Journal of Mathematical Psychology 72 (2016) 144–157

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs

Farouk S. Nathoo, Michael E.J. Masson *

University of Victoria, Canada

HIGHLIGHTS

- We review an approximation method for Bayesian analysis of data from ANOVA designs.
- We derive the correct value for number of observations in the repeated-measures case.
- We derive a closed-form solution for posterior distributions for this approximation.
- We compare this approximation method to another Bayesian method and to NHST.

ARTICLE INFO

Article history:
Available online 8 April 2015

Keywords:
Bayesian estimation
Bayes factors
Null-hypothesis significance testing
Repeated-measures designs

ABSTRACT

We present a mathematical derivation that establishes the validity of a proposed adaptation to repeated-measures designs of Wagenmakers' (2007) Bayesian information criterion (BIC) method for estimating Bayes factors. We also introduce an improved definition of the penalty in this BIC approximation that accommodates the repeated-measures correlation through an effective sample size based on the Fisher Information. Monte Carlo simulations of repeated-measures data were used to compare the BIC method to two Bayesian procedures for analysis of variance (ANOVA) designs and to the standard null-hypothesis significance testing (NHST) approach. When no effects of the independent variable were present in the populations and a reasonable sample size was used, the Bayesian methods consistently yielded posterior probabilities clearly favoring the null model. We discuss two different approaches to comparing the outcome of the Bayesian analyses with NHST results when an effect is present. In general, a direct comparison between NHST p values and Bayesian posterior probabilities indicates that the latter is somewhat conservative when effect size is small. We also derive a closed-form expression for approximating the posterior probability distributions for condition means in one-factor repeated-measures designs and present an R routine for computing these distributions and the posterior probability of H_0 that requires as input nothing more than values from a standard ANOVA.

1. Introduction

A substantial change in how experimental psychologists and cognitive scientists statistically analyze their data and test theoretical propositions is currently underway. One officially sanctioned change is reliance on estimation methods such as confidence intervals for effect sizes (Cumming, 2014), which has been formally adopted by a highly influential journal, *Psychological Science*. Another important alternative that is gaining traction is the Bayesian approach to model comparison and estimation (e.g., Kruschke, 2011, 2013; Rouder, Morey, Speckman, & Province, 2012; Schoot et al., 2014 and Wagenmakers, 2007). There are other model comparison approaches that have been advocated as well, such as likelihood ratios with a correction for number of free parameters in the models (e.g., Glover & Dixon, 2004), but Bayesian methods appear to have achieved a greater degree of acceptance in the behavioral sciences. Our purpose in this article is to encourage the use of Bayesian methods by making available a straightforward method of generating a Bayesian analysis for the standard repeated-measures design commonly used in experimental psychology. This method requires little more than computing the usual analysis of variance (ANOVA). It builds on the proposal by Wagenmakers (2007) by providing validation of an extension of his method to repeated-measures designs.

In large measure, the call for change in how we analyze empirical data is a response to substantial problems associated

* Correspondence to: Department of Psychology, University of Victoria, P.O. Box 1700 STN CSC, Victoria, British Columbia V8W 2Y2, Canada.
E-mail address: mmasson@uvic.ca (M.E.J. Masson).

http://dx.doi.org/10.1016/j.jmp.2015.03.003
0022-2496/

with the widespread use of null-hypothesis significance testing (NHST). We begin by reviewing four of these problems and explaining how a Bayesian approach provides powerful solutions to them. First, the p value generated by a significance test in NHST does not provide the information that researchers actually seek, even though they tend to interpret the result as though it does (Cohen, 1994). In particular, an NHST p value provides the probability that the observed data (D), or a more extreme outcome, would occur, under the assumption that the null hypothesis (H_0) is true. But in fact, our interest is in the viability of a hypothesis given the observed data. That is, NHST delivers $p(D|H_0)$, but researchers wish to draw inferences of the form $p(H|D)$, where H is some hypothesis. This value is readily obtained from a Bayesian analysis.

Second, researchers often fail to obtain evidence that allows rejection of the null hypothesis under NHST methods. Strictly speaking, when this happens no strong conclusions can be drawn from the results (Wilkinson & the Task Force on Statistical Inference, 1999). Yet there are many instances in which researchers actually expect to obtain a null result. In the context of NHST, the best one can do in these cases is to provide a power estimate based on some non-zero effect size. Even when an acceptably high value for statistical power is obtained, however, one must concede than an effect of smaller size might exist. A great benefit of Bayesian analysis is that it provides an estimate of how strongly the empirical results support not only an hypothesized model that assumes an effect is present, but also how strongly the null model or hypothesis is supported.

Third, NHST methods are susceptible to contamination by data collection practices that may be adopted out of ignorance and with completely innocent intent. Specifically, even with an exactly true null hypothesis, a researcher who continues to collect and analyze data as it arrives until a significant p value is obtained using NHST is guaranteed to obtain a significant effect at some point (Armitage, McPherson, & Rowe, 1969; Wagenmakers, 2007). This practice of *optional stopping* can substantially elevate type I error probability in NHST. Although there is practical value in monitoring data as they come in to determine whether sufficient evidence has been obtained to allow a decision between competing hypotheses, this practice is a serious problem within the NHST framework and highly inflates the probability of a type I error. Bayesian inference, however, is compatible with application of the optional stopping heuristic and will yield increasingly reliable results as more data accumulate (Berger & Berry, 1988; Kruschke, 2013; Wagenmakers, 2007). Wagenmakers provides a thorough explanation of this issue in an online Appendix to his 2007 article (www.ejwagenmakers.com/2007/StoppingRuleAppendix.pdf). Essentially, if we monitor $p(H|D)$ for a true hypothesis, H, and stop whenever this probability falls below some low threshold, there is a limit on how often this procedure will succeed. For example, if we monitor $p(H|D)$ for a true hypothesis with the plan to stop collecting data if that probability drops as low as 0.05, then 19 times out of 20 we will never reach that threshold no matter how long we keep collecting data (see also Edwards, Lindman, & Savage, 1963).

