var(~1|country)      0.181         0.284        0.284

-------------------------------------------------------------
  Total            10777         10777        10777
  country             14            14           14
=============================================================
  Significance: *** = p < 0.001; ** = p < 0.01;
                * = p < 0.05
```

Note that the "failure" of the macro-level variables to have statistically significant coefficients arises because of the large standard errors and not because of the small sizes of the coefficients. That is, the negative result could be a consequence of multicollinearity among the macro-level variables. It may be that macro-level predictors are important but their influence is simply difficult to disentangle. This is corroborated by the output from `summary(sj4.reml)`, which shows that the estimates of the four contextual variables are highly correlated. In such situations, it is advisable to conduct a test that involves several coefficients simultaneously. In the following it is shown first how to conduct an F-test with *lmerTest* and Satterthwaite denominator degrees of freedom and second how to do an F-test with Kenward-Roger degrees of freedom and the *pbkrtest* package.

The following code shows such a test with the help of *lmerTest*:

```
library(lme4)
library(lmerTest)
# Note that lmer() now is modified by "lmerTest"
sj4LT.reml <- lmer(support ~ tenurez + tradez + gdpz + inflz
                         + inclow + inchi + lright + olead + male + age
                         + (1|country),
             data = steenbergen_jones_data)
# First construct a matrix for contrasts to be tested
test.matrix <- function(model,test_these){
    cf <- fixef(model)
    cfn <- names(cf)
    L <- diag(nrow=length(cf))
    colnames(L) <- rownames(L) <- cfn
```

```
    L[test_these,]
}
macro.test.matrix <- test.matrix(sj4LT.reml,
                                 c("tenurez","tradez","gdpz","inflz"))
# 'contest() conducts a multi-parameter Wald test
# an F-distribution where the denominator degrees
# of freedom are determined by the Satterthwaite method
contest(sj4LT.reml,
        L = macro.test.matrix)
```

```
    Sum Sq  Mean Sq NumDF    DenDF  F value     Pr(>F)
1 67.55859 16.88965     4 7.955639 4.999273 0.02594672
```

The multi-parameter F-test with Satterthwaite denominator degrees of freedom tells a different story than the t-tests of the individual coefficients of the contextual variables: the null hypothesis test that none of the contextual variables matters is rejected at 5% level of significance.

The following code repeats the test using the *pbkrtest* package and Kenward-Roger degrees of freedom;

```
# For using the Kenward-Roger method we need a "null" model,
# the same set of observations:
sj4LT.reml.0 <- update(sj4LT.reml,
                       .~.-(tenurez + tradez + gdpz + inflz),
                       subset=is.finite(tenurez + tradez + gdpz + inflz))
# and then compare the "full" and the "null" model:
```

```
F-test with Kenward-Roger approximation; computing time: 320.58 sec.
large : support ~ tenurez + tradez + gdpz + inflz + inclow + inchi +
    lright + olead + male + age + (1 | country)
small : support ~ inclow + inchi + lright + olead + male + age + (1 |
    country)
        stat    ndf    ddf F.scaling p.value
Ftest 5.3537 4.0000 8.9962         1  0.0174 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Obviously, the F-test both lead to the same conclusion, irrespective of whether the Satterthwaite or the Kenward-Roger method is used to determine the denominator degrees of freedom: the influence of contextual variables matters, their impact is statistically significant at 5% level. However the test using the Kenward-Roger method takes *much* longer.

Note that the result is not biased by multiple testing, this would be the case if the tests of the null hypothesis with respect to each single coefficient were stochastically independent. This however is not the case. Instead, the joint F-tests takes into account the interdependence among the coefficient estimates.

C.3 Summary

This section can be summarized by the following remarks and recommendations:

1. Obtaining REML estimates in *R* is relatively straightforward for linear multilevel models if one uses the "standard" packages *nlme* (Pinheiro et al. 2018) or *lme4* (Bates et al. 2015). REML is the default estimator for linear multilevel models when these packages are used.

