Bayesian Analysis (2018) **TBA**, Number TBA, pp. 1–32

Bayes Factor Testing of Multiple Intraclass Correlations

Joris Mulder* and Jean-Paul Fox†

Abstract. The intraclass correlation plays a central role in modeling hierarchically structured data, such as educational data, panel data, or group-randomized trial data. It represents relevant information concerning the between-group and within-group variation. Methods for Bayesian hypothesis tests concerning the intraclass correlation are proposed to improve decision making in hierarchical data analysis and to assess the grouping effect across different group categories. Estimation and testing methods for the intraclass correlation coefficient are proposed under a marginal modeling framework where the random effects are integrated out. A class of stretched beta priors is proposed on the intraclass correlations, which is equivalent to shifted F priors for the between groups variances. Through a parameter expansion it is shown that this prior is conditionally conjugate under the marginal model yielding efficient posterior computation. A special improper case results in accurate coverage rates of the credible intervals even for minimal sample size and when the true intraclass correlation equals zero. Bayes factor tests are proposed for testing multiple precise and order hypotheses on intraclass correlations. These tests can be used when prior information about the intraclass correlations is available or absent. For the noninformative case, a generalized fractional Bayes approach is developed. The method enables testing the presence and strength of grouped data structures without introducing random effects. The methodology is applied to a large-scale survey study on international mathematics achievement at fourth grade to test the heterogeneity in the clustering of students in schools across countries and assessment cycles.

Keywords: Intraclass correlations, Bayes factors, stretched beta priors, shifted F priors, hierarchical models.

1 Introduction

The intraclass correlation plays a central role in the statistical analysis of hierarchical data. It quantifies the relative variation between groups or clusters. A large (small) intraclass correlation implies a strong (weak) degree of clustering which implies that there is much (little) variation between groups. In cluster-randomized trials, entire groups (e.g., hospitals, schools) are assigned to the same treatment or intervention. When planning a cluster-randomized experiment, the intraclass correlation is used as an indicator of the level of efficiency of a multistage sample design. Optimal sample size requirements to obtain adequate statistical power and statistical precision depend on the variation between and within groups (Hedges and Hedberg, 2007; Raudenbush, 1997; Spiegelhalter, 2001). When conducting an experiment in different regions and contexts, the statistical variation in intraclass correlations is relevant to optimally plan cluster-randomized

*Tilburg University, Tilburg, The Netherlands, j.mulder3@uvt.nl
†University of Twente, Enschede, The Netherlands, j.p.fox@utwente.nl

 https://doi.org/10.1214/18-BA1115

experiments across regions and to obtain adequate statistical power in each region. Knowledge about the intraclass correlation is also important to verify that conclusions of a statistical analysis are valid. When incorrectly ignoring a grouping effect, standard errors are generally too small and conclusions about the statistical significance of a treatment effect might be incorrect (Raudenbush, 1997).

Testing intraclass correlations can reveal relevant information about the level of heterogeneity between groups and across different group types. For example, Mulder and Fox (2013) tested the intraclass correlation of Catholic schools and public schools to learn that there is more variation in performance of Catholic schools in comparison to public schools. Van Geel et al. (2017) examined differences in intraclass correlations of teacher scores nested in schools in a pretest-posttest study design. After the teachers participated in an intervention program to improve teacher performances, a decrease of the intraclass correlation was measured. It was assumed that at the posttest teachers performed less alike leading to less similarity between teachers in each school, where some teachers did improve their performances while others did not. We will propose Bayes factor tests to formally test differences between intraclass correlations to be able to make inferences about the heterogeneity in teacher improvements.

In this paper, a Bayesian approach is presented for testing multiple precise and order hypotheses on multiple intraclass correlations belonging to different group categories, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_C)'$, where ρ_c is the intraclass correlation in group category c, for $c = 1, \ldots, C$. The intraclass correlation ρ_c is defined as the ratio of the between-groups variance in group category c and the total variance in group category c. The Q hypotheses have the following general form with equality and order constraints on intraclass correlations:

$$H_q : \mathbf{R}_q^E \boldsymbol{\rho} = \mathbf{0}, \mathbf{R}_q^I \boldsymbol{\rho} > \mathbf{0}, \tag{1}$$

for $q = 1, \ldots, Q$, where the rows of the coefficient matrices $\mathbf{R}_q^E$ and $\mathbf{R}_q^I$ are permutations of either $(1, -1, 0, \ldots, 0)$ or $(\pm 1, 0, \ldots, 0)$, $q = 1, \ldots, Q$. Thus, restrictions are considered where intraclass correlations are equal to, larger than, or smaller than zero, or equal to, larger than, or smaller than other intraclass correlations. This class covers the most important hypotheses on intraclass correlations in statistical practice.

A key step in our methodology is the use of a marginal modeling framework, where the random effects in the multilevel model are integrated out. In this marginal modeling framework the intraclass correlations can attain negative values (Searle et al., 1992, p. 60–61). The allowed parameter space under the marginal model is in line with the restriction following from the expression for the intraclass correlation of Harris (1913), which states that the intraclass correlation is greater than $-\frac{1}{p-1}$, where p equals the number of observations per group. Unlike the marginal modeling framework of Liang and Zeger (1986) using generalized estimating equations, our marginal approach connects more closely to integrated likelihood methods where the nuisance parameters are integrated out (Berger et al., 1999). In this integrated likelihood approach inferences concerning the intraclass correlations are also invariant under shifts of the random group means. In our approach, the integrated likelihood is defined for Helmert-transformed grouped observations (Lancaster, 1965). The orthonormal Helmert transformation is

used to partition the integrated likelihood in a component containing the between-groups sum of squares and a component containing the within-groups sum of squares, which are the sufficient statistics for the between-groups variance and within-groups variance, respectively (Fox et al., 2017).

To aid Bayesian estimation and testing a class of stretched beta priors is proposed for the intraclass correlations. This class of priors has positive support for negative intraclass correlations under the marginal model. Furthermore this class of priors is equivalent to shifted F distributions for the between-groups variances which has an additional shift parameter. To our knowledge this class of priors is novel in the Bayesian literature. Note that the F distribution is equivalent with the scaled-beta2 prior (Pérez et al., 2017) and the half-t prior (Gelman, 2006; Polson and Scott, 2012), which are becoming increasingly popular for modeling variance components (Mulder and Pericchi, 2018).

The proposed class of stretched beta priors under the marginal model has several attractive features. By allowing intraclass correlations to be negative it is possible to test the appropriateness of a random effects model using the posterior probability that an intraclass correlation is positive. Moreover using a noninformative improper prior under the marginal model, we can obtain accurate coverage rates for the credible intervals, even in the case of samples of minimal size with two groups and two observations per group for a zero intraclass correlation in the population. Note that frequentist matching priors play an important role in objective Bayesian analysis (Welch and Peers, 1963; Severini et al., 2002; Berger and Sun , 2008). Another consequence of the marginal modeling approach is that significance type tests of whether an intraclass correlation equals zero can be performed using credible intervals with accurate error rates. This is possible because testing whether $\rho = 0$ is not a boundary problem. Another important property of the proposed class of priors is that it can be made conditionally conjugate through a parameter expansion. As will be shown the shifted F distribution on the between-groups variance is equivalent to a gamma mixture of shifted inverse gamma distributions. These shifted inverse gamma priors are conditionally conjugate under the marginal model. This results in efficient posterior sampling with a Gibbs sampler.

For the testing problem (1), a Bayes factor testing procedure is proposed under the marginal model. This test can be applied when prior information about the intraclass correlations is available and when no prior information is available or when a default Bayesian procedure is preferred. In the informative case, proper truncated stretched beta priors are specified on the unique intraclass correlations under each constrained hypothesis H_q where the hyperparameters can be elicited from prior knowledge. A special case is the uniform prior, which assumes that all intraclass correlations are equally likely a priori. In the noninformative case, truncated improper reference priors will be used in combination with a generalized fractional Bayes approach (O'Hagan, 1995; De Santis and Spezzaferri, 2001; Hoijtink et al., 2018).

The paper is organized as follows. First, the marginal model is introduced, where two parameterizations are discussed and the integrated likelihood of the Helmert-transformed observations is given. Then, two prior classes are discussed, where a stretched beta distribution and a shifted F distribution is introduced to describe the distribution of the

intraclass correlation and the between-groups variance, respectively, while taking account of restrictions on the parameter space to ensure that the covariance matrix is positive definite. A Gibbs sampler is then described, and its performance is evaluated through a simulation study. Then a Bayes factor and a generalized fractional Bayes factor are proposed, and their numerical performances are evaluated. Both tests are applied to data from the Trends in International Mathematics and Science Study to evaluate hypotheses concerning the heterogeneity of the intraclass correlation across countries and assessment cycles. Finally, a discussion is given and some generalizations are presented.

2 The marginal model

We focus on the random intercept model, where measurement j in group (or cluster) i in group category c is distributed according to

$$\begin{aligned} y_{cij} &= \mathbf{x}'_{cij}\boldsymbol{\beta} + \delta_{ci} + \epsilon_{cij}, \text{ where} \\ \delta_{ci} &\sim \mathcal{N}(0, \tau_c^2), \\ \epsilon_{cij} &\sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{2}$$

for $j = 1, \ldots, p$ measurements, $i = 1, \ldots, n_c$ groups in category c, and $c = 1, \ldots, C$ categories. In (2), $\boldsymbol{\beta}$ is a vector of K fixed effects with covariates $\mathbf{x}_{cij}$ for measurement j in group i in category c, δ_{ci} is the random intercept of group i in category c, τ_c^2 is the between-groups variance in category c, and σ^2 is the common residual variance, which can be interpreted as the within-groups variance. This random intercept model can be recognized as a two-level multiple-group model, where level-1 units j are nested in level-2 groups i for each group category c. For instance, in each country c, math scores y_{cij} of students j nested in schools i are assumed to be independently distributed given the random school intercept δ_{ci}. The dependencies between student scores within each school can vary across countries.

The marginal model is obtained by integrating out the random effects δ_{ic}. The vectorized version of (2) then has a multivariate normal distribution with a covariance matrix having a compound symmetry structure, i.e.,

$$\mathbf{y}_{ci} \sim N\left(\mathbf{X}_{ci}\boldsymbol{\beta}, \boldsymbol{\Sigma}_c\right), \text{ with } \boldsymbol{\Sigma}_c = \sigma^2\mathbf{I}_p + \tau_c^2\mathbf{J}_p, \tag{3}$$

where $\mathbf{y}_{ci} = (y_{ci1}, \ldots, y_{cip})'$, $\mathbf{X}_{ci}$ is the $p \times K$ stacked matrix of covariates, $\mathbf{I}_p$ is the $p \times p$ identity matrix, and $\mathbf{J}_p$ is a $p \times p$ matrix of ones. In order for the covariance matrix $\boldsymbol{\Sigma}_c$ to be positive definite, it must hold that $\tau_c^2 > -\frac{\sigma^2}{p}$ and $\sigma^2 > 0$, and thus τ_c^2 does not necessarily have to be positive as in (2). For this reason we introduce a more general marginal model with covariance matrix

$$\boldsymbol{\Sigma}_c = \sigma^2\mathbf{I}_p + \eta_c\mathbf{J}_p, \tag{4}$$

where $\eta_c > -\frac{\sigma^2}{p}$. We shall refer to η_c as the generalized between-groups variance in category c. Note that (4) is equivalent to (3) when $\eta_c > 0$. Furthermore when there is support in the data that $\eta_c < 0$, the multilevel model (2) may not be appropriate.

We can reparameterize model (4) using different intraclass correlations for different categories, denoted by $\rho_c = \frac{\eta_c}{\eta_c+\sigma^2}$, for $c = 1, \ldots, C$, and the total variance in group category 1, denoted by ϕ^2, such that

$$\left\{ \begin{array}{rcl} \rho_c & = & \frac{\eta_c}{\eta_c+\sigma^2}, \\ & & \text{for } c = 1, \ldots, C, \\ \phi^2 & = & \eta_1 + \sigma^2 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{rcl} \eta_c & = & \frac{\rho_c}{1-\rho_c}(1-\rho_1)\phi^2, \\ & & \text{for } c = 1, \ldots, C, \\ \sigma^2 & = & (1-\rho_1)\phi^2. \end{array} \right. \tag{5}$$

Note that the total variance ϕ^2 and the fixed effects $\boldsymbol{\beta}$ are considered nuisance parameters in the current paper. The intraclass correlation in group category c is defined as the ratio of the generalized between-groups variance and the total variance. Thus, ρ_c quantifies how much units in the same group resemble each other in category c. If $\rho_c = 0$, then there is no clustering and measurements y_{ijc} are essentially randomly assigned to the groups in category c.