Finally, if a null hypothesis is rejected, the NHST framework offers little or no guidance with respect to a specific alternative hypothesis. Indeed, this is one of the motivations behind the current movement that favors reporting of effects sizes and confidence intervals (Cumming, 2014). This approach, however, has its own shortcomings. One concern is that researchers well acquainted with NHST reasoning are likely to interpret confidence intervals (which often are conveniently defined to provide 95% confidence) as a tool to determine statistical significance of an effect. This is quite easy to do. For example, if an effect size is plotted with a 95% confidence interval, whether or not the interval includes zero determines whether the null hypothesis is rejected under NHST. Second, the confidence interval provides no information about where the probable value of a parameter (a mean or an effect size) lies within that interval (Kruschke, 2013). That is, given a confidence interval of 10–20, a value of 10 is just as credible as a value of 15. Finally, Morey, Rouder, Verhagen, and Wagenmakers (2014) point out that testing a theory requires predictions about what data should be like if the theory is true versus false, and it requires a method for using the data to make an inference about the theory. Estimation procedures such as classic confidence intervals are limited to only the first of these three elements, characterizing data when the theory is true. If a theory predicts, for instance, that an effect should be present and the estimated effect size has a confidence interval that does not include zero, the researcher is likely to conclude that the data support the theory. Morey et al. point out that this conclusion is a logical fallacy (converse error or affirming the consequent) and that a principled method for inferring support for a theory is needed. Finally, confidence intervals are susceptible to the same misuse as NHST with respect to optional stopping. For example, even if the true effect size is zero, if one were to keep sampling and computing a confidence interval after each new subject is tested, it is guaranteed that if one continues long enough, one will obtain a confidence interval that does not include zero.

Fortunately, the Bayesian approach provides a solution to these difficulties as well. It is not tied to an emphasis on the null hypothesis, but instead provides a method for establishing the relative validity of competing hypotheses based on observed data. In addition, Bayesian methods can generate distributions of likely values of an estimated parameter such as an effect size, given the observed data. Moreover, a confidence interval does not provide a mechanism for assigning relative importance to different values lying within the interval, whereas a posterior distribution arising from a Bayesian analysis is more informative. The Bayesian posterior density is typically not uniform, and will be more concentrated in the central region of the distribution, while being sparse at the extreme ends. In addition, the posterior density can be used to assign a posterior probability to any subinterval of parameter values.

2. Practically useful Bayesian methods

A possible obstacle to widespread adoption of Bayesian analyses is the potentially complicated methods that these analyses can require. Chief among these are establishing defensible prior distributions for the relevant model parameters and the computation of Bayesian posterior distributions and posterior probabilities, which will typically involve some form of integration requiring complex numerical methods. A number of practical solutions to this obstacle have recently been made available to general users that do not require sophisticated knowledge of how to construct prior distributions nor implementation of numerical parameter estimation procedures.

Kruschke (2013) provided a Bayesian estimation method that can be used in place of a t test for testing differences between two independent samples. This method generates distributions of credible values for population means and standard deviations as well as the difference between population means. The prior distribution for population means discussed by Kruschke is intended for general applications and so it is generated by assuming little prior knowledge. Thus, the prior is a normal distribution with very high variability (1000 times the observed pooled standard deviation) and mean equal to the observed mean of the pooled data. Such a broad prior distribution is intended to have minimal impact on the posterior distributions that are produced by the Bayesian analysis. Estimation of posterior distributions proceeds using Markov chain Monte Carlo sampling. Kruschke provides a source code for use in the open source statistical program R. The steps required for using this code and entering data are relatively straightforward, and the

program generates posterior distributions that yield rich information about likely parameter values for the populations on which the two groups are based. Although Kruschke noted that the estimation procedure he described can be extended to more complex designs, doing so requires modification to code which may present a serious obstacle to many researchers.

Rouder et al. (2012) introduced a general Bayesian analysis method for a wide range of ANOVA designs. This analysis uses the multivariate generalizations of the Cauchy distribution as the prior for standardized effect size and a noninformative prior for variance. For most types of designs, the Rouder et al. approach uses Markov chain Monte Carlo sampling of the parameter space to produce posterior distributions. The primary results generated by their analysis, however, are Bayes factors for effects of interest in ANOVA designs. A Bayes factor provides the relative likelihood of the observed data under each of two competing hypotheses:

$$BF_{01} = \frac{Pr(Data\,|H_0)}{Pr(Data\,|H_1)} \tag{1}$$

where H_0 is the null hypothesis and H_1 is the alternative hypothesis that assumes an effect is present. This formulation assumes that data are distributed discretely. The probabilities are replaced by probability densities for continuously distributed data. A Bayes factor can be computed using this method to evaluate effects that are typically tested using NHST in any standard ANOVA design, including repeated-measures designs. A great advantage of the Rouder et al. method is that they provide an R package that carries out the computations after the user specifies the relevant independent variables in the design. An additional routine can be used to produce posterior distributions for effect parameters.

Unlike the estimation procedure used by Kruschke (2013), the Rouder et al. (2012) method is a model comparison procedure, and more specifically it compares models with different numbers of free parameters. For example, in a simple ANOVA with a single factor containing two levels, the null hypothesis model (no difference between the conditions represented by the two levels of the factor) has one less parameter than the alternative hypothesis model (a difference between conditions exists). This additional parameter represents the size of the difference between population means. Because of the difference in number of free parameters between these models, the alternative model is bound to give a better account of the data. Model comparison must therefore be based, not only on goodness-of-fit, but also on model complexity, where the preferred model will attain, in some sense, a more optimal balance of fit and complexity. Each of the two marginal likelihoods underlying the Bayes factor (1) contains an implicit penalty for model complexity arising from integration across the parameter space. For example, the numerator can be expressed as,

$$Pr(Data\,|H_0) = \int Pr(Data\,|H_0, \theta_0)p(\theta_0)d\theta_0$$

where θ_0 denotes the parameters of the model underlying H_0, and $p(\theta_0)$ denotes the density of the corresponding prior distribution. This is in contrast to the maximized likelihood function $L_0 = Pr(Data\,|H_0, \hat{\theta}_0)$ which is evaluated at only a single point of the parameter space, namely the maximum likelihood estimator $\hat{\theta}_0$, and thus measures model fit, but does not account for the number of parameters. It is this implicit penalty that is incorporated in the model comparison procedures developed by Rouder et al. (2012).