2. Obtaining t-tests with degrees of freedom determined by the $m - l - 1$ heuristic is also straightforward with the *nlme* package. However, multi-parameter F-tests are not supported with *nlme*. Also, the heuristic used by the *nlme* package to determine the appropriate degrees of freedom breaks down when cross-level interactions are involved.

3. In the example application where the total number of individual observations (at the lowest level) is large, the $m - l - 1$ heuristic, the Satterthwaite method and the Kenward-Roger method lead to very similar results in terms of degrees of freedom and hypothesis tests. However, the Kenward-Roger method is very slow, at least in its current implementation.

4. If contextual variables are correlated, multicollinearity may negatively affect the precision of coefficient estimates and negative outcomes of significance tests of individual coefficients. In this case, joint multi-parameter tests may be preferable.

5. For inference about multilevel models, especially with a small number of large upper-level units, *lme4* (Bates et al. 2015), *lmerTest* (Kuznetsova, Brockhoff, and Christensen 2017), and *iimm* (Elff 2018a) may be the most useful combination. The *pbkrtest* package (Halekoh and Højsgaard 2014) seems to be better suited for cases where both the number of upper-level units and the total sample size is small.

D Improved Methods for Inference about Multilevel Models with Few Clusters in *Stata*

This section explains how to obtain improved likelihood-based inference in *Stata*. The illustration is based on the replication of Steenbergen and Jones' (2002) two-level random intercept model discussed in the main paper. Our replication package contains all information necessary to obtain the analysis data that this illustration is based on. Thus everything that follows can be replicated.

The *Stata* default, which runs the Steenbergen and Jones' two-level random intercept model, is the following call to *Stata*'s `mixed` command and associated output:

```
mixed support tenurez tradez gdpz inflz inclow inchi lright olead male age ///
  || cntry:
```

```
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -21837.019
Iteration 1:   log likelihood = -21837.019  (backed up)

Computing standard errors:

Mixed-effects ML regression                     Number of obs     =     10,777
Group variable: cntry                           Number of groups  =         14

                                                Obs per group:
                                                              min =        621
                                                              avg =      769.8
                                                              max =      1,331

                                                Wald chi2(10)     =     222.01
Log likelihood = -21837.019                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
     support |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tenurez |   .2331334   .1391883     1.67   0.094    -.0396707    .5059374
      tradez |   .3290533   .1314428     2.50   0.012     .0714302    .5866764
        gdpz |  -.3293457   .1815711    -1.81   0.070    -.6852185    .0265272
       inflz |   .3497982   .1704596     2.05   0.040     .0157035    .6838929
      inclow |  -.1301887   .0488669    -2.66   0.008     -.225966   -.0344114
       inchi |   .0896713   .0453986     1.98   0.048     .0006917    .1786509
      lright |   .0345704   .0088255     3.92   0.000     .0172727    .0518681
       olead |   .1050535   .0202984     5.18   0.000     .0652693    .1448377
```

```
        male |   .0258214   .0356687     0.72   0.469    -.0440879    .0957307
         age |  -.0111103   .0010733   -10.35   0.000     -.013214   -.0090066
       _cons |   5.465067   .1308397    41.77   0.000     5.208626    5.721508
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
cntry: Identity              |
                  var(_cons) |   .1809975     .07022      .0846134    .3871737
-----------------------------+------------------------------------------------
               var(Residual) |   3.352813   .0457044       3.26442      3.4436
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 466.50        Prob >= chibar2 = 0.0000
```

But as we explain in detail in the article, this output includes erroneous results. While the estimated slope coefficients are unbiased, standard errors and random-effects parameters are too small, and so are the *p*-values and 95% confidence intervals. As we further elaborate, improving inference for mixed effects models necessitates two steps. First, we need to estimate our mixed effects model using REML. Second, we need to construct confidence intervals and *p*-values based on a *t*-distribution with approximated degrees of freedom.

Stata allows to implement both steps as options of its `mixed` command. For didactic purposes, we will first discuss REML estimation and afterwards complement the degrees of freedom approximation. Of course, practitioners should simply perform one single estimation with both options specified.

