Research Master's programme Methodology and Statistics for the Behavioural, Biomedical and Social Sciences,

Utrecht University, the Netherlands

MSc Thesis Nikola Sekulovski (6465588)

TITLE: A Default Bayes Factor for testing Null Hypotheses about the Fixed Effects of Linear Two-level Models

May 2022

Supervisor: Prof. Dr. Herbert Hoijtink

Second grader: Prof. Dr. Irene Klugkist

Preferred journal of publication: Psychological Methods

Word count: Full text: $\sim$ 10 850; Abstract: $\sim$ 240; Main body: $\sim$ 9500

Abstract

Testing null hypotheses of the form "$\beta = 0$," by the use of various Null Hypothesis Significance Tests (rendering a dichotomous reject/not reject decision), is considered standard practice when evaluating the individual parameters of statistical models. Bayes Factors for testing these (and other) hypotheses allow users to quantify the evidence in the data that is in favour of a hypothesis. Unfortunately, when testing equality contained hypotheses, the Bayes Factors are sensitive to the specification of prior distributions, which may be hard to specify by applied researchers. The paper proposes a default Bayes Factor with clear operating characteristics when used for testing whether the fixed parameters of linear two-level models are equal to zero. This is achieved by generalising an already existing approach for linear regression, presented in Hoijtink (2021). The generalisation requires: (i) the sample size for which a new estimator for the effective sample size in two-level models containing random slopes is proposed; (ii) the effect size for the fixed effects for which the so-called *marginal* R^2 for the fixed effects based on Nakagawa & Schielzeth (2013) is used. Implementing the aforementioned requirements in a small simulation study shows that the Bayes Factor yields clear operating characteristics regardless of the value for sample size and the estimation method. The paper gives practical examples and access to an easy-to-use `wrapper function` to calculate Bayes Factors for hypotheses with respect to the fixed coefficients of linear two-level models by using the `R` package `bain`.

Keywords: Bayes Factor, Effective Sample Size, Null Hypotheses, Prior Sensitivity, Two-level Models.

Introduction

Following an increasing wave of criticism directed toward *Null Hypothesis Significance Testing* (NHST, Cohen, 1994; Wagenmakers, 2007), the *Bayes Factor* (Jeffreys, 1935, further abbreviated as BF), usually considered the cornerstone of Bayesian hypothesis evaluation, is gaining momentum. This paradigm, at least compared to NHST, does not rely on strict cutoff values such as the often vilified "$\alpha = .05$" and can be used to quantify the evidence in the data that is *in favour* of a hypothesis (Kass & Raftery, 1995). Moreover, if the main interest of a study is on the null hypothesis,[1] (for examples of such situations, see, Wainer, 1999), using NHST only allows us to either retain or reject the hypothesis (since by default the null hypothesis is considered to be true), which is not very informative from a substantive point of view. The aforementioned can be seen as an argument in favour of using the BF since it allows researchers to quantify the evidence in the data for the *null* hypothesis. However, as usually is the case, every convenience comes with its price. For the BF this price means that, when used to evaluate null hypotheses, it becomes sensitive to the specification of the prior distribution (discussed further below and also elaborated in Hoijtink, Mulder, et al., 2019).

This paper aims to answer the question of whether the BF can yield clear operating characteristics when used to test if the *fixed effects* of linear two-level models are *equal* to zero, by generalising the work presented in Hoijtink (2021) for multiple linear regression. More specifically, a BF with clear operating characteristics should reach 19, when the effect size for the fixed effects in the data is exactly zero and decrease accordingly as the effect size increases.

It should be noted that inequality constrained hypotheses[2] are *not* sensitive to the prior specification, however, these are not treated in this paper and the interested reader is referred to Hoijtink (2011) for detailed elaborations. Throughout the text, the application of the BF in evaluating *null hypotheses* is sometimes referred to as *Null Hypotheses Bayesian Testing (NHBT)*, a term also used in Hoijtink (2021) and introduced by Tendeiro & Kiers (2019). In the remainder of this section, two-level models, the default BF and the accompanying software used for the aims of this study are introduced.

[1] *Null Hypotheses* impose equality constraints on the parameters of a statistical model.

[2] Hypotheses imposing *inequality* constraints among the parameters of a statistical model are also referred to as *informative hypotheses*.

Two-level Models

Multilevel models (also called mixed models, hierarchical linear models, multilevel regression models etc.), are useful when the data has a hierarchical structure, for example, when individuals are nested within groups. These models enable researchers to take the within-group dependence into account, as well as allow variables to be defined at their original level of measurement, without the need to aggregate or disaggregate them on one single (usually the lowest) level. Two-level models are the most common type of multilevel models, where, as the name suggests, the data has two levels. Practical examples are, among others, students nested within classes, employees nested within companies or, longitudinal designs, where observations are nested within individuals. For an introduction to multilevel models see, Hox et al. (2017, pp. 1–23). Below, the linear equation for a simple two-level model with two continuous level-1 predictors is presented with the aim to illustrate the most important aspects of these models that are relevant to this paper. For the sake of the example, let's assume that we are dealing with persons nested within groups. Then,

$$Y_{ij} = \alpha + \beta_1 \, X_{1,ij} + \beta_2 \, X_{2,ij} + u_{0j} + u_{1j} \, X_{1,ij} + u_{2j} \, X_{2,ij} + \epsilon_{ij}, \tag{1}$$

where Y_{ij} represents the value of the outcome variable for person $i = 1,\ldots,N$ in group $j = 1,\ldots,G$ (where N denotes the number of level-1 observations and G denotes the number of groups); α is the overall (fixed) intercept having a random component u_{0j}, which denotes the deviation of group j from the overall intercept; β_1 and β_2 represent the fixed effect for the continuous level-1 predictors X_1 and X_2, respectively, with random components u_{1j}, u_{2j}, denoting deviations for group j from the overall slopes for both coefficients, respectively; ϵ_{ij} is the standard residual error term for person i in group j.

The random effects (stored in a vector $\boldsymbol{U}$) are assumed to follow a multivariate normal distribution with a mean vector $\boldsymbol{\mu}$ consisting of zeros and a covariance matrix $\boldsymbol{\Sigma}$ containing their variances and the covariances. The residual error term ϵ_{ij} is assumed to be normally distributed around zero, with an estimated residual variance σ^2_ϵ:

$$\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \; \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2_\epsilon),$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{u0} & \sigma_{u0,u1} & \sigma_{u0,u2} \\ \sigma_{u0,u1} & \sigma^2_{u1} & \sigma_{u1,u2} \\ \sigma_{u0,u2} & \sigma_{u1,u2} & \sigma^2_{u2} \end{pmatrix}.$$

All the estimated parameters (both fixed and random effects) are stored in a parameter vector $\boldsymbol{\theta}$. Since the aim of this paper is focused on testing the fixed effects for the slopes, they are defined in a parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$, while the remaining elements of $\boldsymbol{\theta}$ are treated as nuisance parameters.

As it may be clear by now, this model contains a random intercept and random slopes for both of the level-1 predictors. However, one can easily imagine a situation where, for example, only the intercept is random, or only one of the slopes is random and the second one is fixed. In this paper, the focus is mostly on models with level-1 predictors having random slopes. However, from this setup, generalisations for models having different combinations of continuous level-1 predictors with or without random slopes *and* level-2 predictors can quite easily be constructed. In the Examples section, we show that the proposed approach can be used to test whether the coefficients for both level-1 predictors *and* level-2 predictors are equal to zero.

Bayes Factor

In this paper, we introduce a Bayesian approach for testing the fixed effects of continuous predictors that is not sensitive to the specification of the prior distribution when evaluating hypotheses of the form:

$$H_0 : \beta_1 = \beta_2 = 0 \text{ against } H_u : \beta_1, \beta_2. \tag{2}$$

Where H_u is referred to as the *unconstrained* hypothesis in which there are no constraints imposed on the parameters of $\boldsymbol{\beta}$. This approach allows researchers to quantify the evidence in the data that is *in favour* of the null hypothesis, as well as allowing for the possibility to include informative hypotheses.

In order to be able to apply the approach proposed by Hoijtink (2021) which yields BFs with clear operating characteristics, a measure of the effect size for the fixed effects is needed. One such measure is the concept of explained variance expressed through the familiar coefficient of determination R^2 in linear regression. However, in the context of two-level models, the R^2 does not have a straightforward interpretation, since variance can be explained on different levels of the model, for further elaborations see, Hox et al. (2017, pp. 57–64). Since the focus of this study is on the fixed effects, the so-called pseudo R^2 for the fixed effects, as proposed by Nakagawa & Schielzeth (2013) and further expanded by Johnson (2014) and Jaeger et al. (2017), is used as a measure of the effect size. This R^2 represents the proportion of variation in the outcome variable that is attributed to the fixed effects (for further details, see, Equations 26 and 27 in Nakagawa & Schielzeth, 2013). Throughout this text, in accordance with the original paper, this value for the effect size of the fixed effects will be referred to as the *marginal* R^2 and denoted as R^2_m. Additionally, the variance explained by *both* the fixed *and* the random effects is referred to as the *conditional* R^2 (abbreviated as R^2_c). Thus, by taking the difference between R^2_c and R^2_m, we are able to obtain the (partial) effect size of the random effects.