An alternative approach for Bayesian testing known as BIEMS is discussed in Mulder, Hoijtink, and de Leeuw (2012). BIEMS can be used for calculating Bayes factors for multivariate normal linear models with equality or inequality constraints between model parameters against a model with no constraints. The approach is based on using a subset of the data for automatic prior specification, which is for instance the case in the intrinsic Bayes factor (Berger & Pericchi, 1996). More specifically, the BIEMS Bayes factor is based on the conjugate expected-constrained posterior prior, and like the Rouder et al. (2012) approach this Bayes factor is estimated using MCMC.

Although the Mulder et al. (2012) and Rouder et al. (2012) Bayes factors can be applied to a wide range of ANOVA designs, an alternative Bayesian method that is easier to use and understand is the approximation method introduced by Wagenmakers (2007). This method implements the Bayesian information criterion (BIC) to select between two competing models (typically a null and alternative model). The BIC is a measure of model fit, based on the likelihood of the observed data given an optimal set of parameter values, and a penalty driven in part by the number of free parameters used by the model, as shown in the following equation:

$$BIC(H_i) = -2\log(L_i) + \kappa_i \log(n)$$

where L_i is the likelihood of the data for model H_i, κ_i is the number of free parameters for model H_i, and n is the number of observations. This method generates a Bayes factor, as does the Rouder et al. (2012) method, but Wagenmakers also provides a clear explanation of the steps needed to convert the Bayes factor into posterior probabilities of the form $p(H|D)$ for each of the models.

The BIC is based on a Laplace approximation to the Bayes factor and its computation does not require a sampling algorithm to estimate posterior probabilities. Rather, it assumes a unit information prior and an approximation to the Bayes factor can be computed in a few simple steps using sums of squares values drawn from a standard ANOVA. For researchers having little experience with Bayesian methods or simulation-based approximation of posterior distributions, this simplicity is very appealing. The unit information prior is a normal distribution centered at the value of the effect observed in the data and extending over the full distribution of observed data (Kass & Wasserman, 1995; Raftery, 1999). This prior concentrates its density in the region of plausible effect sizes with very little likelihood invested in extreme values of effect size.

The core of the BIC method for Bayesian analysis is the observation that the difference in BIC values for two competing models can be transformed into an approximation of the Bayes factor. Specifically, from Wagenmakers (2007) we have

$$BF_{01} = \frac{p_{BIC}(H_0|D)}{p_{BIC}(H_1|D)} = \exp(\Delta BIC_{10}/2)$$

where $\Delta BIC_{10} = BIC(H_1) - BIC(H_0)$. Critically, the value ΔBIC_{10} can be computed directly from sums of squares components of an ANOVA, making the computation of the approximate Bayes factor straightforward. This computation is justified based on the connection between the BIC and the Bayes factor through large sample theory. In particular, defining the Schwarz criterion as $S = -0.5\Delta BIC_{10}$ it can be shown that $\frac{S-\log B_{01}}{\log B_{01}} \rightarrow 0$ as $n \rightarrow \infty$. A limitation associated with the BIC in this context is that it may be a rough approximation to the Bayes factor in some settings. This method is valid if the errors of measurement are normally distributed, as is assumed when computing standard ANOVAs. Consider a simple example in which we have one independent variable manipulated between subjects, which produces two sum of squares components, one for the observed effect and one for error variability. A model that assumes the null hypothesis (no effect of the manipulation), is unable to account for either source of variability, whereas a model that assumes an effect is present, is able to account for the variability associated with the effect, but not for error variability. Unexplained, or error variability associated with each of the two models is used to compute ΔBIC_{10} as follows:

$$\Delta BIC_{10} = n\log\left(\frac{SSE_1}{SSE_0}\right) + (\kappa_1 - \kappa_0)\log(n), \tag{2}$$

Table 1
Descriptive terms for strength of evidence corresponding to ranges of posterior probability values proposed by Raftery (1995).

$P(H_i\|D)$	Evidence
0.50–0.75	Weak
0.75–0.95	Positive
0.95–0.99	Strong
>0.99	Very strong

where n is the number of observations, SSE_1 is the variability not explained by the model that assumes an effect is present (this would be the SSerror term from the ANOVA), and SSE_0 is the variability not explained by the null hypothesis model (both sums of squares from the ANOVA; SSerror + SSeffect = SStotal).

Taking a hypothetical numerical example for a one-factor between-subjects design, suppose we have three groups of 20 subjects each (60 subjects in all) and the ANOVA produces SSeffect = 220, SSerror = 938, and SStotal = 1158. As an aside, under NHST these results would yield $F(2, 57) = 6.68$, $p = 0.0025$. Note that the model that assumes an effect is present contains two more parameters than the null model, with this difference being equal to the number of degrees of freedom associated with the effect in question. So in the equation for computing ΔBIC_{10}, $\kappa_1 - \kappa_0 = 2$. Applying the BIC approximation to produce the relevant Bayes factor, we have

$$\Delta BIC_{10} = 60 \log\left(\frac{938}{1158}\right) + 2\log(60) = -4.453$$

and $BF_{01} = \exp(-4.453/2) = 0.108$. Computing posterior probabilities requires that the Bayes factor be combined with prior odds (the a priori degree of belief in the null and alternative hypotheses):