As a first step, let us estimate the above-displayed model with REML. Note that currently generalized linear mixed effects models cannot be estimated via REML in *Stata*; Commands like `meologit` offer no `reml` option. But for linear mixed effects models, REML can be performed by simply adding the option `reml` to our command:

```
mixed support tenurez tradez gdpz inflz inclow inchi lright olead male age ///
  || cntry:, reml
```

```
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log restricted-likelihood = -21861.582
Iteration 1:   log restricted-likelihood = -21861.582

Computing standard errors:

Mixed-effects REML regression                   Number of obs     =     10,777
Group variable: cntry                           Number of groups  =         14
```

```
                                                Obs per group:
                                                              min =        621
                                                              avg =      769.8
                                                              max =      1,331

                                                Wald chi2(10)     =     209.35
Log restricted-likelihood = -21861.582          Prob > chi2       =     0.0000

------------------------------------------------------------------------------
     support |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tenurez |   .2329039   .1737206     1.34   0.180    -.1075822      .57339
      tradez |   .3291246   .1640823     2.01   0.045     .0075292      .65072
        gdpz |  -.3288531     .22662    -1.45   0.147      -.77302    .1153139
       inflz |   .3501325   .2127947     1.65   0.100    -.0669374    .7672024
      inclow |  -.1297018   .0488821    -2.65   0.008     -.225509   -.0338946
       inchi |   .0901283    .045413     1.98   0.047     .0011204    .1791362
      lright |   .0346528   .0088285     3.93   0.000     .0173492    .0519565
       olead |   .1050457   .0203059     5.17   0.000      .065247    .1448445
        male |   .0258719   .0356789     0.73   0.468    -.0440574    .0958012
         age |   -.011104   .0010736   -10.34   0.000    -.0132083   -.0089997
       _cons |   5.464579   .1565523    34.91   0.000     5.157742    5.771416
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
cntry: Identity              |
                  var(_cons) |   .2844146   .1363317      .1111567    .7277263
-----------------------------+------------------------------------------------
               var(Residual) |   3.354679   .0457426      3.266213    3.445542
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 485.81        Prob >= chibar2 = 0.0000
```

If we compare this model output to the initial one estimated by ML, the most important difference can be seen in the table displaying the "Random-effects Parameters". Whereas initially the variance of the random intercept (i.e., `var(_cons)`) was estimated as .1809975, the corresponding REML estimate is .2844146. As we explain in the article, ML underestimates random effects when upper-level samples are small. This also affects standard error estimates, and thereby again *p*-values and confidence intervals. Take for instance the "tradez" predictor. Its standard error was initial estimated as .1314428 by ML, but has increased to .1640823 based on REML estimation. As one would expect, the lower-level predictors (e.g., `inchi` or `lright`) remain unaffected.

Nevertheless, `tradez` remains statistically significant with $p = 0.045$ and a 95% confidence interval that does not encompass 0 — despite the fact that Stegmueller's Bayesian

credible interval does entail 0. This is because we have only performed the first of two necessary steps. That is, while our standard error estimates are unbiased now, p-values and confidence intervals are still based on the Normal distribution, although we have only 14 upper-level units (i.e., countries).