Using the parameterization $(\boldsymbol{\beta}, \boldsymbol{\rho}, \phi^2)$, the marginal model in (4) is given by

$$\mathbf{y}_{ci} \sim N\left(\mathbf{X}_{ci}\boldsymbol{\beta}, \boldsymbol{\Sigma}_c\right), \text{ with } \boldsymbol{\Sigma}_c = \phi^2(1-\rho_1)\left(\mathbf{I}_p + \frac{\rho_c}{1-\rho_c}\mathbf{J}_p\right), \tag{6}$$

with $\rho_c \in (-\frac{1}{p-1}, 1)$ in order for $\boldsymbol{\Sigma}_c$ to be positive definite. Hence, the intraclass correlations can attain negative values under this generalized marginal model, which is not the case in the conditional model (2), where $\rho_c \in (0, 1)$, for $c = 1, \ldots, C$. To get some intuition about the impact of a negative intraclass correlation, Figure 1 displays the sampling distribution of the between-groups sums of squares for population values of $\sigma^2 = 1$, $\beta = 0$, $n_1 = 8$, and $p = 4$, and intraclass correlations of $\rho_1 = -.1,\ 0$, or $.3$. As can be seen the between-groups sums of squares is generally smaller in the case ρ_1 is negative in comparison to $\rho_c = 0$ corresponding to random group assignment. Note that the estimated intraclass correlation is negative when the mean between-groups sums of squares is smaller than the mean within-groups sums of squares (Searle et al., 1992, p. 60–62).

Due to the compound symmetry covariance structure, the orthonormal Helmert transformation is useful to obtain transformed outcomes that are independent. The $p \times p$ Helmert transformation matrix is given by (Lancaster, 1965)

$$\mathbf{H}_p = \begin{bmatrix} \frac{1}{\sqrt{p}} & \frac{1}{\sqrt{p}} & \frac{1}{\sqrt{p}} & \cdots & \frac{1}{\sqrt{p}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \frac{1}{\sqrt{p(p-1)}} & \frac{1}{\sqrt{p(p-1)}} & \frac{1}{\sqrt{p(p-1)}} & \cdots & -\frac{p-1}{\sqrt{p(p-1)}} \end{bmatrix}.$$

Subsequently, the transformed observations are independently distributed according to

$$\mathbf{z}_{ci} \quad = \quad \mathbf{H}_p\mathbf{y}_{ci} \sim N\left(\mathbf{W}_{ci}\boldsymbol{\beta}, \phi^2(1-\rho_1)\mathbf{D}_c\right), \tag{7}$$

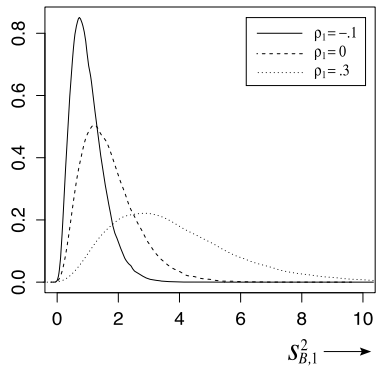

Figure 1: Sampling distribution of the between-groups sums of squares, $s^2_{B,1} = \sum_i (\bar{y}_{1i} - \bar{y}_1)^2$, where $\bar{y}_{1i}$ denotes the sample mean of group i and $\bar{y}_1$ denotes the overall sample mean, for $\phi^2 = 1$, $\beta = 0$, $n_1 = 8$, $p = 4$, and different values of the intraclass correlation $\rho_1 \in \{-.1, 0, .3\}$.

where $\mathbf{W}_{ci} = \mathbf{H}_p \mathbf{X}_{ci}$, and $p \times p$ matrix $\mathbf{D}_c = \text{diag}(\frac{1+(p-1)\rho_c}{1-\rho_c}, 1, \ldots, 1)$, for $i = 1, \ldots, n_c$ and $c = 1, \ldots, C$. From $\mathbf{D}_c$ it can be seen that only the first transformed observation, z_{ci1}, contains information about the intraclass correlation in that group. This can be explained by the fact that z_{ci1} depends on the sum of $\mathbf{y}_{ci}$, which is a key quantity for the between-groups variation.

The likelihood function under the marginal model is given by

$$\begin{aligned} f(\mathbf{z}|\mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\rho}, \phi^2) \quad = \quad & (2\pi)^{-\frac{Np}{2}} (\phi^2)^{-\frac{Np}{2}} (1-\rho_1)^{-\frac{Np}{2}} \\ & \exp\left\{ - \frac{\sum_{c=1}^{C} \sum_{i=1}^{n_c} \sum_{j=2}^{p} (z_{cij} - \mathbf{w}'_{cij}\boldsymbol{\beta})^2}{2\phi^2(1-\rho_1)} \right\} \\ & \prod_{c=1}^{C} \left(\tfrac{1+(p-1)\rho_c}{1-\rho_c} \right)^{-\frac{n_c}{2}} \exp\left\{ - \frac{\sum_{i=1}^{n_c} (z_{ci1} - \mathbf{w}'_{ci1}\boldsymbol{\beta})^2}{2\phi^2(1-\rho_1)\left(\frac{1+(p-1)\rho_c}{1-\rho_c}\right)} \right\}, \end{aligned} \tag{8}$$

where $\mathbf{z}' = (\mathbf{z}'_{11}, \mathbf{z}'_{12}, \ldots, \mathbf{z}'_{Cn_C})$, $\mathbf{W}$ is a stacked matrix of $\mathbf{W}_{ci}$, $\mathbf{w}'_{cij}$ is the j-th row of $\mathbf{W}_{ci}$, and $N = \sum_{c=1}^{C} n_c$. Note that because the Helmert transformation is orthonormal, the likelihood of $\mathbf{z}$ given $\mathbf{W}$ is equivalent to the likelihood of the untransformed $\mathbf{y}$ given $\mathbf{X}$. Further note that inferences are only invariant of the chosen category of η_c in $\phi^2 = \eta_c + \sigma^2$ in (5) when placing a noninformative improper prior on ϕ^2. This can be seen when setting the improper prior $\pi^N(\phi^2) = \phi^{-2}$ and integrating out ϕ^2 in the posterior. In that case each ρ_c will have the same role in the posterior[1].

[1]When $\pi^N(\phi^2) = \phi^{-2}$, it holds that $\int f(\mathbf{z}|\mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\rho}, \phi^2) \pi^N(\phi^2) d\phi^2 = \pi^{-\frac{Np}{2}} \Gamma(\frac{Np}{2}) \prod_{c=1}^{C} (\frac{1+(p-1)\rho_c}{1-\rho_c})^{-\frac{n_c}{2}} (\sum_{c=1}^{C} \sum_{i=1}^{n_c} \sum_{j=2}^{p} (z_{cij} - \mathbf{w}'_{cij}\boldsymbol{\beta})^2 + \frac{\sum_{i=1}^{n_c} (z_{ci1} - \mathbf{w}'_{ci1}\boldsymbol{\beta})^2}{(1+(p-1)\rho_c)(1-\rho_c)^{-1}})^{-\frac{Np}{2}}$.

3 Prior specification

We propose the following class of priors under the marginal model:

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \boldsymbol{\rho}, \phi^2) &= \pi(\boldsymbol{\beta}|\boldsymbol{\rho}, \phi^2)\pi(\phi^2)\prod_{c=1}^{C}\pi(\rho_c), \text{ with} \\
\pi(\rho_c) &= beta(\rho_c|\alpha_c, \zeta_c, -\tfrac{1}{p-1}, 1), \text{ for } c = 1, \ldots, C, \\
\pi(\boldsymbol{\beta}|\boldsymbol{\rho}, \phi^2) &= \mathcal{N}\left(\boldsymbol{\beta}_0, g(\mathbf{X}'\boldsymbol{\Sigma}_N^{-1}\mathbf{X})^{-1}\right) \\
\pi(\phi^2) &= \phi^{-2},
\end{aligned} \tag{9}$$

where the stretched beta distribution with shape parameters α_c and ζ_c in the interval $(-\frac{1}{p-1}, 1)$ is given by,

$$beta(\rho_c|\alpha_c, \zeta_c, -\tfrac{1}{p-1}, 1) = Q(\alpha_c, \zeta_c, p)(1 + (p-1)\rho_c)^{\alpha_c - 1}(1 - \rho_c)^{\zeta_c - 1}, \tag{10}$$

with normalizing constant $Q(\alpha_c, \zeta_c, p) = \frac{\Gamma(\alpha_c+\zeta_c)(p-1)^{\zeta_c}}{\Gamma(\alpha_c)\Gamma(\zeta_c)p^{\alpha_c+\zeta_c-1}}$, for $\alpha_c, \zeta_c > 0$. To our knowledge this prior is novel in the Bayesian literature. In the case of a single intraclass correlation, Spiegelhalter (2001) proposed a beta prior for ρ in the interval $(0, 1)$ under the conditional model (2). The stretched beta prior in (10) in the interval $(-\frac{1}{p-1}, 1)$ seems more natural however, because the prior has common factors as the likelihood function (8). As a result this class of priors is conditionally conjugate for the marginal model by applying a parameter expansion. This will be shown in the following section. Other generalizations that have been proposed for the beta distribution include Armagan et al. (2011).

Further note that the conditional prior for the nuisance parameters $\boldsymbol{\beta}$ is based on Zellner's (1986) g prior with prior guess $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_N = \text{diag}(\mathbf{I}_{n_1} \otimes \boldsymbol{\Sigma}_1, \ldots, \mathbf{I}_{n_C} \otimes \boldsymbol{\Sigma}_C)$ of dimension $Np \times Np$, with $\boldsymbol{\Sigma}_c$ given in (6), and $\mathbf{X}$ is the stacked matrix of $\mathbf{X}_{ci}$. An improper prior is set for the nuisance parameter ϕ^2 (similar as in the g prior). Note that by setting $g = Np$ one would obtain a unit information prior (see also Kass and Wasserman, 1995).

If prior information is available about the relative grouping effect in the different categories, this can be translated to informative stretched beta priors using a method of moments estimator. First note that the first two moments of a stretched beta prior equal

$$\begin{aligned}
E\{\rho_c\} &= \frac{\alpha_c p}{(\alpha_c + \zeta_c)(p-1)} - (p-1)^{-1}, \\
\text{var}(\rho_c) &= \frac{\alpha_c \zeta_c p^2}{(\alpha_c + \zeta_c)^2(\alpha_c + \zeta_c + 1)(p-1)^2}.
\end{aligned}$$

These expressions can be derived by transforming a beta(α_c, ζ_c) distribution in the interval $(0, 1)$ to a stretched beta distribution in the $(-\frac{1}{p-1}, 1)$. Subsequently, the prior hyperparameters α_p and ζ_c can be obtained by setting the prior guess equal to the mean

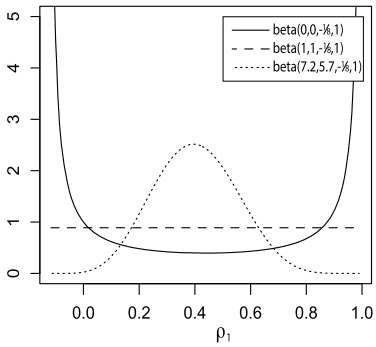

Figure 2: Three examples of stretched beta priors when $p = 9$: the reference prior with $\alpha = \zeta = 0$ (solid line), the uniform prior with $\alpha = \zeta = 1$ (dashed line), and an informative prior that is concentrated around .4 with $\alpha = 7$ and $\zeta = 8$ (dotted line).

and uncertainty about the prior guess equal to the standard deviation[2]. If all values for the intraclass correlations are assumed to be equally likely a priori, the hyperparameters can be set to 1 resulting in uniform priors. Figure 2 displays a uniform prior (dashed line) and an informative prior with prior guess $\rho_1^* = .4$ and standard deviation $s_{\rho_1^*} = .15$ (dotted line) when $p = 9$.

If prior information is absent or if one prefers to adopt an objective Bayesian procedure, the hyperparameters can be set to $\alpha_c = \zeta_c = 0$, for $c = 1, \ldots, C$. The resulting noninformative improper prior is given by

$$\pi^N(\boldsymbol{\beta}, \boldsymbol{\rho}, \phi^2) \quad = \quad \phi^{-2} \prod_{c=1}^{C} (1 + (p-1)\rho_c)^{-1}(1-\rho_c)^{-1}. \tag{11}$$

This is essentially Haldane's (1932) prior for ρ_c in the interval $(-\frac{1}{p-1}, 1)$. Note that (11) corresponds to (10) when defining $Q(0, 0, p) = 1$. In the case of a single intraclass correlation, this corresponds to the reference prior where the intraclass correlation is considered to be the most important parameter (Berger and Bernardo, 1992; Chung and Dey, 1998). This prior is equivalent to the prior considered by (Box and Tiao, 1973, p. 251). Figure 2 displays the reference prior when $p = 9$ (solid line).

In practice, intraclass correlations are generally expected to be positive. Such expectations can be included in the proposed prior by truncating the stretched beta priors on ρ_c in the interval $(0, 1)$. Working with this truncated prior essentially comes down to the marginal model of the random effects model in (3) and (2). Note that this truncated prior differs from a standard beta prior in the interval $(0, 1)$ (except in the case of uniform priors). For example Chung and Dey (1998) truncated the noninformative

[2]If ρ_c^* denotes the prior guess and $s_{\rho_c^*}$ its standard deviation, which reflects the uncertainty about the prior guess, then set $\alpha_c = ((\rho_c^*(p-1)+1)(\rho_c^{*2}(1-p)+\rho_c^*(p-2)-s_{\rho_c^*}^2(p-1)+1))(p-1)^{-1}p^{-1}s_{\rho_c^*}^{-2}$ and $\zeta_c = ((\rho_c^*-1)(\rho_c^{*2}(p-1)-\rho_c^*(p-2)+s_{\rho_c^*}^2(p-1)-1)p^{-1}s_{\rho_c^*}^{-2}$.

reference prior in (11) in the interval $(0, 1)$. Throughout this paper we shall mainly focus on non-truncated priors but we also give some results for the truncated case.