The *Bayes Factor* (Kass & Raftery, 1995) is defined as the ratio of two marginal likelihoods (see, Equation 3). Tendeiro & Kiers (2019) define the marginal likelihood as: "... weighted average of the likelihood over the observed data, where the weights are provided by the (within) priors."

$$BF_{0u} = \frac{P(D|H_0)}{P(D|H_u)} = \frac{\int P(D|\boldsymbol{\theta}, H_0)P(\boldsymbol{\theta}|H_0)d\boldsymbol{\theta}}{\int P(D|\boldsymbol{\theta}, H_u)P(\boldsymbol{\theta}|H_u)d\boldsymbol{\theta}}, \tag{3}$$

where BF_{0u} denotes the Bayes Factor of the null hypothesis against the unconstrained hypothesis; $P(D|H)$ represents the *marginal* likelihood of the data for each of the hypotheses; at the rightmost part of the equation, these marginal likelihoods are defined as the product of the likelihood function, $P(D|\boldsymbol{\theta}, H)$ and the prior, $P(\boldsymbol{\theta}|H)$, integrated with respect to the parameter vector $\boldsymbol{\theta}$. H, in this case, is either H_0 or H_u. The definition given in Equation 3 has two important aspects: (i) it defines the marginal likelihood as the denominator of Bayes' rule; (ii) it stresses the role of the prior distribution on the marginal likelihood and consequently on the value of the BF itself. The second aspect is the overall reason why the BF is sensitive to the specification of the prior distribution.

Straightforward calculation of the BF, based on its mathematical definition, presented in Equation 3, is difficult in most applied (multi-parameter) situations. However, Equation 3 can be approximated as Equation 4. Thus, translating the BF into a so-called *Approximate* (since, due to large sample theory, it uses normal distributions to *approximate* the prior and posterior distributions of the unconstrained hypothesis) *Adjusted* (since the mean of the prior distribution is *adjusted* on the boundary of the hypotheses under consideration) *Fractional* (since it uses a *fraction* of the information in the data to construct a proper prior distribution) *Bayes Factor* (AAFBF),[3] which is defined as the ratio of the *fit* and *complexity* of the null hypothesis, and is an example of one of the so-called *default* BFs (Gu et al., 2018; for a full derivation of this BF see, Mulder, 2014):

$$AAFBF_{0u} = \frac{f_0}{c_0} = \frac{\int_{\beta\in\beta_0} \mathcal{N}(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}_{\beta}) d\boldsymbol{\beta}}{\int_{\beta\in\beta_0} \mathcal{N}(\boldsymbol{\beta}|0, \hat{\boldsymbol{\Sigma}}_{\beta}/b) d\boldsymbol{\beta}}. \tag{4}$$

When testing null hypotheses: (i) *fit* (f_0) is the density of the *normal approximation* of the *posterior* distribution supported by the null hypothesis at hand; (ii) *complexity* (c_0) is the density of the *normal approximation* of the *prior* distribution supported by the null hypothesis at hand (Hoijtink, Mulder, et al., 2019). In Equation 4, $\hat{\boldsymbol{\beta}}$ represents the vector containing the estimated fixed effects and $\hat{\boldsymbol{\Sigma}}_{\beta}$ denotes its respective covariance matrix. For example, a $BF_{0u} = 5$ would mean that the data is five times in favour of H_0 compared to H_u (see, Hoijtink, Mulder, et al., 2019, for more detailed guidelines on interpreting the values for this BF). It should be noted that in the denominator of Equation 4, the *adjusted mean* of the normal approximation of the prior distribution, $\mathcal{N}(\boldsymbol{\beta}|0, \hat{\boldsymbol{\Sigma}}_{\beta}/b)$, is zero *only* in a situation when testing (null) hypotheses where all parameters are equal to zero (for example, the null hypothesis presented in Equation 2). For a more detailed explanation of the prior *adjusted mean* when using this BF for testing other hypotheses, please see, Gu et al. (2018).

The most important aspect for the present study is that this BF uses a *fraction* $b = \frac{J}{N}$ (see, Equation 4) of the information in the data to construct the scaling parameter of the prior distribution. Where: (i) J by default denotes the number of fixed effects that are set equal to zero in the hypothesis (for example, Equation 2 would yield $J = 2$), which will be changed to a so-called *reference* value

[3] Throughout the remaining parts of this text, the AAFBF is referred to simply as the BF.

(further denoted as J_{ref}), by having the BF_{0u} equal to 19, when the observed R^2_m is zero (based on Hoijtink, 2021 and further elaborated in the next sections); (ii) N represents the sample size. It should be noted that situations in which J is small relative to N allow for the interpretation of the resulting BFs as *approximate* BFs. However, when J is large relative to N, then the resulting BFs should be interpreted as information criteria inspired by the BF (a rule of thumb for a cutoff value suggested in Hoijtink, 2021 is $b = 0.05$).

In Equation 5, the BF can be seen as a multiplicative factor that transforms the prior odds, $\frac{P(H_0)}{P(H_u)}$, of two hypotheses to the posterior odds, $\frac{P(H_0|D)}{P(H_u|D)}$, after seeing the data:

$$\frac{P(H_0|D)}{P(H_u|D)} = BF_{01}\frac{P(H_0)}{P(H_u)}. \tag{5}$$

However, if the prior odds of the hypotheses are set to one, by setting the prior probabilities, $P(H_0)$ and $P(H_u)$, of both hypotheses equal to each other, then the BF will equal the posterior odds (Kass & Raftery, 1995).[4] Thus, when using equal prior model probabilities, having BF_{0u} = 19, when R^2_m is zero in the data, renders posterior model probabilities of $P(H_0|D) = .95$ and $P(H_u|D) = .05$, where the latter, *numerically*, mimics the conventional *Type I error rate* in NHST. However, $P(H_u|D) = .05$ represents the probability of incorrectly rejecting H_0 *conditional* on the data, whereas the Type I error rate is the probability of incorrectly rejecting H_0 based on a (theoretical) sampling distribution, constructed from a population in which H_0 is true. In other words the Type I error rate in NHST is *not* dependent on the data.

As depicted in Equation 4, the sample size is an integral part of the computation of the BF introduced in this paper. However, with multilevel models, it is unclear how to quantify the sample size. Ad-hoc values are, for example, the number of level-1 or level-2 observations (referred to as $N_{level-1}$ and $N_{level-2}$, respectively). Another compromise approach is the so-called *effective sample size*, which is based on the *Intraclass Correlation Coefficient*, abbreviated as ICC (Bliese, 1998; Killip et al., 2004). The value for the ICC can be seen as the proportion of total variance that is explained by the group variances and is calculated by fitting a so-called random intercept-only model (i.e., an empty model with a random intercept). Briefly, the effective sample size shrinks the

[4]The concept prior probabilities should not be confused with prior distributions.

number of level-1 observations when taking into account the within-group clustering of the data (as measured by the ICC). However, a drawback of this particular approach is that it can only be calculated for random intercept-only models with equal group sizes. In this paper, a novel method for calculating the effective sample size, which can also be used for two-level models that have random slopes, is introduced and subsequently used for calculating the BFs. In what follows, the effective sample size based on the ICC will be referred to as the ICC-based N_{eff}.

Software

The programming language for statistical computing `R`, version 4.1.2 (R Core Team, 2021) was used to perform all the analyses and simulations presented in this paper. The `R` package `bain` (Gu et al., 2021), computes the BF_{0u} of a hypothesis against the unconstrained hypothesis, using only the estimated parameters, $\hat{\boldsymbol{\beta}}$, and their respective covariance matrix, $\hat{\boldsymbol{\Sigma}}_{\beta}$ (see, Equation 4). A `wrapper function` was programmed specifically for the aims of this paper, to conveniently use `bain` to test hypotheses about the *fixed* parameters of two-level models, built with the `lmer` function from the `R` package `lme4` (Bates et al., 2015). For an introduction on how to use the `wrapper function`, as well as elaborated examples on the methods presented in the subsequent sections, please see the tutorial available on the authors' website.[5] It should be noted that this function can be used to test null hypotheses (stating that the parameters are *equal* to zero) *and* informative hypotheses for *continuous* level-1 *and* level-2 predictors. It should further be noted that the function includes the option to automatically calculate and implement J_{ref}, as proposed by this study.

The rest of this paper is structured as follows: In the next section, a description of the simulated data sets, that are used throughout the paper is given. Subsequently, the novel method for calculating the effective sample size for two-level models containing random slopes is introduced. Thereafter, the results from a small sensitivity analysis are presented, where the values for J and N are varied, in order to illustrate the sensitivity of the BF to the specification of the prior distribution. Thereafter, the details on the derivation of the value for J_{ref} are given. Afterwards, the operating characteristics of the proposed approach are illustrated by means of a small simulation study. Everything presented in this paper is put together in four examples, by using a real openly-available data set, giving

[5] *https://nikolasekulovski.com/tutorials/*

researchers practical guidelines and recommendations on how to make use of and also properly report this BF. The paper ends with a discussion explaining the benefits of the proposed approach and highlighting its limitations, such that future research may focus on addressing these drawbacks. All the code has been made publicly available and can be accessed through the authors' `GitHub` account.[6]

Data

This section gives an overview of the simulated two-level data sets that are used throughout the following sections. A visual depiction of their properties is given in Figure 1 and every time a simulated data set is mentioned, it will be related to its particular cell in this figure. The research protocol has been approved by the Ethical Review Board of the Faculty of Social and Behavioural Sciences of Utrecht University.

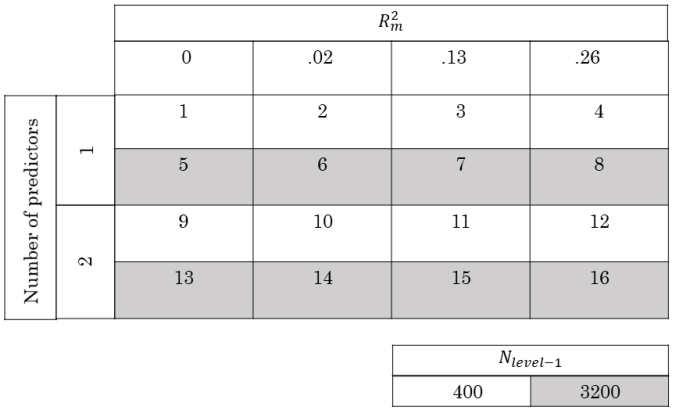

Figure 1: Illustration of the properties of the simulated data sets used throughout the paper. The columns indicate the value for R^2_m and the rows indicate the number of predictors in a data set. The cells in white belong to data sets with $N_{level-1} = 400$ and the cells in grey belong to data sets with $N_{level-1} = 3200$.