The necessary second step is thus to approximate the degrees of freedom and base statistical inference on a t-distribution. This can be achieved by adding `dfmethod()` as additional option. Again, `dfmethod()` currently only works for linear mixed effects models in *Stata*. `dfmethod()` allows for several different methods. Repeated, residual, and ANOVA are all inappropriate for typical comparative research (for an overview see Schaalje, McBride, and Fellingham 2002). Practitioners should instead opt for `satterthwaite` given the usual interest in a single-constraint Wald test:

```
mixed support tenurez tradez gdpz inflz inclow inchi lright olead male age ///
  || cntry:, reml dfmethod(satterthwaite)
```

```
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log restricted-likelihood = -21861.582
Iteration 1:   log restricted-likelihood = -21861.582

Computing standard errors:

Computing degrees of freedom:

Mixed-effects REML regression                   Number of obs     =     10,777
Group variable: cntry                           Number of groups  =         14

                                                Obs per group:
                                                              min =        621
                                                              avg =      769.8
                                                              max =      1,331
DF method: Satterthwaite                        DF:           min =       8.98
                                                              avg =   5,873.41
                                                              max =  10,762.50

                                                F(10,    19.47)   =      20.94
Log restricted-likelihood = -21861.582          Prob > F         =     0.0000

------------------------------------------------------------------------------
     support |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tenurez |   .2329039   .1737206     1.34   0.213    -.1601405    .6259483
      tradez |   .3291246   .1640823     2.01   0.076    -.0421848     .700434
        gdpz |  -.3288531     .22662    -1.45   0.181    -.8415867    .1838805
```

```
       inflz |   .3501325   .2127947     1.65   0.134    -.1314278    .8316929
      inclow |  -.1297018   .0488821    -2.65   0.008    -.2255197   -.0338838
       inchi |   .0901283    .045413     1.98   0.047     .0011104    .1791462
      lright |   .0346528   .0088285     3.93   0.000     .0173473    .0519584
       olead |   .1050457   .0203059     5.17   0.000     .0652425     .144849
        male |   .0258719   .0356789     0.73   0.468    -.0440653     .095809
         age |   -.011104   .0010736   -10.34   0.000    -.0132085   -.0089995
       _cons |   5.464579   .1565523    34.91   0.000     5.125516    5.803642
------------------------------------------------------------------------------


------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
cntry: Identity              |
                  var(_cons) |   .2844146   .1363317      .1111567    .7277263
-----------------------------+------------------------------------------------
               var(Residual) |   3.354679   .0457426      3.266213    3.445542
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 485.81        Prob >= chibar2 = 0.0000
```

In comparing this third output to the prior second one, we see that the random effects and standard error estimates remain unchanged. Also, the t-values remain unchanged, but now they are explicitly called t-, rather than z-values, indicating that *Stata* now constructs p-values and confidence intervals based on a t- and not the Normal distribution. Accordingly, we see that the p-value for `tradez` changed from 0.045 to 0.076, although standard error and t-value have not. Perfectly in line with Stegmueller's Bayesian credible intervals, our likelihood-based 95% confidence interval entails 0. Beware however, that regression tables produced by `esttab` will report inference based on the Normal distribution even after `dfmethod()` has been specified.

If practitioners are interested in the approximated degrees of freedom, they can specify a third option `dftable()` and either request 95% confidence intervals (`ci`) or p-values (`pvalue`) along with the degrees of freedom. Here, we opt for 95% confidence intervals:

```
mixed support tenurez tradez gdpz inflz inclow inchi lright olead male age ///
  || cntry:, reml dfmethod(satterthwaite) dftable(ci)
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log restricted-likelihood = -21861.582
Iteration 1:   log restricted-likelihood = -21861.582

Computing standard errors:

Computing degrees of freedom:
```

```
Mixed-effects REML regression                   Number of obs     =     10,777
Group variable: cntry                           Number of groups  =         14

                                                Obs per group:
                                                              min =        621
                                                              avg =      769.8
                                                              max =      1,331
DF method: Satterthwaite                        DF:           min =       8.98
                                                              avg =   5,873.41
                                                              max =  10,762.50

                                                F(10,    19.47)   =      20.94
Log restricted-likelihood = -21861.582          Prob > F         =     0.0000

------------------------------------------------------------------------------
     support |      Coef.   Std. Err.           DF      [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tenurez |   .2329039   .1737206           9.0     -.1601405    .6259483
      tradez |   .3291246   .1640823           9.0     -.0421848     .700434
        gdpz |  -.3288531     .22662           9.0     -.8415867    .1838805
       inflz |   .3501325   .2127947           9.0     -.1314278    .8316929
      inclow |  -.1297018   .0488821       10759.4     -.2255197   -.0338838
       inchi |   .0901283    .045413       10759.8      .0011104    .1791462
      lright |   .0346528   .0088285       10761.2      .0173473    .0519584
       olead |   .1050457   .0203059       10762.5      .0652425     .144849
        male |   .0258719   .0356789       10757.5     -.0440653     .095809
         age |   -.011104   .0010736       10758.3     -.0132085   -.0089995
       _cons |   5.464579   .1565523          12.7      5.125516    5.803642
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
cntry: Identity              |
                  var(_cons) |   .2844146   .1363317      .1111567    .7277263
-----------------------------+------------------------------------------------
               var(Residual) |   3.354679   .0457426      3.266213    3.445542
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 485.81        Prob >= chibar2 = 0.0000
```