4 Bayesian estimation under the marginal model

A Gibbs sampler is presented for fitting the generalized marginal model using the proposed class of priors in (9). First a parameter transformation is applied to generalized between-groups variances having shifted F priors, which are novel in the Bayesian literature. Subsequently a parameter expansion is applied that results in shifted inverse gamma priors, which are conjugate under the generalized marginal model. Finally a Gibbs sampler is presented.

First the prior in (9) is transformed to the parameters $(\boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2)$.

Lemma 1. *Transforming the prior in (9) from $(\boldsymbol{\beta}, \boldsymbol{\rho}, \phi^2)$ to $(\boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2)$ via (5) yields*

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2) &= \pi(\boldsymbol{\beta}|\boldsymbol{\eta}, \sigma^2)\pi(\sigma^2)\pi(\boldsymbol{\eta}|\sigma^2) && (12) \\ &= \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\beta}_0, g(\boldsymbol{X}'\boldsymbol{\Sigma}_N^{-1}\boldsymbol{X})^{-1}\right)\sigma^{-2}\prod_{c=1}^{C}\pi(\eta_c|\sigma^2), \text{ with} && (13) \\ \pi(\eta_c|\sigma^2) &= \textit{shifted-}\mathcal{F}(\eta_c; 2\alpha_c, 2\zeta_c, \tfrac{p-1}{p}\sigma^2, -\tfrac{\sigma^2}{p}), \end{aligned}$$

where the density of the shifted F distribution is given by

$$\textit{shifted-}\mathcal{F}(\tau^2; \nu_1, \nu_2, s^2, \mu) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)(s^2)^{\frac{\nu_1}{2}}}(\tau^2-\mu)^{\frac{\nu_1}{2}-1}\left(1+\tfrac{\tau^2-\mu}{s^2}\right)^{-\frac{\nu_1+\nu_2}{2}}, \quad (14)$$

where ν_1 is the first degrees of freedom, ν_2 is the second degrees of freedom, s^2 is a scale parameter, and μ is a shift parameter, and $\boldsymbol{\Sigma}_N = diag(\boldsymbol{I}_{n_1} \otimes \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{I}_{n_C} \otimes \boldsymbol{\Sigma}_C)$, with $\boldsymbol{\Sigma}_c = \sigma^2\boldsymbol{I}_p + \eta_c\boldsymbol{J}_p$, for $c = 1, \ldots, C$.

Proof. See Appendix A (Mulder and Fox, 2018). □

Second a parameter expansion is applied to model a shifted F distribution as a gamma mixture of shifted inverse gamma distributions.

Lemma 2. *The shifted F distribution in (14) can be obtained by setting a gamma mixture distribution on the scale parameter of a shifted inverse gamma distribution,*

$$\textit{shifted-}\mathcal{F}(\tau^2; \nu_1, \nu_2, s^2, \mu) = \int \textit{shifted-}\mathcal{IG}(\tau^2; \tfrac{\nu_2}{2}, \psi^2, \mu)\mathcal{G}(\psi^2; \tfrac{\nu_1}{2}, s^{-2})d\psi^2,$$

where the shifted inverse gamma distribution is given by

$$\textit{shifted-}\mathcal{IG}(\tau^2; \alpha, \xi, \mu) = \frac{\xi^\alpha}{\Gamma(\alpha)}(\tau^2-\mu)^{-\alpha-1}\exp\left\{-\tfrac{\xi}{\tau^2-\mu}\right\}, \quad (15)$$

where α is a shape parameter, ξ is a scale parameter, and μ is a shift parameter.

Proof. See Appendix B (Mulder and Fox, 2018). □

By applying Lemma 1 and 2, the joint prior for $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\eta}, \boldsymbol{\psi}^2)$ can be written as

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\eta}, \boldsymbol{\psi}^2) &= \sigma^{-2}\pi(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\eta}) \prod_{c=1}^{C} p(\eta_c|\psi_c^2, \sigma^2)p(\psi_c^2|\sigma^2), \text{ with} \qquad (16)\\
\pi(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\eta}) &= \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\beta}_0, g(\mathbf{X}'\boldsymbol{\Sigma}_N^{-1}\mathbf{X})^{-1}\right),\\
p(\eta_c|\psi_c^2, \sigma^2) &= \text{shifted-}\mathcal{IG}(\eta_c; \zeta_c, \psi_c^2, -\tfrac{\sigma^2}{p}),\\
p(\psi_c^2|\sigma^2) &= \mathcal{G}(\psi_c^2; \alpha_c, \tfrac{p}{p-1}\sigma^{-2}),
\end{aligned}$$

where $\boldsymbol{\psi}^2$ is a vector of length C of auxiliary parameters.

Subsequently by parameterizing the likelihood in (8) in terms of the generalized between-groups variances η_c and within-groups variance σ^2, it can be shown that the conditional posteriors of the parameters have known distributions from which we can sample in a Gibbs sampler (Appendix C; Mulder and Fox, 2018). This can be achieved by splitting the parameters in two blocks $\boldsymbol{\beta}$ and $(\sigma^2, \boldsymbol{\eta}, \boldsymbol{\psi}^2)$. By writing $\tilde{\mathbf{H}}_1 = \mathbf{H}_p'\mathbf{T}_1'\mathbf{T}_1\mathbf{H}_p$, with $\mathbf{T}_1 = (1, 0, \ldots, 0)$ of dimension $1 \times p$, $\tilde{\mathbf{H}}_2 = \mathbf{H}_p'\mathbf{T}_2'\mathbf{T}_2\mathbf{H}_p$, with $\mathbf{T}_2 = (\mathbf{0}\ \mathbf{I}_{p-1})$ of dimension $(p-1)\times p$, $\mathbf{X}_c = [\mathbf{X}_{c1}' \cdots \mathbf{X}_{cn_c}']'$ of dimension $n_c p \times k$, $\mathbf{X} = [\mathbf{X}_{11}'\mathbf{X}_{12}' \cdots \mathbf{X}_{Cn_C}']'$ of dimension $Np \times k$, and $\mathbf{y} = (\mathbf{y}_{11}', \mathbf{y}_{12}', \ldots, \mathbf{y}_{Cn_C}')'$ of length Np, the blocked Gibbs sampler can be written as follows

1. Set initial values for $(\boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2, \boldsymbol{\psi}^2)$, or for $(\boldsymbol{\beta}, \boldsymbol{\rho}, \phi^2, \boldsymbol{\psi}^2)$ and apply the transformation in (5).

2. Draw $\boldsymbol{\beta}$ given ($\boldsymbol{\eta}$, σ^2, $\boldsymbol{\psi}^2$) and $\mathbf{y}$ using

$$\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\eta}, \sigma^2, \boldsymbol{\psi}^2 \sim \mathcal{N}\left((g+1)^{-1}(g\hat{\boldsymbol{\beta}} + \boldsymbol{\beta}_0), \tfrac{g}{g+1}\left(\mathbf{X}'\boldsymbol{\Sigma}_N^{-1}\mathbf{X}\right)^{-1}\right),$$

   where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}_N^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_N^{-1}\mathbf{y}$, and $\boldsymbol{\Sigma}_N = \text{diag}(\mathbf{I}_{n_1} \otimes \boldsymbol{\Sigma}_1, \ldots, \mathbf{I}_{n_C} \otimes \boldsymbol{\Sigma}_C)$ of dimension $Np \times Np$, with compound symmetry covariance matrix $\boldsymbol{\Sigma}_c = \sigma^2\mathbf{I}_p + \eta_c\mathbf{J}_p$, for $c = 1, \ldots, C$.

3. Draw $(\boldsymbol{\eta}, \sigma^2, \boldsymbol{\psi}^2)$ given $\boldsymbol{\beta}$ and $\mathbf{Y}$.

   (a) Draw σ^2 given $\boldsymbol{\beta}$ and $\mathbf{y}$ using

$$\sigma^2|\boldsymbol{\beta}, \mathbf{y} \sim \mathcal{IG}\left(\tfrac{N(p-1)}{2}, \tfrac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{I}_N \otimes \tilde{\mathbf{H}}_2)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right).$$

   (b) Draw ψ_c^2 given σ^2, $\boldsymbol{\beta}$, and $\mathbf{y}$ using

$$\psi_c^2|\sigma^2, \boldsymbol{\beta}, \mathbf{y} \sim \mathcal{G}(\alpha_c, \tfrac{p}{p-1}\sigma^{-2}),$$

   for $c = 1, \ldots, C$.

(c) Draw η_c given ψ_c^2, σ^2, $\boldsymbol{\beta}$, and $\mathbf{Y}$ using

$$\eta_c|\psi_c^2,\sigma^2,\boldsymbol{\beta},\mathbf{y} \sim \text{shifted-}\mathcal{IG}(\tfrac{n_c}{2}+\zeta_c, \tfrac{1}{2p}(\mathbf{y}_c-\mathbf{X}_c\boldsymbol{\beta})'(\mathbf{I}_{n_c}\otimes\tilde{\mathbf{H}}_1)(\mathbf{y}_c-\mathbf{X}_c\boldsymbol{\beta})+\psi_c^2, -\tfrac{\sigma^2}{p}),$$

for $c=1,\ldots,C$.

(d) Compute the c-th intraclass correlation using $\rho_c=\frac{\eta_c}{\eta_c+\sigma^2}$, for $c=1,\ldots,C$.

4. Repeat Steps 2 and 3 until enough draws have been obtained, and exclude a burnin period.

If truncated stretched beta priors would be used for the intraclass correlations in the interval $(0,1)$, this would result in shifted F distributions for a generalized between-groups variances η_c truncated in $(0,\infty)$. Applying the same parameter expansion as above would result in a gamma mixture of truncated shifted inverse gamma priors in $(0,\infty)$. The corresponding conditional posteriors would then also have truncated shifted inverse gamma distributions. Sampling from these truncated shifted inverse gamma distributions can be done by sampling from the nontruncated shifted inverse gamma distribution until a positive value is drawn. This will be fairly efficient because the posterior probability mass in the negative region is generally quite small.

4.1 Frequentist coverage rates

Frequentist coverage rates are useful to investigate the performance of noninformative objective priors (e.g. Stein, 1985; Ghosh and Mukerjee, 1992; Berger et al., 2006). A simulation study was conducted to investigate the coverage rates of the lower 5% and 95% posterior quantiles for ρ_1 in the marginal model with $C=1$ using the reference prior (11), which should ideally be close to .05 and .95, respectively. This was done for population values of $\tau^2\in\{0,.1,.5,1,10\}$ and $\sigma^2=1$, which correspond to intraclass correlations of $\rho\in\{0,.09,.33,.5,.91\}$, and $\mu_1=0$, and for sample sizes of $(n,p)=(2,2),(10,5)$, and $(500,10)$. Note that the first sample size condition corresponds to a minimal balanced design with 2 groups with 2 observations per group. For each condition 50,000 data sets were generated. The coverage rates can be found in Table 1.

As can be seen from Table 1 the coverage rates under the marginal model with the considered reference prior are very accurate, even in the minimal information case with $(n,p)=(2,2)$ and an extreme intraclass correlation of $\rho=0$. These rates are better than previous results using a truncated reference prior under the multilevel model (2) (Berger and Bernardo, 1992; Ye, 1994; Chung and Dey, 1998, which are also presented in Table 1). This illustrates that the marginal model is superior over the multilevel model in terms of coverage rates of interval estimates for the variance components. Hence, the credible intervals can be used for significance type testing, even when testing $\rho=0$. Note that this would not be possible in a multilevel model because testing $\rho=0$ would be a boundary problem. Generally however we recommend using Bayes factors for testing intraclass correlations because significance tests, e.g., using interval estimates, tend to overestimate the evidence against a null hypothesis (Sellke et al., 2001; Pericchi, 2005). Bayes factor tests are proposed in the following section.

	(n, p)			
	marginal model			CD1998
ρ	$(2, 2)$	$(10, 5)$	$(500, 10)$	$(10, 5)$
0	.050(.951)	.050(.950)	.050(.950)	NA
.09	.051(.951)	.049(.950)	.050(.951)	.04(1.00)
.33	.052(.950)	.049(.949)	.050(.951)	.05(.99)
.5	.049(.950)	.051(.951)	.050(.949)	.04(.98)
.91	.048(.950)	.050(.950)	.051(.951)	.04(.94)

Table 1: Frequentist coverage probabilities of lower posterior quantiles of 5%(95%) for ρ for the marginal model (7). The results in the last column were taken from Chung and Dey (1998).