The data sets were sampled from four different populations, defined by two factors: (i) the value for $R^2_m = 0, .02, .13, .26$, which correspond to no effect, small, medium and large effect, respectively

[6] *https://github.com/sekulovskin/research-archive-masters-thesis*

(based on Cohen, 1992 for R^2 in multiple linear regression); (ii) the number of predictors, 1 or 2. The data from these four populations were sampled twice with respect to the sample size: (1) $N_{level-1} = 400$, with $N_{level-2} = 20$ and within-group sample size of 20; (2) $N_{level-1} = 3200$, with $N_{level-2} = 80$ and within-group sample size of 40. This setup yields 16 different combinations of two-level data sets, as enumerated in Figure 1. The values of the fixed effects were chosen such that the desired R^2_ms were achieved, and in the case of having two predictors (i.e., cells 9 through 16), both fixed effects were given the same value. The random effects were simulated independently, i.e., they were generated from univariate normal distributions with a mean of zero. The variance for the normal distribution of the intercept was chosen to be 0.1 and the slope variances were set to 0.01 and 0.04, respectively. The residual variance was given a value of 0.36. These values were inspired by two-level models fitted to openly-available data sets such as the `tutorial` data set from the `R` package `R2MLwiN` (Zhang et al., 2016) and the `popularity` data set from Hox et al. (2017, pp. 317–318). This setup yields an effect size for the random effects (i.e., $R^2_c - R^2_m$) that is approximately the same across the data sets with different R^2_ms. More specifically, the effect size for the random effects for cells 1 through 4 is $\sim$ .09, and for all remaining cells, it is $\sim$ .2. The intercept variance accounts for most of this effect size, followed by the slope variance for the second predictor and lastly, the variance of the first predictor. The variances of the random effects were kept constant among all 16 cells in Figure 1. It should be noted that when $R^2_m = 0$ then R^2_c equals the effect size for the random effects, however, when $R^2_m > 0$ the difference between R^2_c and R^2_m becomes the effect size for the random effects i.e., in that case, the effect size for the random effects is a partial R^2 (however, its value still remains constant across the data sets). The only instance when a value for the variance of the random effects was varied is when obtaining data sets with different values for the ICC (achieved by varying the intercept variance), used to illustrate the calculation of effective sample sizes in the next section.

The predictors were simulated from a normal distribution with a mean of zero, which is equivalent to applying grand-mean centering and standard deviations of 0.8 and 0.6, respectively. The practice of (grand) mean centering in the context of multilevel models yields many practical benefits, one among which is better and more efficient convergence, especially in the case when the means and variances of the predictors differ substantially, see, Hox et al. (2017, pp. 46–52).

A New Estimator of Effective Sample Size

If we were to fit a two-level model containing only a random intercept using data with equal group sizes, then a theoretically sound option for the computation of the effective sample size would be the ICC-based N_{eff}. However, the ICC-based N_{eff} has not yet been generalized to models with random slopes and/or data with unequal group sizes. Inspired by the concept of *Multiple Imputation of Missing Data* (Rubin, 1987), a new method for calculating the effective sample size was developed specifically for the aims of this study. The presence of both fixed and random effects within a two-level model allows us to treat the latter as missing values within the observed data set. By drawing samples of parameter vectors (containing both fixed and random effects) from the posterior distribution of the specified two-level model (discussed further below) and adding each vector to a copy of the original data, it is possible to obtain multiple imputed data sets. More specifically, the sampled random effects are added to the respective Y and X values of the level-1 observations in each group. As will be elaborated now, from these multiple imputed data sets, an estimate of the effective sample size can easily be obtained. This new estimate will be referred to as the Multiple Imputation-based effective sample size (abbreviated as MI-based N_{eff}). For the statistical underpinnings of multiple imputation, see, Van Buuren (2018, Ch. 2).

Procedure

In most applied situations, Bayesian model estimation involves drawing samples from the posterior distribution of the parameters, using different Markov Chain Monte Carlo algorithms (further abbreviated as MCMC) and afterwards summarizing the distribution of the drawn samples to obtain Bayesian point estimates and (credible) intervals (for more details on MCMC sampling, see, for example, Van Ravenzwaaij et al., 2018). In this paper, the MCMC sampling is performed using the program `JAGS` (Plummer, 2003) and the `R` package `rjags` (Plummer, 2021) which implement the *Gibbs sampler* (for more details on the Gibbs sampler, see, for example, Gelfand, 2000). First, a two-level model is fitted with `JAGS` to obtain m sampled parameter vectors from the posterior distribution of the two-level model (where m denotes the number of imputations, which in this case corresponds to the total number of iterations of the MCMC algorithm across the chains i.e.,

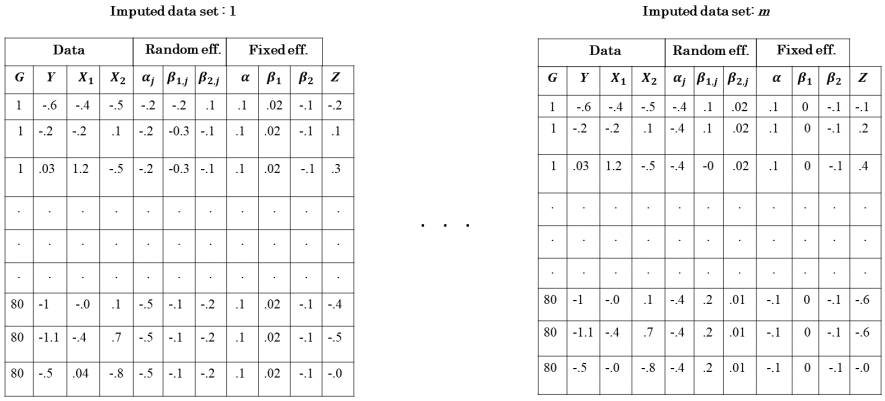

Figure 2: Illustration of the Multiple Imputation process for calculating the effective sample size, where: (i) the first batch of columns represents a copy of the observed data, with a group indicator (G), the outcome variable (Y) and the predictors (X_1 and X_2); (ii) the second batch includes the sampled random effects, with α_j representing the group-specific intercept and β_{1j} and β_{2j} representing the group-specific slopes of the predictors X_1 and X_2, respectively; (iii) the third batch of columns includes the sampled values for the fixed slope α and the fixed intercepts β_1 and β_2; (iv) the last column includes the transformed outcome variable Z, based on Equation 6.

$l = 1,\ldots,m$). Note, in this case, the posterior mean estimates are not of interest. However, the prior distributions of the parameters are chosen to be completely uninformative such that only the data can influence the posterior distribution from which samples of parameter vectors are to be drawn. The interested reader can find the following information in Appendix A: (A1) a detailed specification of the prior distributions, which was inspired by the work presented in the tutorial by Vasishth & Sorensen (2014); (A2) the corresponding `JAGS` code; (A3) a comparison of the (posterior) estimates obtained with `JAGS` and `lmer` by fitting the model on one of the example data sets used further below. Afterwards, each sampled parameter vector is added to a copy of the original data, such that these can be treated as m multiple imputed data sets. A visual depiction of the aforementioned procedure is given in Figure 2. Thereupon, linear regression models can be fitted to each imputed data set. However, in order to be able to fit linear regression models, the value of the outcome variable Y, for each of the m imputed data sets, needs to be transformed in the following manner:

$$Z_{ij} = Y_{ij} - \alpha_j - \beta_{1j}\, X_{1,ij} - \beta_{2j}\, X_{2,ij} + \alpha + \beta_1\, X_{1,ij} + \beta_2\, X_{2,ij}, \tag{6}$$

where Z_{ij} represents the transformed outcome for person i in group j; α_j, β_{1j} and β_{2j} represent the sampled random effects for the intercept and the two predictors, respectively; α, β_1 and β_2 represent the fixed effects for the intercept and the two predictors, respectively. This setting allows us to straightforwardly fit linear regression models to each imputed data set, with Z now serving as the outcome variable (see, Figure 2), that is,

$$Z_i = \eta_0 + \eta_1 \, X_{1,i} + \eta_2 \, X_{2,i} + e_i, \tag{7}$$

where, η_0 represents the linear regression intercept and η_1 and η_2 represent the linear regression slopes for the predictors X_1 and X_2, respectively. e_i is the residual error term, assumed to be normally distributed around zero with variance σ_e.

By using this set-up, for each imputed data set $l = 1,\ldots,m$, we obtain a parameter vector $\hat{\boldsymbol{\eta}}_l$, containing the estimated intercept and slopes, and $\hat{U}_l$ containing their respective covariance matrix. Afterwards, we can use the equations presented in Appendix A4 to obtain the *fraction of missing information*, which will be further denoted as γ.

Finally, the value for the MI-based N_{eff} can be calculated as follows:

$$\text{MI-based } N_{eff} = N_{level-1} - \gamma \, N_{level-1}.$$

MI-based N_{eff} vs ICC-based N_{eff}

In this subsection, we aim to illustrate that this newly developed approach for calculating the effective sample size for two-level models containing random slopes, represents a theoretically grounded alternative relative to the ICC-based N_{eff}. As previously mentioned, ICC-based N_{eff} can only be calculated for a random intercept-only model with equal group sizes. The reason for this is that the value for the ICC is obtained using the variance of the random slopes and the residual variance, estimated from the random intercept-only model:

$$\text{ICC-based } N_{eff} = \frac{N_{level-1}}{1 + (n_c - 1)\text{ICC}}, \tag{8}$$

where n_c denotes the within-group sample size and

$$\text{ICC} = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{\epsilon}}, \tag{9}$$

where σ^2_{u0} denotes the variance for the random slope and σ^2_{ϵ} denotes the residual variance estimated from a random intercept-only model.[7]

Using three simulated data sets corresponding to cell number 13 in Figure 1, with ICC values of .03, .19 and .35, respectively, we first calculated the ICC-based N_{eff} using Equations 8 and 9. Afterwards, by fitting random intercept-only models with `JAGS` and performing the calculations presented in the previous subsection we obtained MI-based N_{eff}s for all three data sets based on a random-intercept-only model. Finally, we calculated the MI-based N_{eff}s based on the full model (containing random slopes, as given in Equation 1). Three more data sets corresponding to cell number 9 of Figure 1, having the same three values for the ICC, were simulated and the calculations were repeated.