$$\frac{p(H_0|D)}{p(H_1|D)} = \frac{p(D|H_0)}{p(D|H_1)}\frac{p(H_0)}{p(H_1)}$$

where the first term on the right is the Bayes factor and the second term is the prior odds. In many cases a neutral position regarding prior odds is quite reasonable, whereby the null and alternative hypotheses are considered equally likely. Under that assumption, the ratio of posterior probabilities simply collapses to the Bayes factor:

$$\frac{p_{BIC}(H_0|D)}{p_{BIC}(H_1|D)} = 0.108\frac{0.5}{0.5} = 0.108.$$

With this ratio of posterior probabilities, the posterior probability for H_0 is simply

$$p_{BIC}(H_0|D) = \frac{BF_{01}}{1 + BF_{01}} = \frac{0.108}{1 + 0.108} = 0.097,$$

and the complement of that (i.e., $1 - p(H_0|D)$) is the posterior probability for H_1 : $p(H_1|D) = 1 - 0.097 = 0.903$. This degree of support for the alternative hypothesis is in the range considered to be "positive" evidence by the standards proposed by Raftery (1995). Table 1 shows the ranges of probability values that correspond to various descriptors of strength of evidence suggested by Raftery. We note that these suggested cutoffs are somewhat arbitrary and they are provided only to give the reader a general sense of how to interpret the range of possible posterior probabilities that might be generated by a Bayesian analysis.

3. Extension of the BIC approximation to repeated-measures

The BIC approximation method introduced by Wagenmakers (2007) was described only for between-subjects designs. This limitation is significant because the formula for computing the key element of the approximation, ΔBIC_{10}, includes n (number of observations) in both of the terms on the right hand side of (2). For an independent-samples design, n is readily interpreted as the number of subjects in the experiment. The situation is less clear in a repeated-measures design where each subject is tested in k different conditions. Wagenmakers pointed out that there is uncertainty regarding what value n should take on in these cases and reported that the standard choice is to set n equal to the number of subjects. Kass and Raftery (1995) is a classic statistical paper reviewing the use and computation of Bayes factors. There, the authors mention that the sample size appearing in the BIC needs to be considered carefully, and suggest that the sample size should be the rate at which the Hessian matrix of the log-likelihood function grows.

In an extension of the BIC approximation for Bayesian analysis, Masson (2011) made a different recommendation for n in the case of repeated-measures designs. He noted that in the BIC computation, n refers to the number of independent observations in the data. For a between-subjects design, this equates to the number of subjects. But in a repeated-measures design, each of n subjects has multiple scores (one per condition in the usual repeated-measures ANOVA). The resulting number of independent observations would then be $n(k - 1)$, where n is the number of subjects and k is the number of conditions. This interpretation of n for a repeated-measures factor was consistent with the definition used by Bortolussi and Dixon (2003) in their computation of likelihood ratios, and it has been implemented in a software package that computes power estimates for various designs along with the BIC approximation for Bayesian analysis (Campbell & Thompson, 2012). Given that n appears in both terms used to compute ΔBIC_{10}, this is quite an important issue that ought to be settled if the BIC approximation is to be used successfully across a range of experimental designs.

To resolve this issue we first considered the standard definition of the BIC, which is $BIC = -2\log(L) + \kappa\log(n)$, where L is the maximum likelihood of the data assuming some model, and κ is the number of free parameters in the model. The question at stake, given the convenient expression (2), is what value should be used in place of n in each of the two terms of this formula when a repeated-measures design is used. For the first term, n plays a hidden role that is revealed when we look at the formula for ΔBIC_{10}. To understand how n should be interpreted in the context of the maximum likelihood estimate, we derived the formulation of ΔBIC_{10} in the context of a repeated-measures design. This derivation appears in Appendix A. As this result shows, the correct expression to be used in place of n in the first term of the ΔBIC_{10} formula is what was suggested by Masson (2011), namely, the number of subjects multiplied by one less than the number of repeated-measures conditions.

As for the second (penalty) term in the BIC formula, Jones (2011) developed a method for computing what he calls "effective sample size" for repeated-measures designs. The approach is based on the Fisher information, and for a single-factor repeated-measures design takes the form $n_{eff} = \frac{nk}{1+\rho(k-1)}$ where ρ is the intraclass correlation. Depending on the correlation, this value varies from a maximum equal to the number of subjects times the number of conditions (i.e., the number of observations) to a minimum equal to the number of subjects. As the degree of correlation between conditions increases from 0 to 1, the effective sample size decreases from its maximum value to its minimum to reflect the reduction in independence among the scores obtained by a given subject. In recent work, Berger, Bayarri, and Pericchi (2014) develop a definition of effective sample size called TESS for use in model selection procedures. This definition applies to a fairly general class of linear models and the authors consider the specific form of TESS for a number of special cases. For the repeated measures mixed model considered in this article, TESS reduces to the effective sample size n_{eff} proposed by Jones (2011).

In practice, the value of ρ is not known and so an estimate based on the fitted model is used in its place when calculating n_{eff}. As the estimates obtained from each of the models corresponding to H_0 and H_1 need not be the same, the value of this effective sample size will depend on whether the model being fit corresponds to H_0 (no effect present) or H_1 (some effect present). Our choice to use different estimates of ρ for each of H_0 and H_1 is based on Jones (2011) who suggested that the maximum likelihood estimate of ρ should be used when computing the BIC for a given model. Doing so also removes the ambiguity in the choice of estimate for ρ, which is an issue that arises if only a single estimate is to be used when computing the BIC for both models. Arguably, a potential drawback of using the two model based estimates of ρ (as opposed to a single estimate) is that the effective sample size will be smaller under H_1 because of the higher estimate of the intraclass correlation arising from reduced error variance, and that this in-turn will reduce the penalty H_1 receives for having more parameters in the model. Practically, for typical datasets the impact of this should be negligible as both estimates will be close and we have verified this using several simulation studies (not reported).