5 Bayes factor testing under the marginal model

When testing statistical hypotheses using the Bayes factor, prior specification plays a more important role than in Bayesian estimation. Instead of having to formulate one prior, which may be improper in Bayesian estimation, proper priors need to be specified for all unique intraclass correlations under all Q equality and order constrained hypotheses in (1). Furthermore, unlike Bayesian estimation, the effect of the priors on the Bayes factor does not fade away as the sample size grows (Jeffreys, 1961; Berger and Pericchi, 2001; Bayarri et al., 2012). Ad hoc or arbitrary prior specification should therefore be avoided. Also note that (objective) improper priors cannot be used in Bayesian hypothesis testing because the resulting Bayes factors would depend on undefined constants (e.g. O'Hagan, 1995; Berger and Pericchi, 1996). These facts have severely complicated the development of (objective) priors in Bayesian hypothesis testing and model selection.

In this section we propose a Bayes factor testing procedure that can be used when prior information about the magnitude of the intraclass correlations under the hypotheses is available or when prior information is too limited for adequate prior specification. When prior information is available this can be translated to proper stretched beta priors for intraclass correlations in (9), similar as in the estimation problem. When prior information is absent or when a default Bayesian method is preferred a generalized fractional Bayesian procedure is proposed. These default Bayes factors are based on the improper versions of stretched beta priors. Note that more examples can be found in the literature where the same family of prior distributions is used for estimation as for hypothesis testing or model selection. For example Cauchy priors with thick tails are useful for estimation in robust Bayesian analyses (Berger, 1994) and in Bayesian regularization problems (Griffin and Brown, 2005), and Cauchy priors are also useful for Bayes factor testing to avoid the information paradox (Zellner and Siow, 1980; Liang et al., 2008). Furthermore, the (matrix) F prior is useful when estimating variance components (Gelman, 2006; Pérez et al., 2017) and for testing variances (Mulder and Pericchi, 2018).

5.1 Prior specification and marginal likelihoods

Under a constrained hypothesis $H_q : \mathbf{R}_q^E\boldsymbol{\rho} = \mathbf{0}, \mathbf{R}_q^I\boldsymbol{\rho} > \mathbf{0}$, let the free intraclass correlations be denoted by the vector $\tilde{\boldsymbol{\rho}}$ of length V (the hypothesis index is omitted to simplify the notation). The inequality constraints on the free intraclass correlations can then be written as $\tilde{\mathbf{R}}_q\tilde{\boldsymbol{\rho}} > \mathbf{0}$. For example, when the first two intraclass correlations are assumed to be equal and larger than the third intraclass correlation, i.e., $H_1 : \rho_1 = \rho_2 > \rho_3$, then $\tilde{\rho}_1 = \rho_1 = \rho_2$ and $\tilde{\rho}_2 = \rho_3$, and $\tilde{\mathbf{R}}_1 = [1 \ -1]$.

If prior information is available under H_q, this can be translated to informative truncated stretched beta priors on the free intraclass correlations,

$$\pi_q(\tilde{\boldsymbol{\rho}}|\boldsymbol{\alpha},\boldsymbol{\zeta}) = I(\tilde{\mathbf{R}}_q\tilde{\boldsymbol{\rho}} > \mathbf{0})P(\tilde{\mathbf{R}}_q\tilde{\boldsymbol{\rho}} > \mathbf{0}|H_q^*)^{-1}\prod_{v=1}^{V} beta(\tilde{\rho}_v|\alpha_v,\zeta_v,-\tfrac{1}{p-1},1), \tag{17}$$

where H_q^* corresponds to hypothesis H_q with the inequality constraints omitted, i.e., $H_q^* : \mathbf{R}_q^E\boldsymbol{\rho} = \mathbf{0}$ (see also Pericchi et al., 2008), and the prior probability that the inequality constraints hold under H_q^*, which serves as a normalizing constant, is given by

$$P(\tilde{\mathbf{R}}_q\tilde{\boldsymbol{\rho}} > \mathbf{0}|H_q^*) = \int_{\tilde{\mathbf{R}}_q\tilde{\boldsymbol{\rho}}>\mathbf{0}}\prod_{v=1}^{V} beta(\tilde{\rho}_v|\alpha_v,\zeta_v,-\tfrac{1}{p-1},1)d\tilde{\boldsymbol{\rho}}.$$

Subsequently, priors need to be specified for the nuisance parameters $\boldsymbol{\beta}$ and ϕ^2 under all hypotheses. First note that the Bayes factor is known to be robust to the choice of the same prior for the common orthogonal nuisance parameters (in the sense of a block-diagonal expected Fisher information matrix; see Jeffreys, 1961; Kass and Vaidyanathan, 1992; Ly et al., 2016). This justifies the use of the same improper prior for the nuisance parameters. First note that the fixed effects $\boldsymbol{\beta}$ are orthogonal to $\boldsymbol{\rho}$, and therefore we can use the improper prior $\pi_q^N(\boldsymbol{\beta}) = 1$. Second, ϕ^2 is not orthogonal to $\boldsymbol{\rho}$[3]. When setting a vague inverse-gamma prior for ϕ^2 however, i.e., $\pi(\phi^2) = \mathcal{IG}(\epsilon,\epsilon)$, with $\epsilon > 0$ small, it can be shown that the resulting Bayes factor will be virtually independent of the exact choice of ϵ as long as ϵ is small enough. Due to this robustness property, we can specify the improper prior $\pi_q^N(\phi^2) = \phi^{-2} = \mathcal{IG}(0,0)$. Hence, the joint prior under H_q is given by

$$\pi_q(\boldsymbol{\beta},\tilde{\boldsymbol{\rho}},\phi^2|\boldsymbol{\alpha},\boldsymbol{\zeta}) = \phi^{-2}\pi_q(\tilde{\boldsymbol{\rho}}|\boldsymbol{\alpha},\boldsymbol{\zeta}). \tag{18}$$

Under each hypothesis H_q, the hyperparameters $\boldsymbol{\alpha},\boldsymbol{\zeta} > \mathbf{0}$ can be specified in a similar manner as was discussed in Section 3. Proper uniform priors also fall in this category which can be specified by setting $\boldsymbol{\alpha} = \boldsymbol{\zeta} = \mathbf{1}$. A uniform prior for the unique intraclass correlations under hypothesis H_q implies that all possible values for the intraclass correlations that satisfy the constraints of H_q are equally likely a priori. Once the priors have been specified, the marginal likelihood of the transformed data $\mathbf{z}$ under hypothesis H_q can be computed according to

$$m_q(\mathbf{z}) = \iiint_{H_q} f_q(\mathbf{z}|\mathbf{W},\boldsymbol{\beta},\tilde{\boldsymbol{\rho}},\phi^2)\pi_q(\boldsymbol{\beta},\tilde{\boldsymbol{\rho}},\phi^2|\boldsymbol{\alpha},\boldsymbol{\zeta})d\boldsymbol{\beta}d\phi^2 d\tilde{\boldsymbol{\rho}}, \tag{19}$$

[3]Note that the prior of García-Donato and Sun (2007) is only asymptotically orthogonal as $p \to \infty$.

where f_q is the likelihood under H_q which is a truncation of the unconstrained likelihood f in (8) in the subspace under H_q. For the above example with $H_1 : \rho_1 = \rho_2 > \rho_3$, the likelihood would be equal to

$$f_1(\mathbf{z}|\mathbf{W}, \boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \psi^2) = f(\mathbf{z}|\mathbf{W}, \boldsymbol{\beta}, (\tilde{\rho}_1, \tilde{\rho}_1, \tilde{\rho}_2)', \phi^2) \times I(\tilde{\rho}_1 > \tilde{\rho}_2).$$

The computation of the marginal likelihood (19) is discussed in the following section.

Formulating informative priors for the intraclass correlations under all hypotheses can be a challenging and time-consuming endeavor (Berger, 2006). To avoid this step, a default Bayesian procedure is proposed. First truncated reference priors will be specified having truncated stretched beta distributions with hyperparameters of zero, i.e.,

$$\pi_q^N(\tilde{\boldsymbol{\rho}}|\boldsymbol{\alpha} = \mathbf{0}, \boldsymbol{\zeta} = \mathbf{0}) = I(\tilde{\mathbf{R}}_q\tilde{\boldsymbol{\rho}} > \mathbf{0}) \prod_{v=1}^{V} (1 - \tilde{\rho}_v)^{-1}(1 + (p-1)\tilde{\rho})^{-1}. \tag{20}$$

To avoid the dependence of the marginal likelihood on the undefined constants in these improper priors, a generalized fractional Bayes procedure is considered using different fractions for different transformed observations. The motivation for using different fractions is that only the first element of the transformed observations $\mathbf{z}_{ci}$ in (7) contains information about ρ_c, and therefore the amount of information in the default prior for the different parameters cannot be properly controlled using one common fraction for all observations, as in the standard fractional Bayes factor (O'Hagan, 1995). Generalized fractional Bayes approaches for normal linear models were for instance considered by Berger and Pericchi (2001), De Santis and Spezzaferri (2001), Mulder (2014), and Hoijtink et al. (2018). The likelihood functions of the different group categories in (8) are raised to different fractions according to (with a slight abuse of notation)

$$f(\mathbf{z}|\mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\rho}, \phi^2)^{\mathbf{b}} \equiv \prod_{c=1}^{C} \prod_{i=1}^{n_c} f(z_{ci1}|\mathbf{W}_{ci}, \boldsymbol{\beta}, \boldsymbol{\rho}, \phi^2)^{b_c} \prod_{j=2}^{p} f(z_{cij}|\mathbf{W}_{ci}, \boldsymbol{\beta}, \boldsymbol{\rho}, \phi^2)^{b_0}, \tag{21}$$

where b_c is the fraction of the data of the c-th category used to identify the parameters that are specific to category c (such as ρ_c, and possibly a category specific intercept), and b_0 is the fraction of the data used to identify the remaining parameters. Generally the use of small fractions is recommended (O'Hagan, 1995; Berger and Mortera, 1995). The choice of the fractions will be motivated in Section 5.3.

Subsequently the marginal likelihood under H_q using the generalized fractional Bayes approach is defined by

$$m_q(\mathbf{y}, \mathbf{b}) = \frac{m_q^N(\mathbf{y})}{m_q^N(\mathbf{y}^{\mathbf{b}})}, \tag{22}$$

where

$$m_q^N(\mathbf{y}^{\mathbf{b}}) = \iiint_{H_q} f_q(\mathbf{y}|\mathbf{W}, \boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2)^{\mathbf{b}} \pi_q^N(\boldsymbol{\beta}, \phi^2, \tilde{\boldsymbol{\rho}}) d\boldsymbol{\beta} d\phi^2 d\tilde{\boldsymbol{\rho}}, \tag{23}$$

symbolizes the marginal likelihood of a fraction $\mathbf{b}$ of the information in the complete dataset $\mathbf{y}$, i.e., $\mathbf{y}^{\mathbf{b}}$, using the truncated noninformative improper prior (20). Note that the numerator in (22) can be obtained by setting $\mathbf{b} = \mathbf{1}$ in (23). Because the same noninformative improper prior is used for computing both marginal likelihoods in (22), the undefined constant in this improper prior cancels out (O'Hagan, 1995).

5.2 Computation of the marginal likelihood

In the following lemma a general result is given for the marginal likelihood for a constrained hypothesis H_q when using proper truncated stretched beta priors for the unique intraclass correlations or when adopting a generalized fractional Bayes approach.

Lemma 3. *Under a constrained hypothesis* $H_q : \boldsymbol{R}_q^E\boldsymbol{\rho} = \boldsymbol{0}, \boldsymbol{R}_q^I\boldsymbol{\rho} > \boldsymbol{0}$*, the marginal likelihood in the informative case with* $\boldsymbol{\alpha}, \boldsymbol{\zeta} > \boldsymbol{0}$ *and in the noninformative case with* $\boldsymbol{\alpha} = \boldsymbol{\zeta} = \boldsymbol{0}$ *are given by*

$$\begin{aligned} m_q(\boldsymbol{y}) &= \pi^{\frac{K}{2}-\frac{1}{2}\sum_{c=1}^{C} n_c+n_c(p-1)}\Gamma\left(-\tfrac{K}{2}+\tfrac{1}{2}\sum_{c=1}^{C} n_c + n_c(p-1)\right) \\ &\quad \int_{H_q^*} h(\tilde{\boldsymbol{\rho}}, \boldsymbol{1}, \boldsymbol{\alpha}, \boldsymbol{\zeta}) d\tilde{\boldsymbol{\rho}}\ \frac{Pr(\tilde{\boldsymbol{R}}_q\tilde{\boldsymbol{\rho}} > \boldsymbol{0}|H_q^*, \boldsymbol{y})}{Pr(\tilde{\boldsymbol{R}}_q\tilde{\boldsymbol{\rho}} > \boldsymbol{0}|H_q^*)}, \qquad (24) \\ m_q^N(\boldsymbol{y}^{\boldsymbol{b}}) &= \pi^{\frac{K}{2}-\frac{1}{2}\sum_{c=1}^{C} n_c b_c+n_c(p-1)b_0}\Gamma\left(-\tfrac{K}{2}+\tfrac{1}{2}\sum_{c=1}^{C} n_c b_c + n_c(p-1)b_0\right) \\ &\quad \int_{H_q^*} h(\tilde{\boldsymbol{\rho}}, \boldsymbol{b}, \boldsymbol{0}, \boldsymbol{0}) d\tilde{\boldsymbol{\rho}}\ Pr(\tilde{\boldsymbol{R}}_q\tilde{\boldsymbol{\rho}} > \boldsymbol{0}|H_q^*, \boldsymbol{y}^{\boldsymbol{b}}), \qquad (25) \end{aligned}$$

where $h(\tilde{\boldsymbol{\rho}}, \boldsymbol{b}, \boldsymbol{\alpha}, \boldsymbol{\zeta})$ *is an analytic function of the unique intraclass correlations under* H_t*, the fractions* $\boldsymbol{b}$*, and the prior hyperparameters* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\zeta}$*.*

Proof. Appendix D (Mulder and Fox, 2018). □

Note that the first part of the marginal likelihood in (24) is equivalent to the marginal likelihood of H_q^* without the inequality constraints, while the second ratio of probabilities quantifies the support for the inequality constraints in the data within hypothesis H_q^* (see also, Pericchi et al., 2008; Consonni and Paroli, 2017; Gu et al., 2017).