Table 1: Results from the effective sample size calculations.

ICC	$N_{level-1}$	ICC-based N_{eff}	MI-based N_{eff} (intercept only)	MI-based N_{eff} (full model)
0.03	3200	1475	1537	1231
0.19	3200	380	316	1009
0.30	3200	252	178	997
0.03	400	255	249	121
0.19	400	87	64	94
0.30	400	60	35	88

The results presented in Table 1 illustrate the following: (i) when calculating the effective sample size for a random intercept-only model, the estimates from both methods tend to be similar, regardless of the value of the ICC and the number of level-1 observations; (ii) when the value for the ICC is close to zero (.03 in this case), the ICC and the new approach, become similar. This observation suggests that when the intraclass correlation coefficient approaches zero, the effective sample size approaches $N_{level-1}$, regardless of the complexity of the model; (iii) most importantly, when a model contains random slopes, the effective sample size obtained by using the new method is larger since it is calculated from a model in which part of the within-group clustering, as indicated

[7]For a random intercept only-model, Equation 1 reduces to $Y_{ij} = \alpha + u_{0j} + \epsilon_{ij}$.

by the random intercept-only model, has been accounted for by including the predictors, thus increasing the effective level-1 observations. As expected based on (ii), (iii) is not the case when having an ICC close to zero. In other words, when having a low value for the ICC the calculated MI-based N_{eff} is slightly lower than its counterparts. These observations should give us confidence in the newly developed MI-based N_{eff}, since, in situations where a two-level model contains random slopes *and* has a medium or large value for the ICC, this approach yields a value for the effective sample size that is larger than the value given by its ICC counterpart. Another reason why this new method should be preferred over the ICC-based N_{eff}, even when dealing with a random intercept model and/or having an ICC close to zero, is that the latter requires the same sample size for each group (see, Equation 8), which is usually never the case in practice.

Sensitivity Analysis

In this section, the sensitivity of the BF to the specification of the values for J and N is briefly illustrated using a simulated data set from cell 13 of Figure 1, where $R^2_m = 0$ (i.e., the null hypothesis presented in Equation 2 is true). This data set has an ICC-based N_{eff} of 413 and an MI-based N_{eff} of 921 (calculated for a model containing a random intercept and random slopes). Table 2 summarizes the results from the fitted two-level model. The parameters were estimated by using *Full Maximum Likelihood Estimation* instead of *Restricted Maximum Likelihood Estimation* (further abbreviated as FML and REML, respectively) since the main interest is in estimating the fixed effects and not the variance components of the random effects, which can be biased when using FML. However, it has been shown that the difference between these two estimation methods with regard to the fixed effects is usually small (Hox, 1998; Kreft & de Leeuw, 1998 as cited in Hox et al., 2017). For further details on these particular estimation procedures, see, Bates et al. (2015) or Hox et al. (2017, pp. 27–29).

As can be seen from the values of their respective standard errors presented in Table 2, both of the fixed effects (β_1 and β_2) are estimated to be around zero. Additionally, the variances of the random effects are also quite low, especially for the random slope of the first predictor.

BFs for the null hypothesis presented in Equation 2 were calculated using the `wrapper function`

Table 2: Estimates from fitting the two-level model with lmer

	Fixed effects		Random effects
	est	SE	var
α	-0.006	0.033	0.076
β_1	0.002	0.017	0.008
β_2	-0.001	0.029	0.041

est = FML estimate; SE = Standard Error; var = Variance of the random effects

while varying both the value of sample size and the value for J. The sample size was set to equal either $N_{level-1}$, $N_{level-2}$, the ICC-based N_{eff} and finally the newly developed MI-based N_{eff}. The value for J was set to equal either the default value (which in this case is two, see, Equation 2) and afterwards, it was iteratively changed to equal 2 J and 3 J, respectively (these choices for J are based on the sensitivity analysis given in Hoijtink, Mulder, et al., 2019). In total 12 BFs, plotted in Figure 1, were obtained.

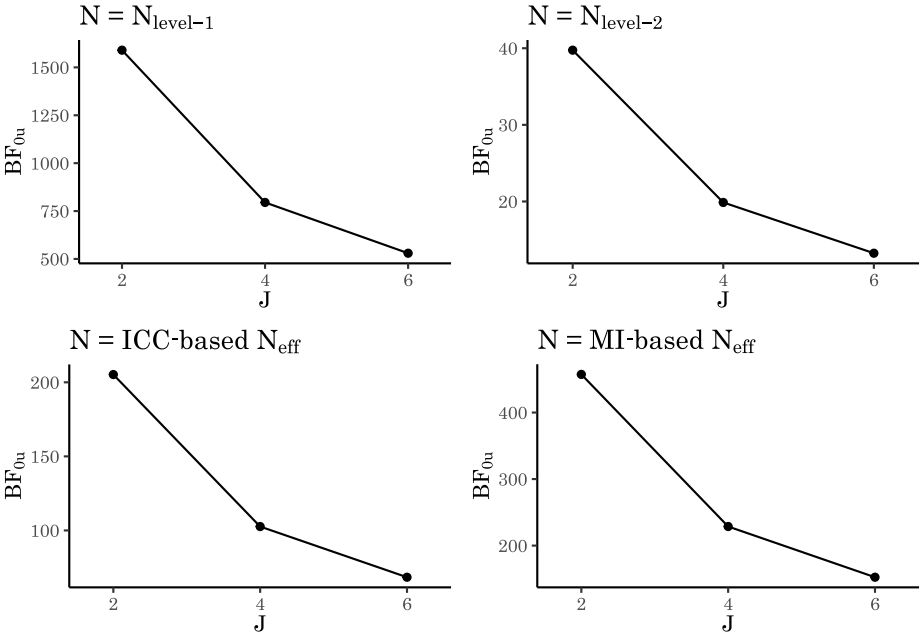

Figure 3: BFs for different valus of J and N.

The results given in Figure 1 illustrate the sensitivity of the BF when testing null hypotheses in the context of two-level models. Even though, all 12 BFs show support for the null hypothesis, there is a large discrepancy when varying J and the value for the sample size, clearly depicting the two issues that this paper is trying to address. Thus, we can conclude the following: (i) regardless

of the choice for the sample size, the BF is sensitive to the value for J; (ii) the magnitude of the BF depends on which sample size is being used. More specifically, the higher the sample size, the larger the magnitude of the BF, even when using the same value for J. Hence, the need to use a theoretically grounded value for the sample size, as the one introduced in the previous section. In the next section the value for J_{ref}, which is a theoretically grounded alternative to J, as proposed by Hoijtink (2021), is also introduced.

How to calculate J_{ref}?

As depicted in the previous section, there is no justification for using the default value for J as implemented in `bain` for NHBT. More specifically, this means that when calculating BFs by using J to test whether the fixed effects are equal to zero, the resulting values of the BFs change when varying the value for J even when testing the same hypotheses on one sample having *exactly* the same value for R^2_m. This allows for J to be chosen such that the BF is deliberately biased in favour or against the null hypothesis (also discussed in Tendeiro & Kiers, 2019). In other words, we say that the BF lacks clear operating characteristics.

As was already mentioned in the introduction section, the value for J, for which the BF = 19 when $R^2_m = 0$, is referred to as *reference* J (abbreviated as J_{ref}). In the context of multiple linear regression, Hoijtink (2021) states that J_{ref} can be derived by choosing a reference sample size N_{ref} and a reference Bayes factor BF_{ref}. This can be achieved by having N_{ref} equal to the observed sample size and choosing $BF_{ref} = 19$ if the observed effect size in the data $R^2 = 0$ (the motivation behind the number 19 has already been given in the introduction section). This setting allows us to straightforwardly generalize the approach given in Hoijtink (2021) in the context of two-level models by having $R^2 = R^2_m$ and N_{ref} = MI-based N_{eff}. Using this set-up and Equation 41 in Hoijtink (2021), given here as Equation 10, we can straightforwardly derive J_{ref}.

$$BF_{01} = (\frac{N_{ref}}{J})^{\frac{M}{2}} \exp(-\frac{N_{ref} - M - 1}{2}\frac{R^2}{1 - R^2}), \tag{10}$$

since $BF_{01} = BF_{ref} = 19$, $R^2_m = 0$ and $J = J_{ref}$, where M denotes the number of predictors in

the model, it follows that

$$J_{ref} = \frac{N_{ref}}{19^{2/M}}. \tag{11}$$

For the data set used in for the sensitivity analysis in the previous section, with an MI-based $N_{eff} = 921$, applying Equation 11 yields $J_{ref} \simeq 48.5$. This renders a $BF_{0u} \simeq 18.9$, that is, the support in the data is almost 19 times in favour of H_0. This result is exactly what we would expect since the data is simulated from a population in which $R^2_m = 0$, which means that the exact R^2_m in the data is slightly above zero, thus the BF is not exactly 19. In this example, the value of the *fraction b* is exactly 0.05, meaning that we can still interpret the resulting BF as an approximate Bayes Factor. All of the aforementioned can be done directly within the `wrapper function`, by setting the argument `jref` to `TRUE`. For the statistical details of this section please see the main paper (Hoijtink, 2021). Functions that calculate J_{ref} for different statistical models are openly available on the `bain` website.[8]

Simulation study

A small simulation study was performed to properly assess the effect of J_{ref} and the sample size on the BF when testing whether the fixed effects of two-level models are equal to zero. A 1000 data sets for each of the 16 cells of Figure 1 were simulated and two-level models of the form given in Equation 1[9] were fitted. Afterwards, the null hypothesis of the form given in Equation 2[10] was tested for each model. First, J_{ref} and the MI-based N_{eff} were used to calculate the BFs. Afterwards, the value for the sample size was changed, to assess its effect when using J_{ref}. Finally, the sensitivity of the BF to the estimation method (FML or REML) was tested.