We use the maximum likelihood estimate of ρ, which leads to expressions that can be obtained from the components of the standard repeated-measures ANOVA. Under H_1

$$n_{eff} = \begin{cases} \dfrac{n(SS_T - SS_C)}{SS_S}, & \text{if } kSS_S - SS_T + SS_C > 0 \\ nk, & \text{otherwise} \end{cases} \tag{3}$$

and under H_0

$$n_{eff} = \begin{cases} \dfrac{nSS_T}{SS_S}, & \text{if } kSS_S - SS_T > 0 \\ nk, & \text{otherwise} \end{cases} \tag{4}$$

where SS_T is the total sum of squares, SS_C is the sum of squares for the effect of the independent variable, and SS_S is the between-subjects sum of squares. The derivation of this formula from the equation given by Jones is shown in Appendix A.

Taking these two new proofs into account, we now offer a revised formula for the first step in computing the BIC approximation of the Bayesian posterior probability of an hypothesis, given observed data, using the results from a standard repeated-measures ANOVA:

$$\Delta BIC_{10} = \begin{cases} n(k-1)\log\left(\dfrac{SS_T - SS_C - SS_S}{SS_T - SS_S}\right) \\ \quad + (k+2)\log\left(\dfrac{n(SS_T - SS_C)}{SS_S}\right) \\ \quad -3\log\left(\dfrac{nSS_T}{SS_S}\right), \quad \text{if } kSS_S > SS_T \\ n\log\left(\dfrac{SS_S}{n}\right) + n(k-1) \\ \quad \times \log\left(\dfrac{SS_T - SS_C - SS_S}{n(k-1)}\right) \\ \quad - nk\log\left(\dfrac{SS_T}{nk}\right) - 3\log(nk) \\ \quad + (k+2)\log\left(\dfrac{n(SS_T - SS_C)}{SS_S}\right), \\ \quad \text{if } SS_T - SS_C < kSS_S \leq SS_T \\ nk\log\left(\dfrac{SS_T - SS_C}{SS_T}\right) + (k-1)\log(nk), \\ \quad \text{if } kSS_S \leq SS_T - SS_C \end{cases} \tag{5}$$

where n is the number of subjects, k is the number of conditions, which is also the number of repeated measurements obtained from each subject, and where log refers to the natural logarithm. We note that the definition that holds under the first case corresponds to the definition given by Masson (2011), except that the penalty term now incorporates Eqs. (3) and (4). We further note that of the three cases appearing in (5), the first case corresponds to the situation where the models underlying H_0 and H_1 both produce non-zero estimates of the intraclass correlation, and this is by far the most likely scenario in behavioral data from repeated-measures designs. The second and third cases are included primarily for completeness. Computation of the Bayes factor and the posterior probabilities proceed as before. Note that as the strength of the correlations between conditions increases, n_{eff} will approach the number of subjects. This relationship means that for relatively strong correlations, the resulting penalty for the alternative hypothesis model will be less severe than in the original formulation recommended by Masson (2011).

4. Posterior distribution of parameters

In addition to using the BIC approximation to compute Bayes factors and posterior probabilities, we provide an approach for generating posterior distributions that show the most likely values of the condition means. Whereas the Rouder et al. (2012) method for computing Bayes factors and posterior distributions for parameter values depends on Markov chain Monte Carlo (MCMC) sampling methods, we adopt an alternative approach that expresses approximate posterior distributions in closed form, and importantly, as a function of only standard statistics that are computed as part of a typical repeated-measures ANOVA. Although sampling methods based on MCMC are far more general, our approach is specifically developed for users who may be transitioning to the use of Bayesian methods, and who may be more comfortable generating Bayesian solutions that can be expressed in closed form as a function of only the condition means and elements of the standard ANOVA.

The approximations are derived in Appendix B and are based on the Bayesian Central Limit Theorem (see e.g., Carlin & Louis, 1996, pp. 142–145) which leads to a large sample Gaussian approximation. For each of the k conditions we let $\overline{Y}_j$, $j = 1, \ldots, k$ denote the sample mean with $\overline{\boldsymbol{Y}} = (\overline{Y}_1, \ldots, \overline{Y}_k)$, μ_j the corresponding population mean with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$, and SS_c, SS_T, SS_S are the effect, total, and between-subjects sum of squares respectively. The approximation takes the form

$$\boldsymbol{\mu}|\text{Data} \overset{approx}{\sim} MVN\left(\overline{\boldsymbol{Y}}, \frac{1}{n}\hat{\boldsymbol{\Sigma}}\right)$$

where

$$\hat{\boldsymbol{\Sigma}} = \begin{cases} \dfrac{(SS_T - SS_C - SS_S)}{n(k-1)}\boldsymbol{I}_k + \dfrac{(kSS_S - SS_T + SS_C)}{nk(k-1)} \\ \quad \times \mathbf{1}_k\mathbf{1}_k', \quad \text{if } kSS_S - SS_T + SS_C > 0 \\ \dfrac{(SS_T - SS_C)}{nk}\boldsymbol{I}_k, \quad \text{otherwise.} \end{cases}$$

We note that the form presented above does not depend on any prior distribution assumed for the condition means. Being a large sample approximation, our approach obviates the need to consider priors. In part because of the convenience of the closed-form solutions offered by the BIC and large sample Gaussian approximations, we are able to provide a set of commands that can be run as a source file in the open source statistical program R to compute posterior probabilities for competing hypotheses and posterior distributions for condition means for simple one-factor, repeated-measures designs. The supplementary material provided online includes the source file for making these computations using only condition means and elements of the standard ANOVA as input.