In (24) and (25), the posterior probabilities can be computed as the proportion of draws satisfying the inequality constraints under H_q^*. The Gibbs sampler for obtaining draws under H_q^* given $\mathbf{y}^{\mathbf{b}}$ can be found in Appendix E (Mulder and Fox, 2018). The integrals in (24) and (25) can be computed using the following importance sample estimate

$$\int_{H_q^*} h(\tilde{\boldsymbol{\rho}}, \mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\zeta}) d\tilde{\boldsymbol{\rho}} = \int_{H_q^*} \frac{h(\tilde{\boldsymbol{\rho}}, \mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\zeta})}{q(\tilde{\boldsymbol{\rho}})} q(\tilde{\boldsymbol{\rho}}) d\tilde{\boldsymbol{\rho}} \approx S^{-1}\sum_{s=1}^{S} \frac{h(\tilde{\boldsymbol{\rho}}^{(s)}, \mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\zeta})}{q(\tilde{\boldsymbol{\rho}}^{(s)})},$$

for $t = 1$ or 2, where $q(\tilde{\boldsymbol{\rho}})$ is a proposal density under H_q^*, and $\tilde{\boldsymbol{\rho}}^{(s)}$ is the s-th draw from $q(\tilde{\boldsymbol{\rho}})$. The proposal density is a product of stretched beta distributions, $beta(\alpha_v^*, \zeta_v^*, -\frac{1}{p-1}, 1)$, for $v = 1, \ldots, V$, which is tailored to $h(\tilde{\boldsymbol{\rho}}, \mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\zeta})$. First a posterior sample is drawn for $\tilde{\boldsymbol{\rho}}$ under H_q^* (Appendix E in Mulder and Fox, 2018). Then the shape parameters of the proposal distribution are computed with a method of moments estimator using the estimated posterior mean and variance as in footnote 2. By

multiplying the shape parameters of the proposal density with, say, .7, the proposal density gets heavier tails than the kernel of the posterior h which ensures a stable and consistent estimate of the integral.

In the special case where $\mathbf{X}_{ci}$ is a $p \times c$ matrix with ones in column c and zeros elsewhere, which implies that fixed intercepts per group category are the only covariates (as in a standard random intercept model), the marginal likelihood based on the truncated reference prior (20) has an analytic form. The expression can be found in Appendix F (Mulder and Fox, 2018). Consequently the generalized fractional Bayes factor has an analytic solution when testing equality and/or order constraints on multiple intraclass correlations in the random intercept model.

5.3 Choice of the fractions

In the fractional Bayes factor a fraction of the data is used to implicitly construct a default prior that is concentrated around the likelihood (e.g. Gilks, 1995). This is also the case for the generalized fractional Bayes factor as can be seen below

$$\begin{aligned} m_q(\mathbf{y}, \mathbf{b}) &= \frac{\iiint f_q(\mathbf{y}|\boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2)\pi_q^N(\boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2)d\boldsymbol{\beta}d\tilde{\boldsymbol{\rho}}d\phi^2}{\iiint f_q(\mathbf{y}|\boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2)^{\mathbf{b}}\pi_q^N(\boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2)d\boldsymbol{\beta}d\tilde{\boldsymbol{\rho}}d\phi^2} \\ &= \iiint f_q(\mathbf{y}|\boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2)^{\mathbf{1}-\mathbf{b}}\pi_q(\boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2|\mathbf{y}^{\mathbf{b}})d\boldsymbol{\beta}d\tilde{\boldsymbol{\rho}}d\phi^2, \end{aligned}$$

where the proper updated prior is defined by

$$\pi_q(\boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2|\mathbf{y}^{\mathbf{b}}) = \frac{f_q(\mathbf{y}|\boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2)^{\mathbf{b}}\pi_q^N(\boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2)}{\iiint f_q(\mathbf{y}|\boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2)^{\mathbf{b}}\pi_q^N(\boldsymbol{\beta}, \tilde{\boldsymbol{\rho}}, \phi^2)d\boldsymbol{\beta}d\tilde{\boldsymbol{\rho}}d\phi^2}. \tag{26}$$

In the original papers of the fractional Bayes factor, it was argued that the choice of the fraction should depend on the uncertainty about the employed improper prior: In the case of much (little) uncertainty, a relatively large (small) fraction should be used to update the improper prior (O'Hagan, 1995, 1997; Conigliani and O'Hagan, 2000). Because the improper prior seems a reasonable objective choice (Section 4.1) and because larger fractions for prior specification would result in less information for hypothesis testing, we focus on minimal fractions in this paper (see also Berger and Mortera, 1995). A minimal fraction is based on the minimal amount of observations that are needed to obtain a proper updated prior. In practice each group category often has its own fixed intercept, which implies that $\mathbf{X}_{ci}$ contains a column with only ones. After the Helmert transformation in (7), this column becomes $(\sqrt{p}, 0, \ldots, 0)'$ in $\mathbf{W}_{ci} = \mathbf{H}_p\mathbf{X}_{ci}$. Thus, only the intercept and intraclass correlation of each group category are identified by the first transformed observations, z_{ci1}, for $c = 1, \ldots, C$ and $i = 1, \ldots, n_c$. This implies that two observations are needed of the first transformed observations in each group, which corresponds to a minimal fraction of $b_c = \frac{2}{n_c}$. The remaining $K - C$ fixed effects (where the groups specific intercept are excluded) and the total variance parameter ϕ^2 are then identified by the $N(p-1)$ transformed observations, z_{cij}, for $c = 1, \ldots, C$, $i = 1, \ldots, n_c$, and $j = 2, \ldots, p$, which implies a minimal fraction of $b_0 = \frac{K-C+1}{N(p-1)}$.

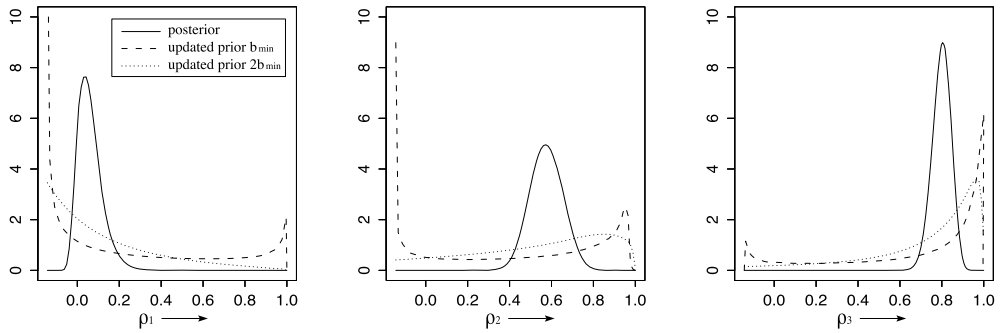

Figure 3: Estimated posterior densities and default prior densities based on minimal fractions, $b_c = \frac{2}{n_c}$ and $b_0 = \frac{K-C+1}{N(p-1)}$, and twice the minimal fractions, $b_c = \frac{4}{n_c}$ and $b_0 = \frac{2K-2C+2}{N(p-1)}$, for randomly generated data with $C = 3$ groups, $K = 3$ fixed intercepts that are group type specific, intraclass correlations of size $\boldsymbol{\rho} = (.1, .6, .8)$, groups of size $p = 8$, $\phi^2 = 1$, and $\mathbf{n} = (20, 25, 30)$.

To get an idea about the effect of the choice of the fractions on the proper default prior, Figure 3 displays the estimated marginal posterior densities (solid line) of the intraclass correlations (ρ_1, ρ_2, ρ_3) (left, middle, and right panel, respectively) and the estimated marginal updated prior densities based on minimal fractions (dashed line) and twice the minimal fractions (dotted line), all based on the noninformative improper prior. These densities were estimated from a randomly generated data set with $\boldsymbol{\rho} = (.1, .6, .8)$, $\mathbf{n} = (20, 25, 30)$, $p = 8$, and group type specific intercepts $\boldsymbol{\beta} = (0, 0, 0)'$. As can be seen the proper updated prior based on minimal fractions are very similar to the noninformative reference priors. The updated priors based on twice the minimal fractions are more concentrated around the likelihood. In the remaining part of the paper we use minimal fractions so that most information in the data is used for hypothesis testing.

6 Numerical performance

A multiple hypothesis test is considered on $C = 2$ group specific intraclass correlations. The following hypotheses are being tested:

$$\begin{aligned} H_1 &: \rho_1 = 0, \rho_2 > 0, \\ H_2 &: \rho_1 > 0, \rho_2 = 0, \\ H_3 &: \rho_1 = \rho_2, \\ H_4 &: \rho_1 > \rho_2, \\ H_5 &: \rho_1 < \rho_2. \end{aligned} \tag{27}$$

Our interest is in the default relative evidence based on the generalized fractional Bayes factor while varying the (unconstrained) posterior modes of the intraclass correlations

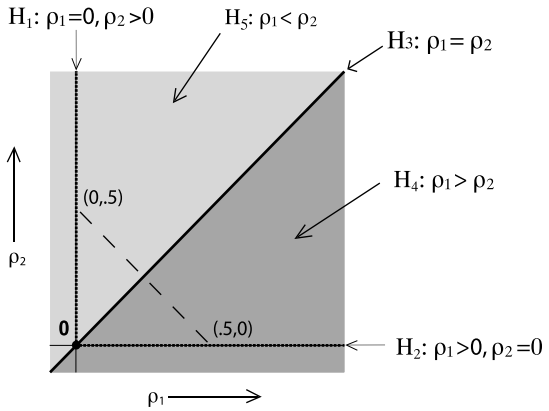

Figure 4: Graphical representation of the subspaces of the intraclass correlations $(\rho_1, \rho_2) \in (-\frac{1}{9}, 1) \times (-\frac{1}{9}, 1)$ under five different hypotheses in (27) and the trajectory of the estimated intraclass correlations $(\bar{\rho}_1, \bar{\rho}_2) = (0, 0.5)$, to $(0.5, 0)$ (dashed line).

for different group sizes (n_1, n_2). As fixed effects only group specific intercepts were included. Therefore the marginal likelihoods can be computed by simply plugging in the group specific between groups sums of squares, $s^2_{B,c}$ for $c = 1$ and 2, and the within groups sums of squares s^2_W in (8) in Appendix F of Mulder and Fox (2018). The sums of squares were varied according to $s^2_W = (p-1)N\bar{\sigma}^2$ and $s^2_{B,c} = n_c(\bar{\tau}^2_c + \bar{\sigma}^2/p)$, for $c = 1, 2$, where $\bar{\sigma}^2$ and $\bar{\boldsymbol{\tau}}^2$ are the unconstrained posterior modes, which were varied over $\bar{\boldsymbol{\tau}}^2 = (\bar{\tau}^2_1, 1-\bar{\tau}^2_1)'$, for $\bar{\tau}^2_1 = 0, \ldots, 1$ and $\bar{\sigma}^2 = 1$, so that $(\bar{\rho}_1, \bar{\rho}_2) = (0, .5), \ldots, (.5, 0)$. Thus, when $\bar{\rho}_1 \approx (0, .5)$, $(.5, 0)$, or $(.25, .25)$, it is expected to receive most evidence for H_1, H_2, or H_3, respectively, and between these regions it is expected to either receive most evidence for H_4 or H_5. The subspaces under the hypotheses and the trajectory of unconstrained estimated intraclass correlations are displayed in Figure 4. The group size was set to $p = 10$, and the number of groups in each category was set to $n_1 = n_2 =$ 30, 300, and 3000.