The output shows that the approximated degrees of freedom are far from what inference based on the Normal distribution would require. Moreover, in this example the Satterthwaite method approximates the same degrees of freedom as our $m - l - 1$ rule: $14 - 4 - 1 = 9$. Unfortunately, our intuitive and computationally very fast $m - l - 1$ rule is not among the methods offered by `dfmethod()`, yet. This is particularly unfortunate, because `dfmethod()`

does not work with weighted estimation, or the `mi` prefix for multiply imputed data. In these cases, practitioners will have to apply it by hand. For the slope coefficient of the `tradez` predictor, this comes down to approximating the degrees of freedom using the $m - l - 1$ rule (which gives 9 in this case), collecting the t-value from `mixed` output (which is 2.01 in this example) and plugging in both to the `ttail` command, which needs to be multiplied by 2 to give a two-tailed test:

```
display ttail(9, 2.01) * 2
.07532804
```

The following remarks and recommendations may summarize this section:

1. Obtaining REML estimates in *Stata* is relatively straightforward for linear multilevel models. While ML is the default estimator of the `mixed` command, simply adding the option `reml` changes that.

2. Obtaining t-tests with degrees of freedom determined by the $m - l - 1$ heuristic can only be implemented by hand. Yet, this is the most robust option that will also work for weighted estimation, or multiply imputed data.

3. Alternatively, the `dfmethod()` option allows to approximate degrees of freedom by the Satterthwaite or Kenward Roger methods.

References

Aitken, Alexander C. 1936. "On Least Squares and Linear Combination of Observations." *Proceedings of the Royal Society of Edinburgh* 55:42–48.

Baltagi, Badi. 2008. *Econometric Analysis of Panel Data.* Hoboken, NJ: Wiley.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. doi:10.18637/jss.v067.i01.

Booth, James G, and James P Hobert. 1999. "Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 61 (1): 265–285.

Breslow, N. E., and D. G. Clayton. 1993. "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association* 88 (421): 9–25.

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

Caffo, Brian S., Wolfgang Jank, and Galin L. Jones. 2005. "Ascent-based Monte Carlo expectation-maximization." *Journal of the Royal Statistical Society. Series B (Methodological)* 67:235–251.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference.* Second. Pacific Grove, CA: Duxbury.

Cox, D. R., and N. Reid. 1987. "Parameter Orthogonality and Approximate Conditional Inference." *Journal of the Royal Statistical Society. Series B (Methodological)* 49 (1): 1–39.

Elff, Martin. 2018a. *iimm: Facilitating Improved Infrence for Multilevel Models with Few Clusters.* R package version 0.1. https://github.com/melff/iimm.

———. 2018b. *memisc: Management of Survey Data and Presentation of Analysis Results.* R package version 0.99.14. https://cran.r-project.org/package=memisc.

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

Firth, David. 1993. "Bias reduction of maximum likelihood estimates." *Biometrika* 80 (1): 27–38. Accessed January 25, 2014.

Giesbrecht, F. G., and J. C. Burns. 1985. "Two-Stage Analysis Based on a Mixed Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results." *Biometrics* 41 (2): 477.

Gourieroux, Christian, and Alain Monfort. 1995a. *Statistics and Econometric Models, Volume 1: General Concepts, Estimation, Predicitions, and Algorithms.* Cambridge: Cambridge University Press.

———. 1995b. *Statistics and Econometric Models, Volume 2: Testing, Confidence Regions, Model Selection, and Asymptotic Theory.* Cambridge: Cambridge University Press.

Greene, William H. 2012. *Econometric Analysis.* 7th. Upper Saddle River, NJ: Prentice Hall.

Halekoh, Ulrich, and Søren Højsgaard. 2014. "A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest." *Journal of Statistical Software* 59 (9): 1–30. http://www.jstatsoft.org/v59/i09/.