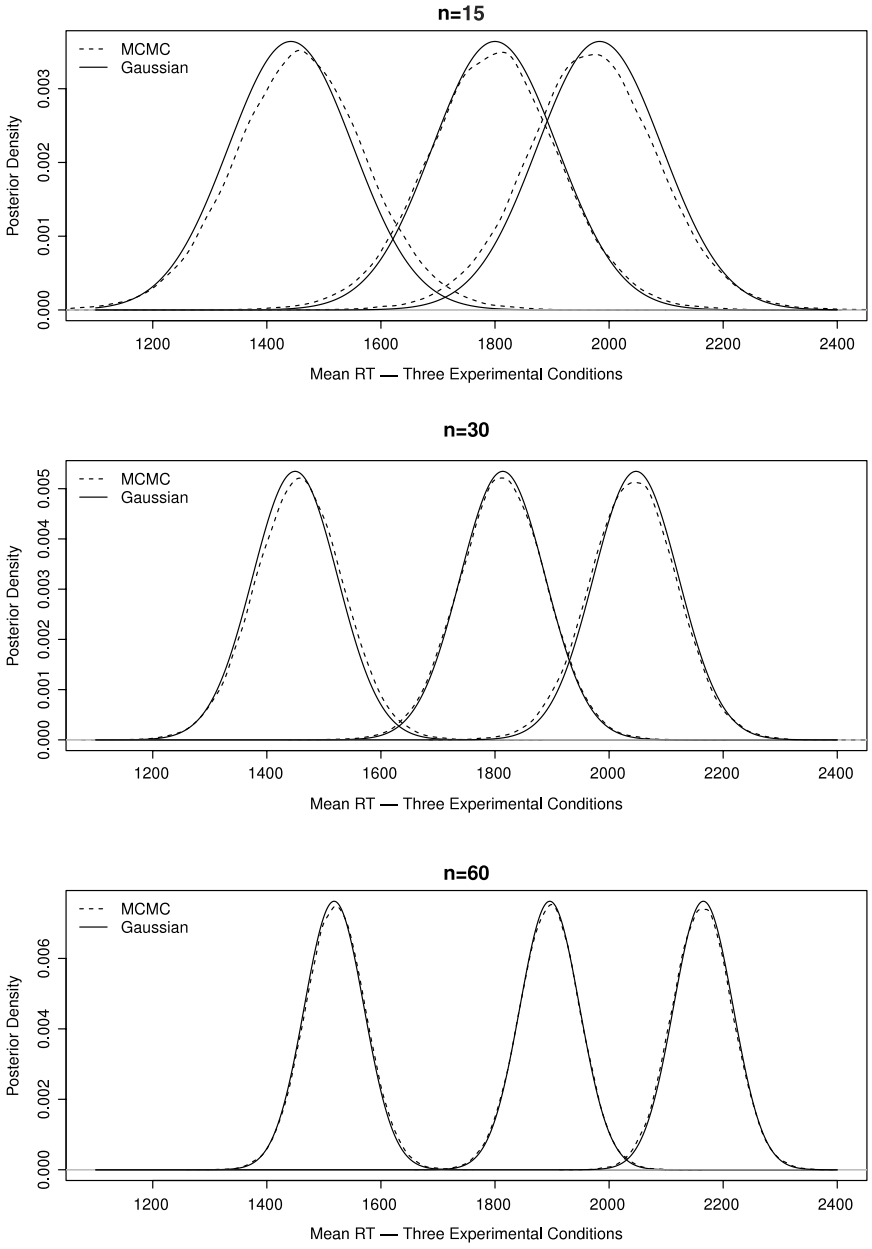

Fig. 1. Comparison of MCMC and the large-sample Gaussian approximation to the posterior densities of the condition means in a simulated experiment having three conditions. Data representing reaction times (RT) are simulated from the linear mixed model $Y_{ij} = \mu_j + b_i + \epsilon_{ij},\ i = 1, \ldots, n; j = 1, \ldots, 3$ with $b_i \overset{i.i.d.}{\sim} N(0, \sigma_b^2)$ independent of $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$. Simulated data are based on parameter values $\mu_1 = 1532$, $\mu_2 = 1832$, $\mu_3 = 2132$, $\sigma_b = 347$, $\sigma_\epsilon = 173$, and the results obtained from three datasets based on 15, 30, and 60 subjects are depicted.

As an illustration, Fig. 1 presents the posterior distributions associated with a single-factor repeated-measures design with three conditions. The posterior distributions are computed based on repeated-measures data simulated for designs having $n = 15$, $n = 30$, and $n = 60$ subjects, and in each case our approximations are compared with posterior densities obtained from MCMC sampling. Even in the case of a fairly small sample size, our approximations match those obtained from MCMC very well, and have the advantages of simplicity and a lack of dependence on any assumed prior. Despite the accuracy seen in this example, it should be noted that posterior distributions estimated using MCMC are more accurate than large sample approximations in general.

In an informal correction to his 2011 article, Wagenmakers provides on his web site (http://www.ejwagenmakers.com/2007/CorrigendumPvalues.pdf) information about how one can use the nlme function in R to compute a BIC value for any design, not just the one-factor repeated-measures design we have emphasized here. Our approach, however, has a number of advantages over using the nlme function in R, at least for the one-factor case. First, the BIC command in R uses the total number of observations as the sample size when computing the penalty term for the BIC. As we have seen, this will be overly severe in the typical case where at least a modest correlation between conditions holds. Second, using the nlme approach requires the user first to fit the null and alternative models using nlme by specifying them within R. This may not be an appealing prospect for researchers who are much more comfortable with computing standard ANOVAs. Our methodology provides a closed-form expression for BIC (and therefore the Bayes factor) based on the sums of squares available from ANOVA summary tables generated by commonly used statistical packages. Finally, the R source file that we provide uses components generated by ANOVA to produce not only

Table 2
ANOVA summary table for example data set.

Source	SS	df	MS	F	p
Subjects	16 877	11			
Conditions	3 196	2	1598.0	14.96	<0.0001
S × C	2 349	22	106.8		
Total	22 422	35			

posterior probabilities, but also posterior distributions and 95% highest posterior density intervals for condition means. These distributions cannot be obtained from the nlme method.