Figure 5 (left columns) displays the logarithms of generalized fractional Bayes factors of hypothesis H_1 (dashed line), H_2 (dash-dotted line), H_3 (thick solid line), H_4 (dotted line), and H_5 (thin solid line) versus an unconstrained (reference) hypothesis H_u : $(\rho_1, \rho_2) \in (-\frac{1}{9}, 1) \times (-\frac{1}{9}, 1)$ as a function of the unconstrained estimates of the intraclass correlations $(\bar{\rho}_1, \bar{\rho}_2)$ for $n_1 = n_2 = 30$ (upper panels), $n_1 = n_2 = 300$ (middle panels), and $n_1 = n_2 = 3{,}000$ (lower panels). Figure 5 (right columns) displays the corresponding posterior probabilities of the hypotheses based on equal prior probabilities, which can be computed as $P(H_q|\mathbf{y}) = \frac{B_{qu}}{\sum_{q'=1}^{5} B_{q'u}}$, with $B_{q'u} = m_q(\mathbf{y}, \mathbf{b}_{\min})/m_u(\mathbf{y}, \mathbf{b}_{\min})$. The plots show desirable default behavior of the generalized fractional Bayes factors as a function of the effects and sample size: The evidence is largest for the hypothesis that is also most supported by the data and the posterior probability for the true hypothesis goes to 1 as the number of groups increases, which implies consistency. Also note that the evidence for a true precise hypothesis with equality constraints (i.e., H_1, H_2, and H_3)

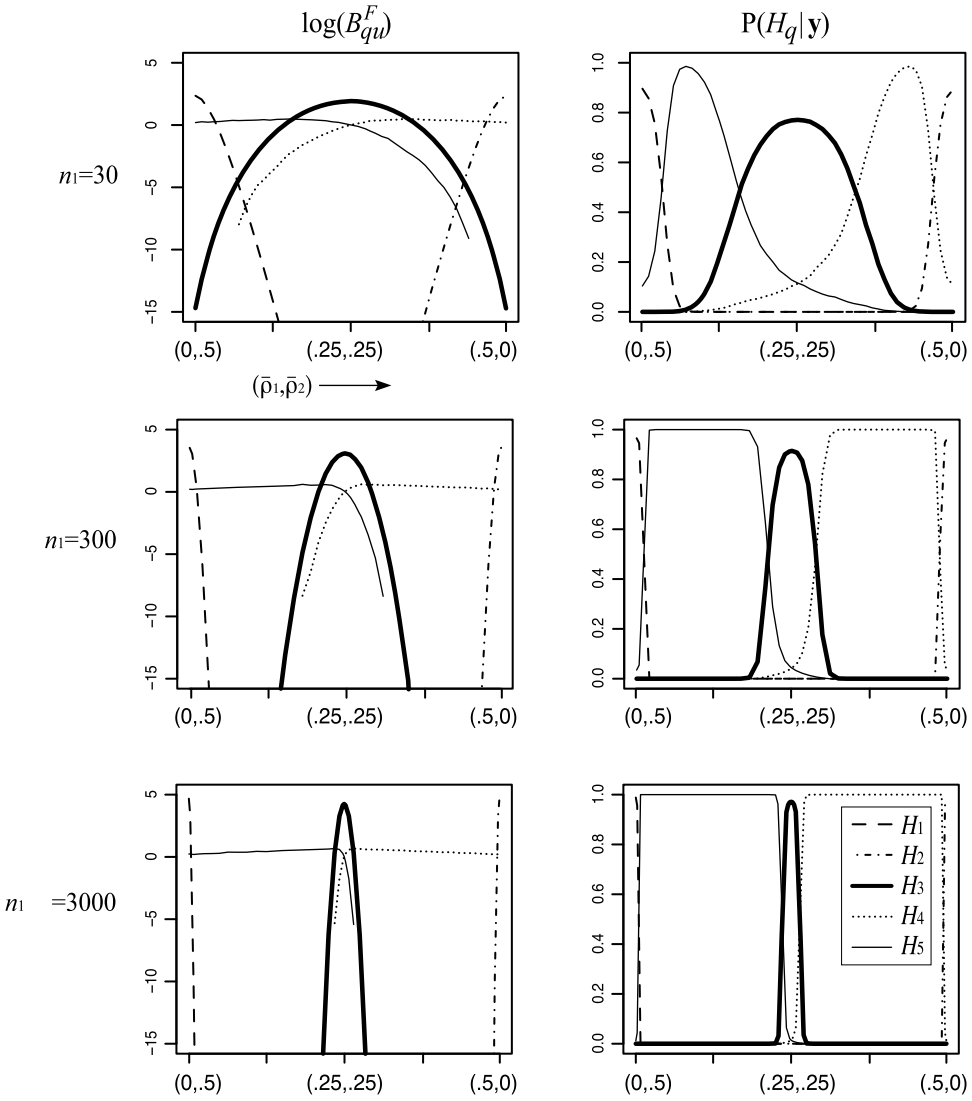

Figure 5: Logarithms of generalized fractional Bayes factors (left column) of hypothesis H_1 (dashed line), H_2 (dash-dotted line), H_3 (thick solid line), H_4 (dotted line), and H_5 (thin solid line) against an unconstrained hypothesis, and corresponding posterior probabilities of the hypotheses (right column) as a function of estimated intraclass correlations $(\bar{\rho}_1, \bar{\rho}_2)$ which varied from $(0, .5)$ to $(.5, 0)$ for $n_1 = n_2 = 30$ (upper panels), $n_1 = n_2 = 300$ (middle panels), and $n_1 = n_2 = 3{,}000$ (lower panels).

accumulates with a slower rate than for the other hypotheses. This is commonly observed behavior of Bayes factor methodology (e.g., Johnson and Rossell, 2010). The evidence

would increase with a faster rate when testing interval hypotheses instead of precise hypotheses (see Appendix G in Mulder and Fox, 2018). Finally note that the lines for H_4 and H_5 in Figure 5 are incomplete because, in the case of misfit of the inequality constraints, the proportion of 10,000 posterior draws that satisfy the constraints that is used for estimating the posterior probabilities is equal to zero.

7 Testing intraclass correlations in TIMSS

The Trends in International Mathematics and Science Study (TIMSS) measures the performances of fourth and eight graders in more than 50 participating countries around the world (http://www.iea.nl/timss). TIMSS is conducted regularly on a 4-year cycle, where mathematics and science has been assessed in 1995, 1999, 2003, 2007, 2011, and 2015. The fourth grade is a reference to a year in elementary eduction, where in North America the fourth grade is the fifth school year and in The Netherlands it is called group 6. The children are usually around 9–10 years old. The assessment data of each cycle can be found in the TIMSS's International Database.

When considering the international mathematics achievement of 2015 at the fourth grade, 21 countries improved their average performance, 15 countries had the same average achievement, and 5 countries had a lower average achievement compared to the mathematics achievement of 2011. The average 4th-grade mathematics scores in 2015 were lower for Germany and the Netherlands, scoring 6 and 10 points lower on average, respectively. To provide a reference point, the TIMSS achievement scale is centered at 500 and the standard deviation is equal to 100 scale score points. The TIMSS data set has a three-level structure, where students are nested within classrooms/schools, and the classrooms/schools are nested within countries. Only one classroom is sampled per school, so it is not possible to model variability among classrooms within schools.

For the TIMSS 2011 and 2015 assessment, the changes in the mathematics achievement were investigated by examining the grouping of students in schools across countries. The object was to evaluate whether a specific selection of schools (i.e., particular subpopulation) performed less in 2015, or whether the drop in performance applied to the entire population of schools of the considered country. Therefore, changes in the country-specific intraclass correlation coefficient from 2011 to 2015, representing the heterogeneity in mathematic achievements within and between schools across years, were tested. When detecting a decrease in average performance together with an increase of the intraclass correlation, a subset of schools performed worse. For a constant intraclass correlation across years the drop in performance applied to the entire population of schools. For different countries, changes in the intraclass correlation across years were tested concurrently to examine also differences across countries.

From a sampling perspective, the size of the intraclass correlation is also of specific interest, since sampling becomes less efficient when the intraclass correlation increases. Countries with low intraclass correlations have fewer restrictions on the sample design, where countries with high intraclass correlations require more efficient sample designs, larger sample sizes, or both. Knowledge about the size of the heterogeneity provide useful

information to optimize the development of a suitable sample design and to minimize the effects of high intraclass correlations.

Four countries were considered, The Netherlands (NL), Croatia (HR), Germany (DE), and Denmark (DK), where Croatia improved their average achievement and Denmark had the same average achievement. The achievement scores of overall mathematics were considered and the first plausible value was used as a measure of the mathematic achievements of the population (Olson et al., 2008). A stratified sample was drawn by country and school to obtain a balanced sample of $p = 15$ grade-4 students per school for each of the four countries and two measurement occasions.

The final sample consisted of C=8 group categories, by crossing the four countries with the two measurement occasions, which are referred to as group category $c = 1$ $(NL, 11)$, $c = 2$ $(NL, 15)$,..., $c = 8$ $(DK, 15)$. The data was retrieved from schools from The Netherlands ($n_{NL,11} = 93$, $n_{NL,15} = 112$), Croatia ($n_{HR,11} = 139$,$n_{HR,15} = 106$), Germany ($n_{DE,11} = 179$, $n_{DE,15} = 170$), and Denmark ($n_{DK,11} = 166$,$n_{DK,15} = 153$) with the sampled number of n schools in brackets for 2011 and 2015, respectively. Although often unconditional intraclass correlations are the object of study to explore variations (Hedges and Hedberg, 2007), differences in intraclass correlations were tested conditional on several student variables (e.g., gender, student sampling weight variable). The marginal model represented in (6) was fitted to obtain the parameter estimates, where 10,000 iterations were made and a burnin period of 1,000 iterations was used.

The following hypotheses were considered in the analyses. Hypothesis H_1 represents a common positive (invariant) intraclass correlation across countries and years. Positive country-specific and time-invariant intraclass correlations are represented by hypothesis H_2. Variation in intraclass correlation across years (i.e., a time-variant intraclass correlation) is represented by Hypothesis H_3, while assuming a common (invariant) positive intraclass correlation across countries per year. Finally, hypothesis H_4 represents the complement of H_1, H_2, and H_3 with unique (variant) intraclass correlations across countries and years.

Next to the assumed heterogeneity in country-specific intraclass correlations of H_2, an ordering in the correlations can also be hypothesized. The variance of the mean from a balanced clustered sample each of size p is larger than the variance of the mean of a simple random sample by a factor $1 + (p - 1)\rho$ (Kish, 1965, p. 162–163), which is known as the design effect. So, the intraclass correlation modifies the variance of the mean, given the number of schools and students per school. In the Netherlands, the variance of the average mathematic achievements of fourth graders is known to be relatively low. This can be inferred from the reported standard errors of the Netherland's average mathematics achievement during the cycles from 2003 to 2015, which were usually one of the lowest and ranged from 1.7 to 2.1. The standard errors for Denmark were much higher and ranged from 2.4 to 2.7. For Germany they ranged from 2.0 to 2.3. For the cycles in 2011 and 2015, Croatia had a standard error of 1.8 to 1.9, where the Netherlands had a standard error of 1.7 (Mullis et al., 2011, Exhibit 1.5) (http://timssandpirls.bc.edu/timss2015/international-results/timss-2015/mathematics/student-achievement). It can be expected that the variation in scores

across schools was higher for countries with higher reported standard errors of the average mathematics achievement. This implies an ordering of the country-specific intraclass correlations (from high to low) of Denmark, Germany, Croatia, and The Netherlands. Furthermore, the reported country-specific mathematics achievement distribution also revealed this ordering in the spread of student scores across countries.

To summarize, the following hypotheses were tested to examine differences in intraclass correlations:

H_1 : $0 < \rho_{NL,11} = \rho_{NL,15} = \rho_{HR,11} = \rho_{HR,15} = \rho_{DE,11} = \rho_{DE,15} = \rho_{DK,11} = \rho_{DK,15}$
(invariant positive intraclass correlations)

H_2 : $0 < \rho_{NL,11} = \rho_{NL,15} < \rho_{HR,11} = \rho_{HR,15} < \rho_{DE,11} = \rho_{DE,15} < \rho_{DK,11} = \rho_{DK,15}$
(time-invariant, country-ordered and -variant positive intraclass correlations)

H_3 : $0 < \rho_{NL,11} = \rho_{DE,11} = \rho_{HR,11} = \rho_{DK,11}$,
$0 < \rho_{NL,15} = \rho_{DE,15} = \rho_{HR,15} = \rho_{DK,15}$
(time-variant, country-invariant positive intraclass correlations)

H_4 : not H_1, H_2, H_3
(time- and country-variant intraclass correlations).