[8] *https://informative-hypotheses.sites.uu.nl/software/bain/*

[9] In the case of one predictor the model specified in Equation 1 reduces to $Y_{ij} = \alpha + \beta_1 X_{1,ij} + u_{0j} + u_{1j}\ X_{1,ij} + \epsilon_{ij}$.

[10] In the case of one predictor, the null hypothesis given in Equation 2, reduces to $H_0 : \beta_1 = 0$.

Using the MI-based N_{eff}

Summaries of the calculated BFs for the fixed effects are given in Figures 4 and 5. As with the sensitivity analyses, the fixed effects were estimated using FML estimation.

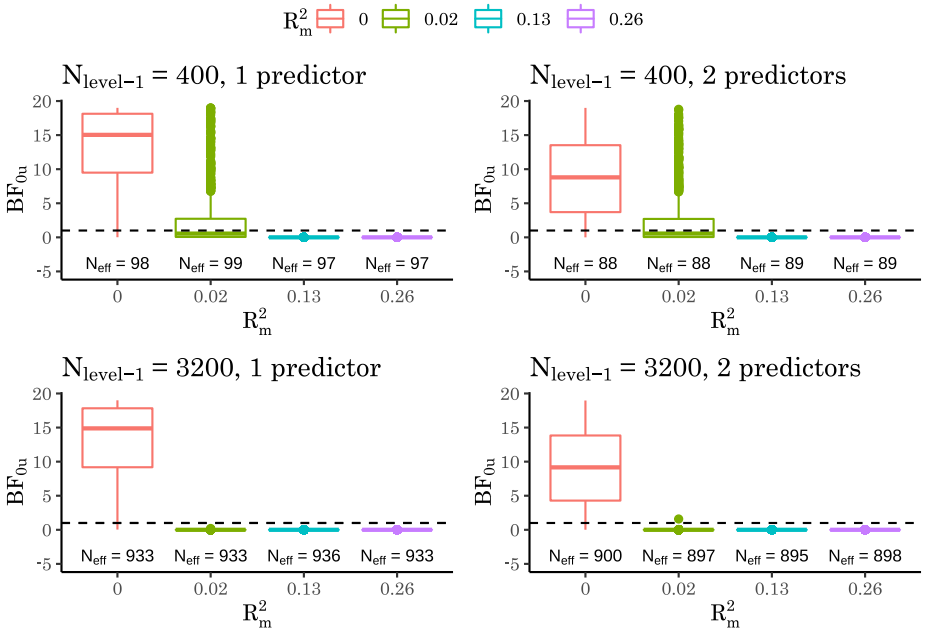

Figure 4: Boxplots of the resulting BFs for combinations of number of level-1 observations and number of predictors for each value of R^2_m. Below each box the respective average MI-based N_{eff} is given. The horizontal dashed line indicates where $BF_{0u} = 1$ i.e., there is equal support in the data for both H_0 and H_u.

- First, from Figure 4, it can be observed that in all situations where the data sets are sampled from a population where $R^2_m = 0$ (the boxes in pink), regardless of the sample size and the number of predictors, the BFs tend to range from around 1 to 19 which indicates that when applying J_{ref} the resulting BF_{0u}'s approach 19 as R^2_m tends to zero. It should be noted, yet again, that the data sets are repeatedly sampled from populations where $R^2_m = 0$, which results in fluctuations of the observed R^2_m's slightly above zero, hence explaining why the BFs are not *exactly* equal to 19. However, by extracting a data set where the R^2_m is a very small number, we can see that the resulting $BF_{0u} = 18.9$, this was also the case with the example data set in the previous section where the calculation of J_{ref} was illustrated.

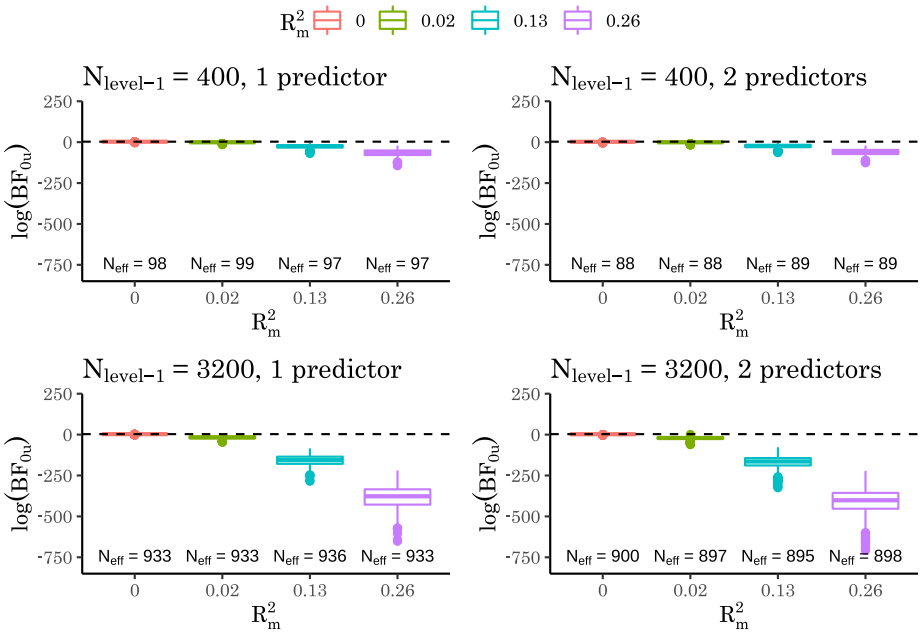

Figure 5: Boxplots of the natural logarithms of the resulting BFs for combinations of number of level-1 observations and number of predictors for each value of R^2_m. Below each box the respective average MI-based N_{eff} is given. The dashed horizontal line indicates the location of the natural logarithm of $BF_{0u} = 19$ i.e., $log(BF_{0u}) = 2.9$.

- Secondly, when having a medium (.13) and large (.26) effect sizes (i.e., H_0 is false) we can see that the BF consistently rejects the null. However, by comparing the boxes in green from the top two plots of Figure 4 with the bottom two plots, we can see that the value for the (effective) sample size plays an important role in rejecting the null when there is a small effect size (.02). More specifically, we can see that in situations when having an MI-based N_{eff} of 100 to 200 we do not have enough power to always reject the null (i.e., show support for H_u). This observation is in line with Cohen (1992), where in the context of multiple linear regression with two predictors, for a small effect size, a sample size of around 480 is needed to reject the null using the standard $\alpha = .05$ threshold. As the effect size becomes larger, the sample size needed to reject the null steeply drops to 67 and 30, for medium and large effect sizes, respectively. It should be noted, however, that this isn't an exact one-on-one comparison with the BF, since the guidelines given in Cohen (1992) are in terms of NHST p-values. In the case of testing only one fixed effect (i.e., having only one predictor) the BFs tend to be stronger on average, relative to the situation when testing two coefficients. In order to "zoom in" on the situations in which there is a medium or large effect size, the natural

logarithms of the same BFs are plotted in Figure 5.

- Finally, as the sample size increases so does the strength of the BFs, this is especially evident in the bottom subfigures of Figure 5 for the situation when $N_{level-1} = 3200$ and $N_{eff} \simeq 900$. It should be noted that this is only the case when having a non-zero effect size i.e., when H_0 is false. As the value of the effect size goes from small (.02) to medium (.13) and finally to strong (.26) the support in the data for the unconstrained hypothesis increases (i.e., there is no support in the data for the null hypothesis). Moreover, with a higher sample size, the distance between the boxes representing the BFs for the different values of the effect size becomes larger. In other words, as the effect size for the fixed effects and the sample size become bigger the resulting BFs become smaller and smaller (numbers very close to zero). It should also be noted that the inverse relationship between the strength of the BF and the number of parameters being tested diminishes as the value of the sample size increases.

Thus, we can conclude that when implementing J_{ref} and the MI-based N_{eff}, to test whether the fixed effects of linear two-level models are equal to zero, this BF yields clear operating characteristics. In other words, when R^2_m is very close to zero, the BF tends to 19, and as the R^2_m becomes larger the BF becomes smaller and tends to zero.

The sample size revisited

The analyses from the previous subsection were repeated three times using the remaining three options for the sample size of two-level models i.e., $N_{level-1}$, $N_{level-2}$ and the ICC-based N_{eff} were used to calculate J_{ref} and the BFs. The results were *exactly* the same as the ones presented in Figures 4 and 5. This represents a valuable observation since it means that when using J_{ref} the sample size does not have any influence on the BF anymore. The reasoning behind this is that the same value for the sample size is used twice: (i) first, as N_{ref} for the calculation of J_{ref} based on Equation 11; (ii) secondly, for the calculation of the BF within `bain`, based on Equation 4. This means that the ratio $\frac{J_{ref}}{N}$ remains constant, regardless of which value for N is used. However, the value for the sample size still remains relevant, since as described in the previous subsection, using an appropriate value for the sample size for the particular model is crucial in terms of having

enough power for the BF to correctly reject the null hypothesis when having a small effect size.

The Estimation Method: FML vs REML

For all 16000 data sets, the same models were fitted again by using REML estimation, with the aim of testing whether the estimation method affects the resulting BFs calculated when implementing J_{ref}. The BFs were calculated by using the MI-based N_{eff}. The results presented in Figures 6 and 7 quite clearly illustrate that there is almost no difference between the BFs obtained for models fitted with REML compared to the BFs presented in the previous subsections (for models fitted using FML). Moreover, it can be seen that the BF is slightly stronger on average when using REML, however, these differences are negligible. Thus, we conclude that the estimation method does not have an influence on the resulting BFs and users can make informed decisions about which estimation procedure to use based on other methodological considerations that are not related to testing the fixed effects.