Harville, David A. 1997. *Matrix Algebra From a Statistician's Perspective.* New York: Springer.

Huber, Peter J. 1967. "The Behavior of Maximum Likelihood Estimators under Non-Standard Conditions." In *Proceedings of the 5. Berkeley Symposium on Mathematical Statistics and Probability, Held at the Statistical Laboratory University of California June 21–July 18, 1965 and Dec. 27, 1965–Jan. 7, 1966,* edited by Lucien M. LeCam, 221–233. Berkeley, CA: University of California Press.

Jeffreys, Harold. 1946. "An Invariant Form for the Prior Probability in Estimation Problems." *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 186 (1007): 453–461. ISSN: 0080-4630.

Jiang, Jiming. 1999. "On Unbiasedness of the Empirical BLUE and BLUP." *Statistics & Probability Letters* 41 (1): 19–24.

Kackar, Raghu N., and David A. Harville. 1981. "Unbiasedness of Two-stage Estimation and Prediction Procedures for Mixed Linear Models." *Communications in Statistics-Theory and Methods* 10 (13): 1249–1261.

Kenward, Michael G., and James H. Roger. 1997. "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood." *Biometrics* 53 (3): 983–997.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. "lmerTest Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software* 82 (13): 1–26. doi:10.18637/jss.v082.i13.

Kuznetsova, Alexandra, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. 2015. *lmerTest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (lmer Objects of lme4 Package).*

Lee, Woojoo, and Youngjo Lee. 2012. "Modifications of REML Algorithm for HGLMs." *Statistics and Computing* 22 (4): 959–966.

Lee, Youngjo, and John A. Nelder. 1996. "Hierarchical Generalized Linear Models." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (4): 619–678.

Lehmann, E. L., and George Casella. 2011. *Theory of Point Estimation.* 2nd ed. Springer New York.

Lin, Xihong, and Norman E. Breslow. 1996. "Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion." *Journal of the American Statistical Association* 91 (435): 1007–1016.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models. Second Edition.* London, New York: Chapman / Hall.

McCullagh, Peter, and Robert Tibshirani. 1990. "A Simple Method for the Adjustment of Profile Likelihoods." *Journal of the Royal Statistical Society. Series B (Methodological)* 52 (2): 325–344.

McCulloch, Charles E. 1997. "Maximum Likelihood Algortithms for General Linear Mixed Models." *Journal of the American Statistical Association* 92 (437): 162–170.

Patterson, H. D., and R. Thompson. 1971. "Recovery of Inter-Block Information When Block Sizes Are Unequal." *Biometrika* 58 (3): 545–554.

Pinheiro, José C., and Douglas M. Bates. 1995. "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model." *Journal of Computational and Graphical Statistics* 4 (1): 12–35.

———. 2000. *Mixed-Effects Models in S and S-PLUS* [in en]. New York: Springer.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. 2018. *nlme: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-137. https://CRAN.R-project.org/package=nlme.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Second. Thousand Oaks: Sage Pubn Inc.

Raudenbush, Stephen W., Anthony S. Bryk, and Richard Congdon. 2004. *HLM 6.* Skokie, IL: Scientific Software International.

Rönnegård, Lars, Moudud Alam, and Xia Shen. 2015. *The Hglm Package (Version 2.0).* https://cran.r-project.org/package=hglm.

Satterthwaite, F. E. 1946. "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin* 2 (6): 110–114.

Schaalje, G. Bruce, Justin B. McBride, and Gilbert W. Fellingham. 2002. "Adequacy of approximations to distributions of test statistics in complex mixed linear models." *Journal of Agricultural, Biological, and Environmental Statistics* 7 (4): 512–524.

Snijders, Tom A. B., and Roel J. Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* [in en]. London: Sage.

Steenbergen, Marco R., and Bradford S. Jones. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46 (1): 218–237.

Stegmueller, Daniel. 2013. "How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches." *American Journal of Political Science* 57 (3): 748–761.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50 (1): 1–25.

Zorn, Christopher. 2005. "A Solution to Separation in Binary Response Models." *Political Analysis* 13 (2): 157–170.