The unconstrained posterior distributions of the intraclass correlations for each country and occasion are given in Figure 6. It can be seen that the posterior distribution of the intraclass correlations show an ordering, where the Netherlands show the lowest level of clustering and Denmark the highest level of clustering. This ordering in intraclass correlations appears to be similar in 2011 and 2015. For the Netherlands, the posterior means of the intraclass correlations are around $\rho_{NL,11} = .089$ and $\rho_{NL,15} = .082$, for Croatia, $\rho_{HR,11} = .118$ and $\rho_{HR,15} = .117$, for Germany, $\rho_{DE,11} = .153$ and $\rho_{DE,15} = .150$, and for Denmark $\rho_{DK,11} = .189$ and $\rho_{DK,15} = .222$, for 2011 and 2015, respectively, where the estimated posterior standard deviation is around .02. From Figure 6 and the estimated intraclass correlations it follow that the change in heterogeneity across years is rather small, where only Denmark's estimated posterior distribution of the intraclass correlation for 2011 and 2015 show some differences. In the Netherlands, in 2011 and 2015, around 8–9% of variation in student achievements is explained by differences across schools, and in Germany around 15%, which shows that the decrease in average performance cannot be identified by a poorer performance of a particular subset of schools. In Denmark, the increase in performance was associated with an increase of the intraclass correlation of around 14.9%, indicating that a subset of schools performed much better in 2015.

The different hypotheses were formally tested using the Bayes factor with a uniform prior and the generalized fractional Bayes factor with an improper prior. In Table 2 the results of the Bayes factor based on uniform priors, referred to as BF, and the generalized fractional Bayes factor, referred to as FBF, are reported, including the posterior probability of each hypothesis. First, the invariant positive intraclass hypothesis was evaluated against the variant intraclass hypothesis. For the BF, it was concluded that

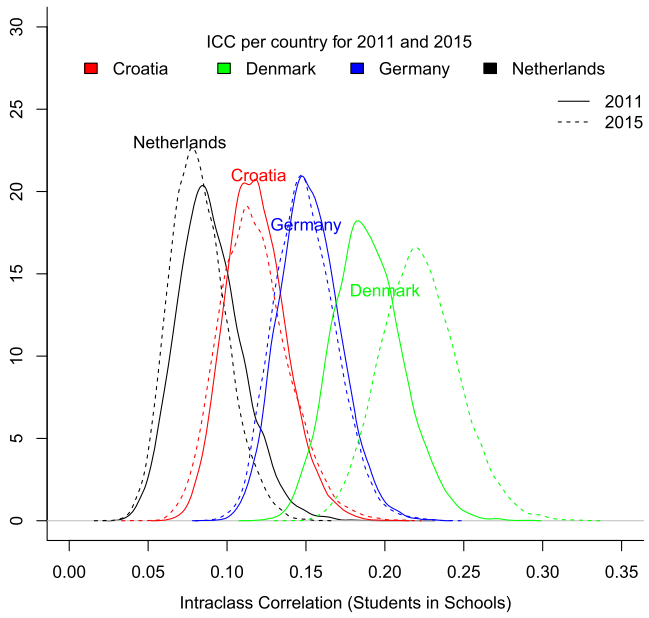

Figure 6: Posterior distribution of the intraclass correlation of 4th-graders nested in schools for four countries assessed by TIMSS 2011 and 2015.

	$\log B_{14}$	$\log B_{24}$	$\log B_{34}$	$P(H_1 \mid \mathbf{y})$	$P(H_2 \mid \mathbf{y})$	$P(H_3 \mid \mathbf{y})$	$P(H_4 \mid \mathbf{y})$
BF	2.24	13.55	-0.95	.00	1.00	.00	.00
FBF	0.55	13.42	-1.86	.00	1.00	.00	.00

Table 2: Results of the tests on intraclass correlations for the four countries (Netherlands, Croatia, Germany, and Denmark) in TIMSS 2011 and 2015.

there was positive evidence for the H_1 ($B_{14} > 3$) representing invariant positive intraclass correlations across years and countries, when comparing it to H_4. For the FBF, there was less evidence in favour of H_0, although both the BF and the FBF indicated support for the invariant hypothesis. Second, for the BF and the FBF, the (positive) time-invariant and country-ordered-variant hypothesis H_2 was very strongly preferred over the variant hypothesis. Finally, the (positive) time-variant country-invariant hypothesis H_3 was not preferred over H_4, where the FBF showed positive evidence for H_4 and the BF showed some evidence in favour of H_4. When also including the results from the posterior probabilities of the hypotheses, it was concluded that the positive intraclass correlations differed across countries, and that an order in intraclass correlations was identified. Within each country, the intraclass correlations did not appear to differ across years.

The present analysis showed that having accurate information about the stratification can be beneficial across years, since changes in the intra-correlation coefficient were invariant over time. The intraclass correlations differed across countries, although

the estimated correlations did not differ that much and varied from .08 to .22. Nevertheless, efficient sampling strategies are needed in countries with positive intraclass correlations, where countries with higher intraclass correlations will benefit more from efficient stratification strategies. Hedges and Hedberg (2007) also reported intraclass correlations for different large-scale surveys to provide information for employing randomized experiments in eduction, where schools are assigned to treatments. However, only pairs of intraclass correlations were compared using a Bonferroni adjustment, and the estimated intraclass correlations were assumed to be approximately normally distributed to evaluate the significance of a difference in correlations. These limitations do not apply to the developed generalized fraction Bayes factor and Bayes factor test for intraclass correlations.

8 Discussion

Currently there are two well-known approaches to model grouped data. In the population-average approach the correlation is treated as a nuisance and the marginal expectation of the outcome is modeled as a function of explanatory variables (Liang and Zeger, 1986). In the conditional or group-specific approach, the variability between groups is explicitly modeled using random effects, which measure directly the heterogeneity between groups. The marginal modeling approach outlined in this paper introduces a third approach. The random effects are integrated out in the conditional model, and the marginal mean and implied covariance structure are directly modeled to make inferences about the correlation structure. Under the integrated likelihood, a prior class is presented which can be used when prior information is available or absent. The new shifted F prior can be seen as an extension of popular priors for variance components (Gelman, 2006; Polson and Scott, 2012; Pérez et al., 2017; Mulder and Pericchi, 2018). The prior can be used when prior information is available or absent. In the latter case an improper prior is considered which results in frequentist matching credible intervals. Furthermore, posterior sampling is efficient using a Gibbs sampler via a parameter expansion. It is also straightforward to compute the probability of whether a between-groups variance parameter is less than (or greater than) zero under the proposed marginal approach. Support for a negative between-groups variance can indicate that a random effects model is not appropriate for the data or that the sample size is too small. Unlike the methodology of Kinney and Dunson (2007), no proper prior has to be specified to obtain such posterior probabilities. Finally the numerical performances of the proposed Bayes factor and generalized fractional Bayes factor showed accurate and consistent results.

Although other methods have been proposed for testing intraclass correlations, no general method has been proposed for the testing problem in (1). The classical significance tests proposed by Donner and Zou (2002) and Konishi and Gupta (1989) are limited to testing a null hypothesis of equal intraclass correlations against an unconstrained alternative. Note that significance tests in general are not designed for testing multiple hypotheses simultaneously or for testing nonnested hypotheses with order constraints on the parameters of interest (Silvapulle and Sen, 2004). Furthermore, the Bayesian information criterion (BIC; Schwarz, 1978; Raftery, 1995) is also

not suitable for this testing problem because (i) the Gaussian approximation of the posterior of the intraclass correlations, $\boldsymbol{\rho}$, would be inaccurate for small to moderate samples, (ii) the normally distributed unit-information prior may not be suitable for the bounded interval of intraclass correlations, and (iii) the number of free parameters is ill-defined for hypotheses with order constraints on the parameters (Mulder et al., 2009). Furthermore Bayes factors have been proposed for testing whether a single intraclass correlation equals zero (García-Donato and Sun, 2007; Pauler et al., 1999; Westfall and Gönen, 1996). The Bayes factor test of Mulder and Fox (2013) assumes uniform priors for the intraclass correlations which are not suitable for general use. Pauler et al. (1999) proposed a tailor-made prior, based on the unit of information, to use MCMC for calculating Bayes factors while dealing with the boundary null-hypothesis. However, this truncated-normal prior is not appropriate for order-constrained hypotheses, since the number of groups and the number of within-group observations can vary across different types of groups. This complicates the specification of a noninformative prior to evaluate inequality constrained hypotheses. Furthermore, these authors considered Bayes factors from a multilevel modeling framework. The integration of the joint posterior with respect to the random effect parameters however is computationally challenging and also requires specification of priors for the (random effect) nuisance parameters whose choice might be ambiguous (Berger et al., 1999). Therefore, a more general Bayesian testing framework was presented to make inferences when testing multiple hypotheses with equality constraints and/or order constraints on the intraclass correlations when prior information about the intraclass correlations is available, weak or completely unavailable. Thereby the paper contributes to the increasing literature on Bayes factor tests of equality and order constrained hypotheses (e.g. Hoijtink, 2011; Gu et al., 2014; Braeken et al., 2015; Mulder, 2016; Böing-Messing et al., 2017, and the references therein), which are becoming increasingly popular in the social and behavioral sciences.

The Bayes factor tests have been developed for continuous data. Future research will focus on extending the tests to categorical and count data by using an appropriate data augmentation scheme (Albert and Chib, 1993; Fox, 2010). Fox et al. (2017) proposed Bayes factor tests for the covariance parameter in a multivariate probit model, with a compound-symmetry covariance structure using data augmentation. For categorical data, the intraclass correlation is often used to determine, for instance, the test reliability of a scoring system, where the object is to obtain compatible results in different statistical trials. When the measurement error remains stationary, the intraclass correlation increases in line with increasing subject variability, which demonstrates that subjects can be better distinguished from each other.

In the psychometric application, the Bayes factor was used as a confirmatory tool to determine which hypothesis of a set of four hypotheses with competing constraints on the intraclass correlations receives most evidence from the data. The proposed Bayes factors can also be used for a more exploratory analysis to find the best fitting hypothesis of all possible equality/order constrained hypotheses, similar as in a variable selection problem. In such an exploratory approach it would be recommended to correct for multiple testing, e.g., using the work of Scott and Berger (2006). How to do this in the case of equality and order constrained models on intraclass correlations is an open topic for further research.

For unbalanced designs the number of observations can vary across groups. The distribution of the between-groups variance is then a mixture of shifted inverse-gamma distributions where the shift parameter depends on the group size. The closed-form distributions from the balanced case can be used to generate proposals for a Metropolis-Hastings algorithm. Furthermore, they can also serve as importance sampling functions to compute Bayes factors concerning hypothesis of the intraclass correlation for the unbalanced situation. More research is needed to examine the numerical performances and appropriate priors for making inferences about the intraclass correlation in an unbalanced design.

Supplementary Material

The supplementary material for "Bayes Factor Testing of Multiple Intraclass Correlations" (DOI: 10.1214/18-BA1115SUPP; .pdf). The supplementary material for "Bayes factor testing of multiple intraclass correlations" contains the proof of Lemma 1, the proof of Lemma 2, the conditional posterior distributions for the Gibbs sampler, the proof of Lemma 3, the Gibbs sampler under a constrained model, the analytic expression of the marginal likelihood (with derivation) for a standard random intercept model using fractional Bayes methodology, and a simulation study when testing interval hypotheses.

References

Albert, J. H. and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88(422): 669–679. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476321 MR1224394. 25

Armagan, A., Dunson, D. B., and Clyde, M. (2011). "Generalized Beta Mixtures of Gaussians." In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems 24*, 523–531. 7

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40: 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 12

Berger, J. O. (1994). "An overview of robust Bayesian analysis (with discussion)." *Test*, 3: 5–124. MR1293110. doi: https://doi.org/10.1007/BF02562676. 12

Berger, J. O. (2006). "The case for objective Bbayesian analysis." *Bayesian Analysis*, 1: 385–402. MR2221271. doi: https://doi.org/10.1214/06-BA115. 14

Berger, J. O. and Bernardo, J. (1992). *Reference priors in a variance component's problem*, 323–340. London: Oxford University Press. MR1194392. doi: https://doi.org/10.1007/978-1-4612-2944-5_10. 8, 11

Berger, J. O., De Oliveira, V., and Sansó, B. (2006). "Objective Bayesian analysis of spatially correlated data." *Journal of the American Statistical Association*, 96: 1361–1374. MR1946582. doi: https://doi.org/10.1198/016214501753382282. 11

Berger, J. O., Liseo, B., and Wolpert, R. L. (1999). "Integrated likelihood methods for eliminating nuisance parameters." *Statistical Science*, 14(1): 1–22. URL http://www.jstor.org/stable/2676641 MR1702200. doi: https://doi.org/10.1214/ss/1009211803. 2, 25

Berger, J. O. and Mortera, J. (1995). "Discussion to fractional Bayes factors for model comparison (by O'Hagan)." *Journal of the Royal Statistical Society Series B*, 56: 130. MR1325379. 14, 16

Berger, J. O. and Pericchi, L. (2001). "Objective Bayesian methods for model selection: Introduction and comparison (with discussion)." In Lahiri, P. (ed.), *Model Selection*, volume 38 of *Monograph Series*, 135–207. Beachwood Ohio, institute of mathematical statistics lecture notes edition. MR2000753. doi: https://doi.org/10.1214/lnms/1215540968. 12, 14

Berger, J. O. and Pericchi, L. R. (1996). "The intrinsic Bayes factor for model selection and prediction." *Journal of the American Statistical Association*, 91: 109–122. MR1394065. doi: https://doi.org/10.2307/2291387. 12

Berger, J. O. and Sun, D. (2008). "Objective priors for the bivariate normal model." *The Annals of Statistics*, 36: 963–982. MR2396821. doi: https://doi.org/10.1214/07-AOS501. 3