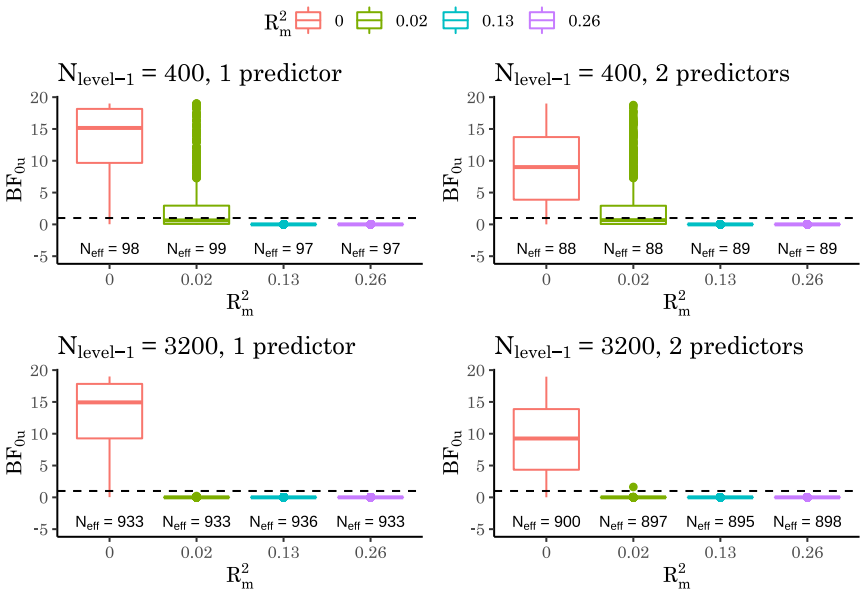

Figure 6: Boxplots of the resulting BFs based on models estimated with REML for combinations of number of level-1 observations and number of predictors for each value of R^2_m. Below each box the respective average MI-based N_{eff} is given. The horizontal dashed line indicates where $BF_{0u} = 1$ i.e., there is equal support in the data for both H_0 and H_u.

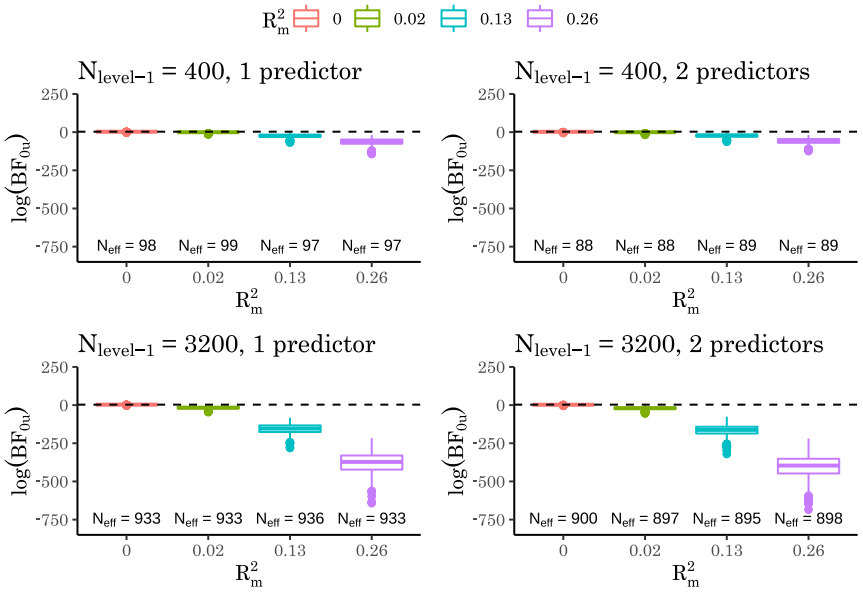

Figure 7: Boxplots of the natural logarithms of the resulting BFs based on models estimated with REML for combinations of number of level-1 observations and number of predictors for each value of R^2_m. Below each box the respective average MI-based N_{eff} is given. The dashed horizontal line indicates the location of the natural logarithm of $BF_{0u} = 19$ i.e., $log(BF_{0u}) = 2.9$.

Examples

By using the `tutorial` data, which is an openly-available two-level data set from the `R` package `R2MLwiN` (Zhang et al., 2016), we aim to illustrate how to use the default BF proposed in this paper. The data represents a subset from a larger data set of examination results from six inner London Education Authorities consisting of 4059 students nested within 65 schools. The variables used for the aims of this example are: (i) the standardized exam score for each student (`normexam`) which will serve as the outcome variable; (ii) the standardized score at age 11 on the London Reading Test (`standlrt`), which will serve as the level-1 predictor; (iii) the group indicator (`school`) with 65 schools of varying size. In Examples 3 and 4, a level-2 predictor in the form of the average LRT score for each school (`avslrt`) is included to illustrate that this approach can also be extended to test the coefficients of level-2 predictors. Additionally, an informative hypothesis stating that the parameters are *larger* than zero is included in all four examples. This section contains the most important parts of the `R` code for the first example only. However, interested researchers are referred to the aforementioned tutorial available on the authors' website. All the models were estimated using FML estimation and the results are presented in Table 3.

Example 1

```
# load the wrapper & packages
source("wrapper_function.R") # the wrapper
library(lme4)                # for fitting the model
library(jtools)              # for model fit summary and R^2_m
library(R2MLwiN)             # for the data

# load the data
data("tutorial")

# fit the two-level model
model.1 <- lmer(normexam ~ standlrt + (standlrt | school), REML = FALSE, data = tutorial)

# inspect model fit and R^2_m
summ(model.1)  # R^2_m = .32 (large effect size)

# calculate the MI-based N_eff (see the linked tutorial)

# define the hypotheses
hypotheses <- "standlrt = 0;
               standlrt > 0"

# calculate the BFs using Jref
BFs.1 <- bain_2lmer(model.1, hypotheses, standardize = FALSE,
                    N = MI_N_eff, seed = 123, jref = TRUE)
print(BFs.1)
# inspect the value of the fraction b
BFs.1$b
```

The code above shows all the necessary steps to obtain the BFs in `R`. After loading the data, we fit a two-level model with a random intercept and random slope for `standlrt`. In other words the model specified in Equation 1 becomes:

$$\text{normexam}_{ij} = \alpha + \beta_1 \text{ standlrt}_{ij} + u_{0j} + u_{1j} \text{ standlrt}_{ij} + \epsilon_{ij}. \tag{12}$$

The hypotheses of interest for this example are $H_0 : \beta_1 = 0$ and $H_i : \beta_1 > 0$. The `summ` function from the `R` package `jtools` (Long, 2020) automatically renders the R^2_m, however, how to calculate this manually is illustrated in the linked tutorial. In this case, $R^2_m = .32$, which represents a large effect size (according to Cohen, 1992, for linear regression). Thus, we expect to reject the null hypothesis, which states that the fixed effect is equal to zero. Afterwards, we calculate the MI-based N_{eff}, by specifying the same model in `JAGS`, which in this case renders a value of 874. Finally, we calculate the BFs by using the `wrapper function` with the argument `jref` set to `TRUE`. The resulting BF_{0u} is a very small number close to zero, which indicates there is no support in the data for the null hypothesis. We can take the inverse of BF_{0u} to obtain BF_{u0}, which in this case equals 1.092e+169, i.e., there is overwhelming support in the data for the unconstrained hypothesis. The BF_{iu} is around 2, which indicates that the support in the data is two times in favour of H_i. Moreover, we can easily obtain the BF of the informative hypothesis against the null hypothesis, by taking the ratio of their respective BFs against the unconstrained hypothesis. In this case, $BF_{i0} = 2.5\text{e+}181$, indicating that H_i is strongly preferred by the data. Finally, we inspect the value for the *fraction b*, which is around 0.003. This allows us to interpret the resulting BFs as *Approximate BFs*, since $b < 0.05$. Thus, we can state the following: by using the AAFBF (Gu et al., 2018) set to equal 19 when the *marginal* R^2 for the fixed effects is zero in the data (Hoijtink, 2021), the *approximate* BF of the informative hypothesis against the null hypothesis is 2.5e+181. In other words, we conclude that, *given the data*, the fixed coefficient for the predictor `standlrt` is larger than zero.

Example 2

In order to further illustrate the clear operating characteristics of this BF, we simulate the outcome `normexam`, by having the fixed effect for `standlrt` equal to zero and redo all the analyses. Now, $R^2_m = 0$ and MI-based $N_{eff} = 1070$. The $BF_{0u} = 18.93$ and the $BF_{iu} = 1.1$. In this case we conclude that *given the data*, the fixed coefficient for the predictor `standlrt` is is not different from zero. Subsequently, the BF of the null hypothesis against the informative hypothesis, $BF_{0i} = 16.5$. Thus, we say that the support in the data is 16 times in favour of the null hypothesis against the informative hypothesis, that is, *given the data* the fixed effect for `standlrt` is equal to zero, based on the *approximate* BF_{iu} (the value for the fraction b is still 0.003).

Example 3

Now, we fit a new model with a random intercept and random slope for `standlrt` and well as including the level-2 predictor `avslrt`. The model given in Equation 1 becomes:

$$\text{normexam}_{ij} = \alpha + \beta_1 \text{ standlrt}_{ij} + \beta_{1,2} \text{ avslrt}_j + u_{0j} + u_{1j} \text{ standlrt}_{ij} + \epsilon_{ij}, \tag{13}$$

where $\beta_{1,2}$ denotes the estimated coefficient for the level-2 predictor `avslrt`.

The R^2_m for this model is .35. Now, $H_0 : \beta_1 = \beta_{1,2} = 0$ and $H_i : \beta_1 > 0$ & $\beta_{1,2} > 0$. We calculate the BFs by using the ICC-based $N_{eff} = 358$ (calculated using the average n_c, see Equation 8), since currently, the MI-based N_{eff} has not yet been extended to include level-2 predictors and, as shown in the simulation study, the sample size does not influence the BF when using J_{ref}. This yields $BF_{0u} \simeq 0$ and $BF_{iu} \simeq 4.2$, with a value for b exactly equal to 0.05. Thus, we say that based on the approximate BF, there is substantial evidence in the data that both the fixed effect for `standlrt` and the level-2 coefficient for `avslrt` are larger than zero.