5. An example application

To illustrate the use of the BIC approximation method we have described, we will consider an experiment in which eye movements were monitored while subjects searched for a target object in a photograph of a scene. In an initial phase of the experiment, subjects searched one set of photographs. In the final phase, subjects searched those same scenes, but either for the same target as in the initial phase, or for a new target. In a third condition, the scenes were new. One of the dependent measures assessed the average duration of the eye fixations made during search. For a sample of 12 subjects, the mean fixation durations were 204 ms for the same-target condition, 224 for the new-target condition, and 225 for the new-scene condition. A standard repeated-measures ANOVA generated the summary table shown in Table 2. Based on the values in Table 2, we can see that the condition $kSS_s > SS_T$ holds, so we can proceed with the first case of Eq. (5), as will usually happen. In this example, $n = 12$ and $k = 3$. Applying Eq. (5), we have

$$\begin{aligned}\Delta BIC_{10} &= 12(3-1)\log\left(\frac{22422-3196-16877}{22422-16877}\right)\\ &\quad+(3+2)\log\left(\frac{12(22422-3196)}{16877}\right)\\ &\quad-3\log\left(\frac{12(22422)}{16877}\right)\\ &= 12(2)(\log(0.4236))+5(\log(13.670))3(\log(15.943))\\ &= -20.615+13.0768.307=-15.855.\end{aligned}$$

Using the equation presented above for converting ΔBIC_{10} to the Bayes Factor for expressing strength of evidence in favor of the null hypothesis relative to the hypothesis that assumes an effect is present, we have $BF_{01} = \exp(-15.855/2) = 0.00036$. This is powerful evidence in favor of the alternative model (which assumes that an effect is present). Converting the BF_{01} value to posterior probabilities, we have $p(H_0|D) = 0.00036/(1+0.00036) = 0.00036$ and $p(H_1|D) = 1-0.00036 = 0.9996$. This outcome qualifies as "very strong" evidence according to Table 1 (see Raftery, 1995). In this example, if we had used n, as proposed by Masson (2011), instead of n_{eff}, we would have obtained $p(H_1|D) = 0.9992$ (slightly more conservative than the result produced by the method we advocate here).

The posterior densities for each of the three condition means are computed based on the large sample approximation described in Section 4 and these densities are depicted in Fig. 2. These posterior densities are obtained using R along with our *rmBayes*() function. The use of this function for producing posterior probabilities, densities, and interval estimates is illustrated in Appendix C.

6. Comparison of Bayesian and frequentist methods of hypothesis testing

Now that we have established the validity of a revised method for computing a BIC approximation for Bayesian analysis of the repeated-measures design, we turn to a comparison between this, the method developed by Rouder et al. (2012), the BIEMS method of Mulder et al. (2012), and the standard NHST approach. It is instructive at this point to re-iterate the differences between the three Bayesian methods. First, the three approaches are based on different priors. The BIC is based on the unit information prior, a data-dependent multivariate normal prior, with mean equal to the maximum likelihood estimator, and variance chosen so that the information brought by the prior is equivalent to one observation. The Bayes factor of Rouder et al. (2012) is based on a modification of the Zellner and Siow g-priors for the effects in the linear model combined with a Jeffreys prior for the overall mean and error variance. BIEMS is based on the conjugate expected-constrained posterior prior, a prior that uses minimal training subsets of the data that contain enough information to obtain a proper posterior prior, and the prior is centered on the boundary of the constrained parameter space under investigation. With regard to computation of the Bayes factor, both Rouder et al. (2012) and BIEMS use MCMC, whereas the BIC is based on a large sample approximation. In general, the latter will be less accurate than the former. Although the prior distributions are different, all offer the advantages of an objectively determined prior, namely, that the researcher does not have to construct his or her own prior distribution and that different researchers applying the analysis to the same set of data will reach the same statistical conclusion if they compare the same two models.

To examine the performance of these four inference methods, we conducted a Monte Carlo simulation study in which sets of data were generated from three hypothetical, normally distributed,

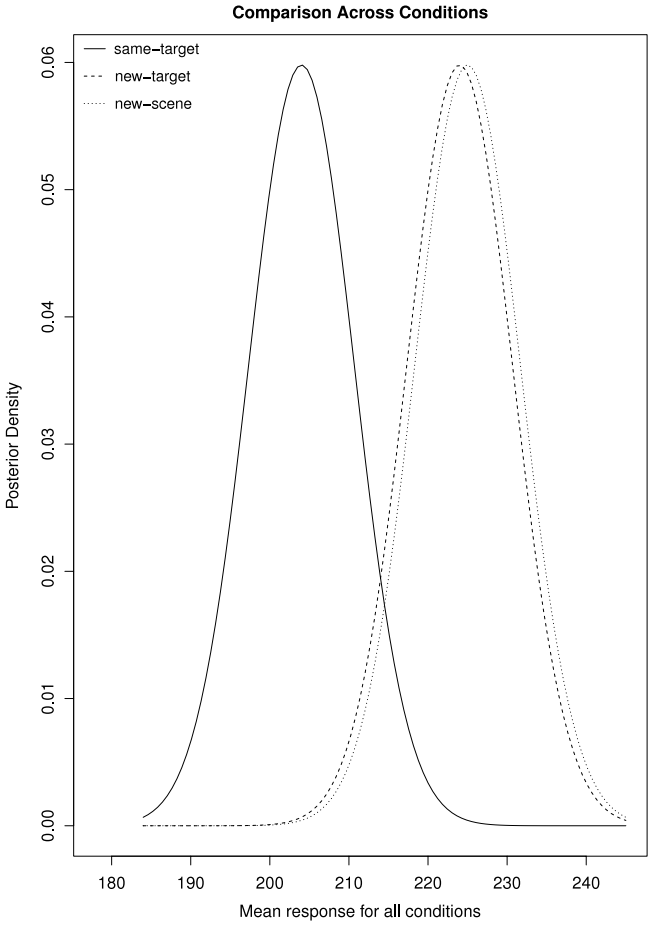

Fig. 2. Approximate posterior densities of the three condition means in the example application of Section 5. The R commands to conduct the Bayesian analysis using the *rmBayes*() function are provided in Appendix C.