Böing-Messing, F., van Assen, M. A. L. M., Hofman, A., Hoijtink, H., and Mulder, J. "Bayesian evaluation of constrained hypotheses on variances of multiple independent groups." *Psychological Methods*, 22: 262–287. doi: https://doi.org/10.1037/met0000116. 25

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Snalysis*. Reading, MA: Addison-Wesley. MR0418321. 8

Braeken, J., Mulder, J., and Wood, S. (2015). "Relative effects at work: Bayes factors for order hypotheses." *Journal of Management*, 41: 544–573. doi: https://doi.org/10.1177/0149206314525206. 25

Chung, Y. and Dey, D. K. (1998). "Bayesian approach to estimation of intraclass correlation using reference prior." *Communications in Statistics*, 26: 2241–2255. MR1646661. doi: https://doi.org/10.1080/03610929808832225. 8, 11, 12

Conigliani, C. and O'Hagan, A. (2000). "Sensitivity of the fractional Bayes factor to prior distributions." *The Canadian Journal of Statistics*, 28: 343–352. MR1792054. doi: https://doi.org/10.2307/3315983. 16

Consonni, G. and Paroli, R. (2017). "Objective Bayesian Comparison of Constrained Analysis of Variance Models." *Psychometrika*, 82: 589–609. MR3688962. doi: https://doi.org/10.1007/s11336-016-9516-y. 15

De Santis, F. and Spezzaferri, F. (2001). "Consistent fractional Bayes factor for nested normal linear models." *Journal of Statistical Planning and Inference*, 97: 305–321. MR1861156. doi: https://doi.org/10.1016/S0378-3758(00)00240-8. 3, 14

Donner, A. and Zou, G. (2002). "Testing the equality of dependent intraclass correlation coefficients." *Journal of the Royal Statistical Society Series D*, 51: 379–379. MR1920764. doi: https://doi.org/10.1111/1467-9884.00324. 24

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer. MR2657265. doi: https://doi.org/10.1007/978-1-4419-0742-4. 25

Fox, J.-P., Mulder, J., and Sinharay, S. (2017). "Bayes factor covariance testing in item response models." *Psychometrika*, 82: 979–1006. MR3736338. doi: https://doi.org/10.1007/s11336-017-9577-6. 3, 25

García-Donato, G. and Sun, D. (2007). "Objective priors for hypothesis testing in one-way random effects models." *Canadian Journal of Statistics*, 35: 302–320. MR2393611. doi: https://doi.org/10.1002/cjs.5550350207. 13, 25

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1: 515–534. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 3, 12, 24

Ghosh, J. K. and Mukerjee, R. (1992). *Non-informative priors*, 195–210. Oxford University Press. MR1380277. 11

Gilks, W. R. (1995). "Discussion to fractional Bayes factors for model comparison (by O'Hagan)." *Journal of the Royal Statistical Society Series B*, 56: 118–120. MR1325379. 16

Griffin, J. and Brown, P. (2005). "Alternative prior distributions for variable selection with very many more variables than observations." Technical report, University of Warwick. 12

Gu, X., Mulder, J., and Hoijtink, H. (2014). "Bayesian evaluation of inequality constrained hypotheses." *Psychological Methods*, 9(4): 511-527. doi: https://doi.org/10.1037/met0000017. 25

Gu, X., Mulder, J., Decovíc, M., and Hoijtink, H. (2017). "Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses." *British Journal of Mathematical and Statistical Psychology*. 15

Haldane, J. B. S. (1932). "A note on inverse probability." *Mathematical Proceedings of the Cambridge Philosophical Society*, 28: 55–61. 8

Harris, J. A. (1913). "On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large." *Biometrika*, 9: 446–472. 2

Hedges, L. V. and Hedberg, E. C. (2007). "Intraclass Correlation Values for Planning Group-Randomized Trials in Education." *Educational Evaluation and Policy Analysis*, 29(1): 60–87. URL http://www.jstor.org/stable/30128045 1, 21, 24

Hoijtink, H. (2011). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. New York: Chapman & Hall/CRC. 25

Hoijtink, H., Gu, X., and Mulder, J. (2018). "Bayesian evaluation of informative hypotheses for multiple populations." *British Journal of Mathematical and Statistical Psychology*. 3, 14

Jeffreys, H. (1961). *Theory of Probability-3rd ed*. New York: Oxford University Press. MR0187257. 12, 13

Johnson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society Series B*, 72: 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 19

Kass, R. E. and Vaidyanathan, S. K. (1992). "Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions." *Journal of the Royal Statistical Society, Series B*, 54: 129–144. MR1157716. 13

Kass, R. E. and Wasserman, L. (1995). "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association*, 90: 928–934. MR1354008. 7

Kinney, S. and Dunson, D. B. (2007). "Fixed and Random Effects Selection in Linear and Logistic Models." *Biometrics*, 63: 690–698. MR2395705. doi: https://doi.org/10.1111/j.1541-0420.2007.00771.x. 24

Kish, L. (1965). *Survey sampling*. Chichester: Wiley. 21

Konishi, S. and Gupta, K. A. (1989). "Testing the equality of several intraclass correlation coefficients." *Journal of Statistical Planning and Inference*, (1): 93–105. MR0995594. doi: https://doi.org/10.1016/0378-3758(89)90022-0. 24

Lancaster, H. O. (1965). "The Helmert matrices." *The American Mathematical Monthly*, 72: 4–12. MR0170899. doi: https://doi.org/10.2307/2312989. 2, 5

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection." *Journal of American Statistical Association*, 103(481): 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 12

Liang, K.-Y. and Zeger, S. L. (1986). "Longitudinal data analysis using generalized linear models." *Biometrika*, 73(1): 13. MR0836430. doi: https://doi.org/10.1093/biomet/73.1.13. 2, 24

Ly, A., Verhagen, J., and Wagenmakers, E.-J. (2016). "An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys." *Journal of Mathematical Psychology*, 72: 43–55. MR3506025. doi: https://doi.org/10.1016/j.jmp.2016.01.003. 13

Mulder, J. (2014). "Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses." *Computational Statistics and Data Analysis*, 71: 448–463. MR3131982. doi: https://doi.org/10.1016/j.csda.2013.07.017. 14

Mulder, J. (2016). "Bayes factors for testing order-constrained hypotheses on correlations." *Journal of Mathematical Psychology*, 72: 104–115. doi: https://doi.org/10.1016/j.jmp.2014.09.004. 25

Mulder, J. and Fox, J.-P. (2013). "Bayesian tests on components of the compound symmetry covariance matrix." *Statistics and Computing*, 23: 109–122. MR3018353. doi: https://doi.org/10.1007/s11222-011-9295-3. 2, 25

Mulder, J. and Fox, J.-P. (2018). "Supplemetary material for "Bayes factor testing of multiple intraclass correlations"." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1115SUPP. 9, 10, 15, 16, 18, 20

Mulder, J., Klugkist, I., van de Schoot, A., Meeus, W., Selfhout, M., and Hoijtink, H. (2009). "Bayesian Model Selection of Informative Hypotheses for Repeated Measurements." *Journal of Mathematical Psychology*, 53: 530–546. MR2574692. doi: https://doi.org/10.1016/j.jmp.2009.09.003. 25

Mulder, J. and Pericchi, L. R. (2018). "The matrix-F prior for estimating and testing covariance matrices." *Bayesian Analysis*. 3, 12, 24

Mullis, I., Martin, M. O., Foy, P., and Arora, A. (2011). *TIMSS 2011 International Results in Mathematics*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College Chestnut Hill, MA, USA and International Association for the Evaluation of Educational Achievement (IEA) IEA Secretariat, Amsterdam, the Netherlands, first edition. 21

O'Hagan, A. (1995). "Fractional Bayes Factors for Model Comparison (with discussion)." *Journal of the Royal Statistical Society Series B*, 57: 99–138. MR1325379. 3, 12, 14, 16

O'Hagan, A. (1997). "Properties of intrinsic and fractional Bayes factors." *Test*, 6: 101–118. MR1466435. doi: https://doi.org/10.1007/BF02564428. 16

Olson, J. F., Martin, M. O., and Mullis, I. V. S. (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College, first edition. 21

Pauler, D. K., Wakefield, J. C., and Kass, R. E. (1999). "Bayes Factors and Approximations for Variance Component Models." *Journal of the American Statistical Association*, 94(448): 1242–1253. MR1731486. doi: https://doi.org/10.2307/2669938. 25

Pérez, M. E., Pericchi, L. R., and Ramirez, I. C. (2017). "The Scaled Beta2 Distribution as a Robust Prior for Scales." *Bayesian Analysis*, 12: 615. MR3655869. doi: https://doi.org/10.1214/16-BA1015. 3, 12, 24

Pericchi, L. R. (2005). "Model selection and hypothesis testing based on objective probabilities and Bayes Factors." In D. K. Dey, C. R. Rao (eds.), *Bayesian Thinking: Modeling and Computation*, 115–149. Elsevier, Amsterdam, The Netherlands. 11

Pericchi, L. R., Liu, G., and Torres, D. (2008). "Objective Bayes factors for informative hypotheses: "Completing" the informative hypothesis and "splitting" the Bayes

factors." In H. Hoijtink, I. Klugkist, and P. Boelen (eds.), *Bayesian Evaluation of Informative Hypotheses*, 131–154. New York: Springer. 13, 15

Polson, N. G. and Scott, J. G. (2012). "On the Half-Cauchy Prior for a Global Scale Parameter." *Bayesian Analysis*, 7. MR3000018. doi: https://doi.org/10.1214/12-BA730. 3, 24

Raftery, A. E. (1995). "Bayesian model selection in social research." *Sociological Methodology*, 25: 111–163. 24

Raudenbush, S. W. (1997). "Statistical analysis and optimal design for cluster randomized trials." *Psychological Methods*, (2): 173–185. URL http://dx.doi.org/10.1037/1082-989X.2.2.173 1, 2

Schwarz, G. E. (1978). "Estimating the dimension of a model." *Annals of Statistics*, 6: 461–464. MR0468014. 24

Scott, J. G. and Berger, J. O. (2006). "An exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136: 2144–2162. MR2235051. doi: https://doi.org/10.1016/j.jspi.2005.08.031. 25

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. John Wiley & Sons, Inc., Hoboken, NJ, USA, first edition. MR1190470. doi: https://doi.org/10.1002/9780470316856. 2, 5

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). "Calibration of p Values for Testing Precise Null Hypotheses." *The American Statistician*, 55: 62–71. doi: https://doi.org/10.1198/000313001300339950. 11

Severini, T. A., Mukerjee, R., and Ghosh, M. (2002). "On an exact probability matching property of right-invariant priors." *Biometrika*, 89: 952–957. MR1946524. doi: https://doi.org/10.1093/biomet/89.4.952. 3

Silvapulle, M. J. and Sen, P. K. (2004). *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*. Hoboken, NJ: John Wiley, second edition. MR2099529. 24

Spiegelhalter, D. J. (2001). "Bayesian methods for cluster randomized trials with continuous responses." *Statistics in Medicine*, 20(3): 435–452. URL http://dx.doi.org/10.1002/1097-0258(20010215)20:3<435::AID-SIM804>3.0.CO;2-E 1, 7

Stein, C. (1985). *On the coverage probability of confidence sets based on a prior distribution*, 485–514. PWN-Polish Scientific Publishers. MR0847495. 11

Van Geel, M., Keuning, T., Visscher, A., and Fox, J.-P. (2017). "Changes in educators' data literacy during a data-based decision making intervention." *Teaching and Teacher Education*, 64: 187–198. 2

Welch, B. and Peers, H. W. (1963). "On formulae for confidence points based on integrals of weighted likelihoods." *Journal of the Royal Statistical Society Series B*, 25: 318–29. MR0173309. 3

Westfall, P. and Gönen, M. (1996). "Asymptotic properties of ANOVA Bayes factors." *Communications in Statistics: Theory and Methods*, 25: 3101–3123. MR1422324. doi: https://doi.org/10.1080/03610929608831888. 25

Ye, K. (1994). "Bayesian reference prior analysis on the ratio of variances for the balanced one-way random effect model." *Journal of Statistical Planning and Inference*, 41: 267–280. MR1309613. doi: https://doi.org/10.1016/0378-3758(94)90023-X. 11

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." In Goel, P. K. and Zellner, A. (eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti*, 233–243. Amsterdam: Elsevier. MR0881437.

Zellner, A. and Siow, A. (1980). "Posterior odds ratios for selected regression hypotheses." In J. M. Bernardo, D. V. Lindley, A. F. M. Smith, M. H. DeGroot (ed.), *Bayesian statistics: proceedings of the first international meeting held in Valencia*, 585–603. Spain: University of Valencia. MR0862503. 12

Acknowledgments

We would like to thank Luis Raul Pericchi for insightful discussions about the F distribution. Furthermore we would like to thank three anonymous reviewers and the editor for insightful comments which greatly improved the quality of the paper. The first author was supported by a Veni grant from the Netherlands Organization for Scientific Research (NWO).