Example 4

We repeat this analysis by, yet again, simulating the outcome where both coefficients are zero i.e., the $R^2_m = 0$. This time we obtain a $BF_{0u} = 12.8$ and a $BF_{iu} = 0.3$, which translates to a

$BF_{0i} = 53.2$. In other words, the evidence in the data is 53 times in favour of H_0 (i.e., that the effects of both `standlrt` and `avslrt` are zero) against H_i (i.e., the effects of both `standlrt` and `avslrt` are larger than zero).

Table 3: Estimated fixed effects and their respective SE's for the four models

	Example 1		Example 2		Example 3		Example 4	
	est	SE	est	SE	est	SE	est	SE
α	-0.01	0.04	0.01	0.04	-0.00	0.04	-0.01	0.04
β_1	0.56	0.02	0.00	0.02	0.55	0.02	-0.01	0.02
β_{21}	/	/	/	/	0.29	0.11	-0.08	0.10

Recommendations for researchers

Researchers who want to use the approach proposed in this paper, to test whether the fixed effects of two-level models are equal to zero, can do so by using the `wrapper function` for `bain` which includes an option to directly implement J_{ref}. Concerning the value for the sample size, for now, users should decide for themselves whether they want to use the level-1 observations, the level-2 observations, the ICC-based N_{eff} or they can adjust the available code from the example in the linked tutorial to calculate the MI-based N_{eff}. In the future, user-friendly `R` functions might be available, such that researchers can also use the MI-based N_{eff} proposed in this paper for the sample size when calculating BFs or in general when they would want to calculate the effective sample size for two-level models that have random slopes. Since it was shown that the value of the sample size does not have an effect on the BF calculated when using J_{ref}, users who do not want to calculate the MI-based N_{eff} are advised to use the ICC-based N_{eff} instead. However, using the MI-based N_{eff} should be preferred since it represents a theoretically grounded value for the sample size of two-level models that include (random) predictors, since, as shown in the simulation, the sample size is important when capturing a small effect size.

Researchers should always explicitly state that they are using the AAFBF, with a value for J chosen such that it is required that the BF = 19 when the effect size (R^2_m) for the fixed effects is zero in the data and cite all relevant references. Honest reporting of such choices is crucial for the advancement of transparent practices in psychological research and research in the social and behavioural sciences in general. Furthermore, researchers should be careful when interpreting

the resulting BFs with regard to the value of the fraction b, which as previously explained, is an indication of the relative size of J vs N. As a rule of thumb, they are advised to use the value of 0.05, as proposed by Hoijtink (2021), such that when $b < 0.05$ they are able to interpret the resulting Bayes factors as an *approximate* BF; and when $b > 0.05$ they should interpret it as an information criterion inspired by the BF. As shown in the example code, the exact value for b can easily be obtained from the `R` list, where the results from calling the `wrapper function` are stored.

Discussion

Why use this BF?

When testing whether the fixed effects of mixed models are equal to zero, some statistical software packages perform a t-test based on Satterthwaite or Kenward-Roger approximations for the degrees of freedom or simply treat the t-value as a z-value and perform a Wald test (Luke, 2017), while others only report the t-statistic and completely omit the p-values due to issues regarding the calculation of the degrees of freedom (for example the `R` package `lme4`, Bates et al., 2015). A completely different approach is to use the Likelihood Ratio test (LR) and compare two models which only differ in the fixed effects of interest. Luke (2017) compares these different NHST approaches for testing the fixed effects of multilevel models and concludes that all suffer from some combination of low sample size and/or choice of the estimation method. More specifically, after the Wald test (which showed the worst performance out of all the considered approaches), the LR test suffered from low sample size and could only be used to test fixed effects from models estimated using FML. Additionally, it is not completely clear how to specify models that differ *only* in the fixed effects. The Satterthwaite approximation for the degrees of freedom performed better than the LR, however, when REML was applied, it yielded p-values that were not conservative enough. The use of Kenward-Roger approximation of the degrees of freedom and parametric bootstrapping were shown to yield the best results, however, the former required the use of REML and the latter suffered when having a small sample size.

The present study has successfully been able to calibrate a default BF for testing null hypotheses of the form given in Equation 2, by having the $BF_{0u} = 19$ when the effect size for the fixed effects

(R^2_m) is zero in the data. Additionally, it was shown that when using the newly-proposed J_{ref} the value for the *fraction b* remains constant regardless of the sample size. However, the effective sample size for two-level models containing random slopes, calculated with the newly proposed estimator is valuable in itself since it gives an indication of the effective level-1 observations after accounting for (part of) the within-group clustering by introducing the (random) predictors. Moreover, as it was shown, using an appropriate value for the sample size is crucial when having a small effect size.

Thus, with the above in mind, it is fair to conclude the following:

1. By using this default BF, researchers can choose between REML or FML to estimate the models, without any impact on the outcome of NHBT;
2. The sample size does not influence the value of the BF when using J_{ref}, however, a theoretically grounded value for the effective sample size of two-level models should be preferred;
3. Using this BF allows us to quantify the evidence in the data that is *in favour* of the null hypothesis, which is especially valuable when the null hypothesis is of main interest;
4. Using the BF helps researchers to move away from dichotomous decisions which are inherent to NHST;
5. This BF can be extended to include informative hypotheses;
6. Most importantly, the prior sensitivity of the BF has been adequately dealt with when testing whether the fixed effects of two-level models are *equal* to zero.

Recommendations for future research

This paper has successfully managed to derive a BF that has clear operating characteristics *only* when testing null hypotheses which state that the fixed effects of two-level models are *equal to zero* (as in Equation 2). This gives researchers a Bayesian alternative for the most widely used hypotheses about the fixed effects (i.e., no effect of the predictor). However, due to various theoretical reasons, we can quite easily imagine a situation where researchers would want to test whether the fixed effects are *equal to each other* i.e., they would want to test $H_0 : \beta_1 = \beta_2$. In that case, simply using

BF = 19, when $R^2_m = 0$ is not appropriate. A new approach where an additional parameter (for the situation with two predictors this will be β_3) should be introduced, which will represent the difference between β_1 and β_2 and the resulting R^2_m would represent the effect size for β_3. Thus, the BF will reach 19 when β_1 and β_2 are exactly equal to each other. Moreover, when testing null hypotheses of the aforementioned form, it only makes sense to use the *standardized* coefficients, since due to the scale of the predictors the resulting fixed effects might seem different while at the same time having the same magnitude. Standardizing the fixed parameters of a multilevel model, as with many of the other issues tackled in this paper, is not straightforward. First, we cannot use the standardized regression coefficients (like the ones in linear regression) directly, since there exists no way of obtaining their *standardized* covariance matrix (which is needed to calculate the BFs, see Equation 4). Secondly, when standardizing the data beforehand (which directly results in standardized coefficients), there remains an open question of whether to use, so-called, *overall standardization* or *within-group standardization* (see, Schuurman et al., 2016). The `wrapper function` already includes an option to test standardized fixed effects by using overall standardization of the data, which can be used as a starting point for future research.

One of the main limitations of the current approach is that it can *only* be used for testing *continuous* predictors. The introduction of categorical predictors, such as, for example, sex, assumes that the level-1 observations are drawn from different populations (defined by the categories of the factor variable). In such cases, the calculation of the BF in `bain` based on Equation 4 requires separate covariance matrices ($\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$) of the estimated parameters for each category of the predictor (for the details, see, Hoijtink, Gu, et al., 2019). Ignoring this requirement has been shown to yield inconsistent BFs.

Finally, expanding the new sample size calculation for higher-level models would be valuable and welcome. For now, the approach is limited to two-level models having continuous level-1 predictors, preferably with random slopes.

It is the authors' sincerest hope that: (i) applied researchers will make use of the work presented in this paper; (ii) in the near future the above-highlighted limitations of the present study will be addressed and the approach will be expanded to include more complex multilevel models.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods*, *1*(4), 355–373. https://doi.org/10.1177/109442819814001

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997. https://doi.org/10.1037/0033-2909.112.1.155

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association*, *95*(452), 1300–1304. https://doi.org/10.1080/01621459.2000.10474335

Gu, X., Hoijtink, H., Mulder, J., & van Lissa, C. J. (2021). *Bain: Bayes factors for informative hypotheses*. https://CRAN.R-project.org/package=bain

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 229–261. https://doi.org/10.1111/bmsp.12110

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Chapman & Hall/CRC. https://doi.org/10.1201/b11158

Hoijtink, H. (2021). Prior sensitivity of null hypothesis bayesian testing. *Psychological Methods*. https://doi.org/10.1037/met0000292

Hoijtink, H., Gu, X., & Mulder, J. (2019). Bayesian evaluation of informative hypotheses for multiple populations. *British Journal of Mathematical and Statistical Psychology*, *72*(2), 219–243. https://doi.org/10.1111/bmsp.12145

Hoijtink, H., Mulder, J., Lissa, C. van, & Gu, X. (2019). A tutorial on testing hypotheses using the bayes factor. *Psychological Methods*, *24*(5), 539. https://doi.org/10.1037/met0000201

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge. https://doi.org/10.4324/9781315650982

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, *44*(6), 1086–1105. https:

//doi.org/10.1080/02664763.2016.1193725

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*, 203–222. https://doi.org/10.1017/S030500410001330X

Johnson, P. C. (2014). Extension of nakagawa & schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946. https://doi.org/10.1111/2041-210X.12225

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *The Annals of Family Medicine*, *2*(3), 204–208. https://doi.org/10.1370/afm.141

Long, J. A. (2020). *Jtools: Analysis and presentation of social scientific data.* https://cran.r-project.org/package=jtools

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in r. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Mulder, J. (2014). Prior adjusted default bayes factors for testing (in) equality constrained hypotheses. *Computational Statistics & Data Analysis*, *71*, 448–463. https://doi.org/10.1016/j.csda.2013.07.017

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Plummer, M. (2003). *JAGS: A program for analysis of bayesian graphical models using gibbs sampling.*

Plummer, M. (2021). *Rjags: Bayesian graphical models using MCMC.* https://CRAN.R-project.org/package=rjags

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rubin, D. B. (1987). *Multiple imputation for survey nonresponse.* New York: Wiley.

Schuurman, N. K., Ferrer, E., Boer-Sonnenschein, M. de, & Hamaker, E. L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychological Methods*,

21(2), 206. https://doi.org/10.1037/met0000062

Tendeiro, J. N., & Kiers, H. A. (2019). A review of issues about null hypothesis bayesian testing. *Psychological Methods*, *24*(6), 774. https://doi.org/10.1037/met0000221

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman & Hall/CRC Press. https://doi.org/10.1201/9780429492259

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to markov chain monte–carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143–154. https://doi.org/10.3758/s13423-016-1015-8

Vasishth, S., & Sorensen, T. (2014). *Fitting linear mixed models using JAGS and stan: A tutorial*. https://www.ling.uni-potsdam.de/~vasishth/JAGSStanTutorial/SorensenVasishthMay12014.pdf

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/bf03194105

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*(2), 212. https://doi.org/10.1037/1082-989X.4.2.212

Zhang, Z., Parker, R. M. A., Charlton, C. M. J., Leckie, G., & Browne, W. J. (2016). R2MLwiN: A package to run MLwiN from within R. *Journal of Statistical Software*, *72*(10), 1–43. https://doi.org/10.18637/jss.v072.i10

Appendix A

A1. Prior distributions

This model specification is based on the tutorial by Vasishth & Sorensen (2014), with two notable differences: (i) we give the precisions gamma priors with shape and scale hyperparameters equal to 0.001; (ii) we do not set a normal prior for the correlation between the random effects (ρ), with a uniform hyperprior for its mean, instead we give a uniform prior to ρ directly.

The random intercepts and slopes are assumed to follow a multivariate normal distribution

$$\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma^{-1}}), \tag{14}$$

with a mean vector $\boldsymbol{\mu}$ containing the fixed effects

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_\alpha \\ \mu_{\beta_1} \\ \mu_{\beta_2} \end{pmatrix}. \tag{15}$$

These fixed effects are drawn from uninformative normal hyperprior distributions

$$\mu_\alpha \sim \mathcal{N}(0, 0.0001);\ \mu_{\beta_1} \sim \mathcal{N}(0, 0.0001);\ \mu_{\beta_2} \sim \mathcal{N}(0, 0.0001). \tag{16}$$

Note, the scaling hyperparameters of the normal hyperpriors are defined in terms of precisions (the inverse of the variance).

The prior of the inverse of the covariance matrix $\boldsymbol{\Sigma^{-1}}$ (i.e., the precision matrix) follows a Wishart distribution:

$$\boldsymbol{\Sigma^{-1}} \sim \mathcal{W}(\boldsymbol{\tau}, 3) \tag{17}$$

where $\boldsymbol{\tau}$ is a scale matrix containing the scaled (to the power of $-\frac{1}{2}$) precisions of the random effects,

$$\boldsymbol{\tau} = \begin{pmatrix} \tau_{\alpha}^2 & \tau_{\alpha,\beta_1}\rho_1 & \tau_{\alpha,\beta_2}\rho_2 \\ \tau_{\alpha,\beta_1}\rho_1 & \tau_{\beta_1}^2 & \tau_{\beta_1,\beta_2}\rho_3 \\ \tau_{\alpha,\beta_2}\rho_2 & \tau_{\beta_1,\beta_2}\rho_3 & \tau_{\beta_2}^2 \end{pmatrix}. \tag{18}$$

Each precision of the random effects follows an uninformative gamma distribution with a shape and rate hyperparameters of 0.001:

$$\tau_{\alpha}^2 \sim \mathcal{G}(0.001, 0.001);\ \tau_{\beta_1}^2 \sim \mathcal{G}(0.001, 0.001);\ \tau_{\beta_2}^2 \sim \mathcal{G}(0.001, 0.001). \tag{19}$$

The correlation coefficients are given a uniform prior distribution on the range from -1 to 1:

$$\rho_1 \sim \mathcal{U}(-1, 1);\ \rho_2 \sim \mathcal{U}(-1, 1);\ \rho_3 \sim \mathcal{U}(-1, 1). \tag{20}$$

Finally, the prior of the residual precision is also follows a gamma distribution with the same hyperparameters as the precisions of the random effects:

$$\tau_{\epsilon}^2 \sim \mathcal{G}(0.001, 0.001). \tag{21}$$

A2. JAGS Model Specification

```
model {
  for (i in 1:3200){   # Likelihood
    y[i] ~ dnorm(mu[i], tau)
    mu[i] <- alpha[group[i]] + beta_1[group[i]] * X1[i]
                             + beta_2[group[i]] * X2[i]
  }

  for(j in 1:80){   # level 2
    alpha[j] <- U[j,1]
    beta_1[j]  <-  U[j,2]
```

```
    beta_2[j]  <-  U[j,3]
    U[j,1:3] ~ dmnorm (MU[j,], invSigma[,])
    MU[j,1] <- mu.alpha
    MU[j,2] <- mu.beta_1
    MU[j,3] <- mu.beta_2
  }
  mu.alpha  ~ dnorm(0, 0.0001)
  mu.beta_1 ~ dnorm(0,  0.0001)
  mu.beta_2 ~ dnorm(0,  0.0001)
  tau ~ dgamma (0.001, 0.001)
  invSigma[1:3,1:3] ~ dwish(Tau, 3)
  tau.alpha  ~ dgamma (0.001, 0.001)
  tau.beta_1 ~ dgamma (0.001, 0.001)
  tau.beta_2 ~ dgamma (0.001, 0.001)
  Tau[1,1] <- pow(tau.alpha,  -1/2)
  Tau[2,2] <- pow(tau.beta_1, -1/2)
  Tau[3,3] <- pow(tau.beta_2, -1/2)
  Tau[1,2] <- rho_1*tau.alpha*tau.beta_1
  Tau[2,1] <- Tau[1,2]
  Tau[1,3] <- rho_2*tau.alpha*tau.beta_2
  Tau[3,1] <- Tau[1,3]
  Tau[2,3] <- rho_3*tau.beta_1*tau.beta_2
  Tau[3,2] <- Tau[2,3]
  rho_1 ~ dunif(-1, 1)
  rho_2 ~ dunif(-1, 1)
  rho_3 ~ dunif(-1, 1)
}
```

A3. Bayesian and FML Estimates

Table 4: Estimates from fitting the two-level model with JAGS (Bayesian est.) and lmer (FML est.)

	JAGS			lmer		
	Fixed effects		Random effects	Fixed effects		Random effects
	est	SD	var	est	SE	var
α	0.03	0.04	0.09	0.03	0.03	0.08
β_1	0.03	0.02	0.01	0.03	0.02	0.01
β_2	-0.03	0.03	0.03	-0.03	0.03	0.03

Estimates from fitting the model on a data set corresponding to cell 13 of Figure 1. On the left hand side: the posterior mean estimates for the fixed effects obtained by fitting the specified model given in Appendix A2 with JAGS. On the right hand side: the FML estimates obtained by fitting the same model using lmer. As can be seen, the point estimates and their Standard Deviations/Errors are approximately the same, hence illustrating the use of uninformative prior distributions.

A4. Multiple Imputation Equations

The following equations are taken from Van Buuren (2018, Ch. 2.3).

Combined estimate:

$$\bar{\boldsymbol{\eta}} = \frac{1}{m}\sum_{l=1}^{m}\hat{\boldsymbol{\eta}}_l, \tag{22}$$

where $\bar{\boldsymbol{\eta}}$ denotes the combined estimate over all m imputed data sets and $\hat{\boldsymbol{\eta}}_l$ denotes the estimate of the l^{th} imputed data set. In the case of more than one parameter (as in this paper) $\bar{\boldsymbol{\eta}}$ is a vector.

Average of the complete-data variances:

$$\bar{U} = \frac{1}{m}\sum_{l=1}^{m}\hat{U}_l, \tag{23}$$

where $\bar{U}$ represents the average covariance matrix of the estimated parameters coming from all the m imputed data sets and $\hat{U}_l$ denotes the covariance matrix coming from the l^{th} imputed data set.

Unbiased estimate of the variance between the m complete estimates:

$$B = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{\boldsymbol{\eta}}_l - \bar{\boldsymbol{\eta}})(\hat{\boldsymbol{\eta}}_l - \bar{\boldsymbol{\eta}})^T. \tag{24}$$

Total variance:

$$T = \bar{U} + (1 + \frac{1}{m})B. \tag{25}$$

Proportion of variation attributable to the missing data (a compromise over all estimates):

$$\bar{\lambda} = (1 + \frac{1}{m})tr(BT^{-1})/k, \tag{26}$$

where k denotes the number of parameters in $\bar{\boldsymbol{\eta}}$.

Degrees of freedom:

$$\nu_{old} = \frac{m-1}{\lambda^2};\ \nu_{com} = n - k;\ \nu_{obs} = \frac{\nu_{com}+1}{\nu_{com}+3}\nu_{com}(1-\lambda);\ \nu = \frac{\nu_{old}\nu_{obs}}{\nu_{old}+\nu_{obs}}. \tag{27}$$

Fraction of missing information:

$$\gamma = \frac{\nu+1}{\nu+3}\lambda + \frac{2}{\nu+3}. \tag{28}$$

Effective sample size:

$$\text{MI-based } N_{eff} = N_{level-1} - \gamma\ N_{level-1} \tag{29}$$