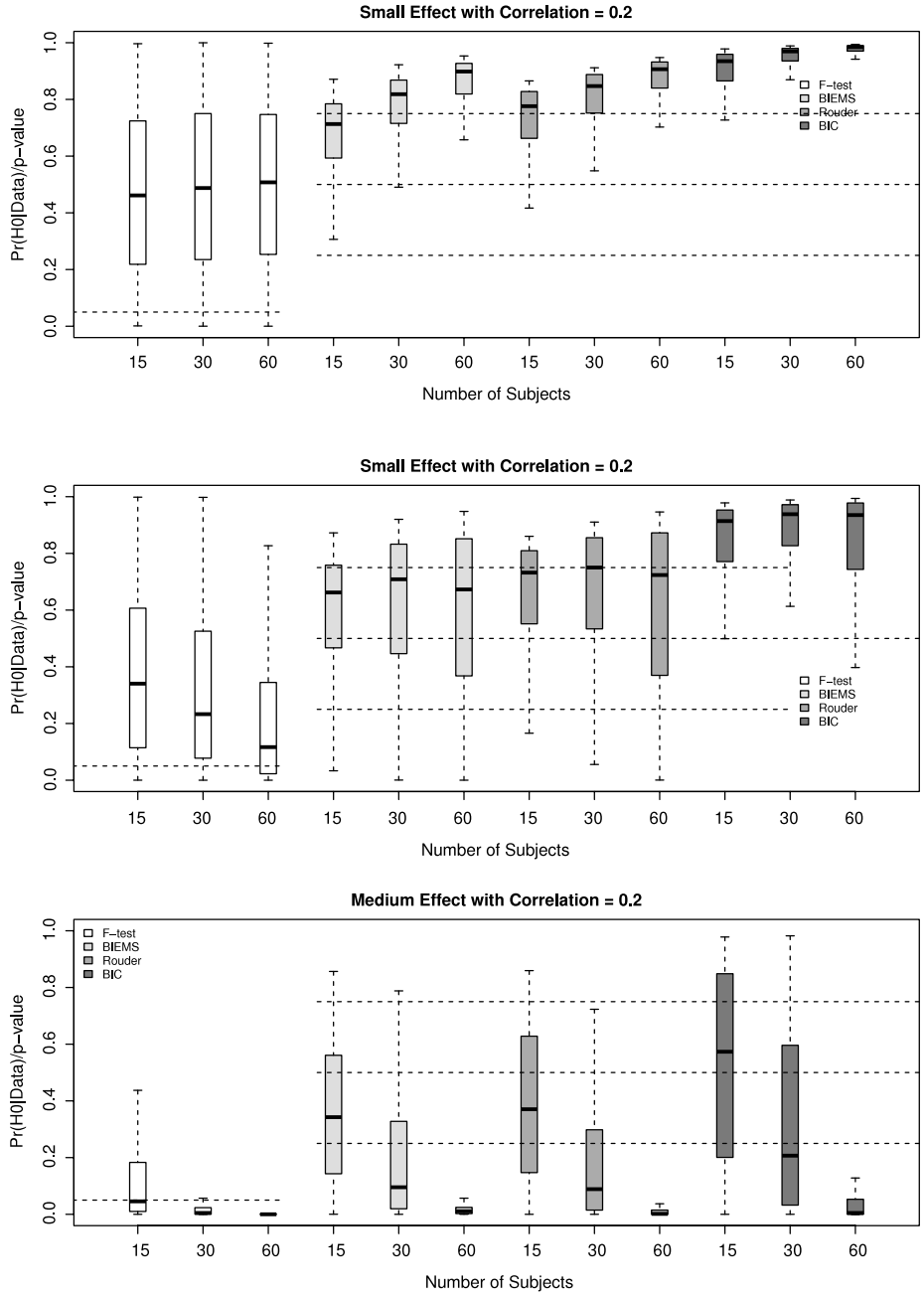

Fig. 3. Results from the simulation study with $\rho = 0.2$. Each boxplot depicts the sampling distribution of the p-value/posterior probability based on 1000 Monte Carlo replications. Horizontal dashed lines correspond to values 0.05 for the F test and 0.75, 0.5 and 0.25 for the Bayesian posterior probability.

equal variance, correlated populations with prescribed means. The data are simulated from the linear mixed model

$$Y_{ij} = \mu_j + b_i + \epsilon_{ij}, \quad i = 1, \ldots, n;\ j = 1, \ldots, k$$

where $k = 3$ and $b_i \overset{i.i.d.}{\sim} N(0, \sigma_b^2)$ independent of $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$. The populations of scores were correlated to varying degrees to allow us to simulate realistic data for a repeated-measures design with a single factor having three levels. Population means were adjusted to produce three different effect sizes: null effect (the null hypothesis model was correct and all population means were equal), small effect, and medium effect, where we define the effect size as $ES = (\mu_{largest} - \mu_{smallest})/\sqrt{(\sigma_b^2 + \sigma_\epsilon^2)}$ which can also be expressed as $ES = (\mu_{largest} - \mu_{smallest})\sqrt{1-\rho}/\sigma_\epsilon$ where $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_\epsilon^2)$ is the intraclass correlation. Within the simulation study we simulated data based on effect sizes of $ES = 0$ (null), $ES = 0.258$ (small) and $ES = 0.645$ (medium), and for each effect size we consider two values for the correlation, either $\rho = 0.2$ or $\rho = 0.8$, with the larger value corresponding to a larger value of σ_b^2 and a correspondingly smaller value of σ_ϵ^2 such that the marginal variance $\sigma_b^2 + \sigma_\epsilon^2$ remains constant under all of the simulation settings considered. Datasets consisted of 15, 30, or 60 simulated subjects, with each subject having a score in each of three conditions. We drew 1000 such datasets for each combination of effect size, sample size, and correlation between conditions, and analyzed each dataset using the three methods. The results are depicted in Figs. 3 and 4.

Our measure of how the Bayesian methods and the F test performed in the simulations consisted of assessing the sampling distribution of posterior probability values produced by each method under a particular set of circumstances (e.g., effect size, sample size, and degree of correlation between conditions). For

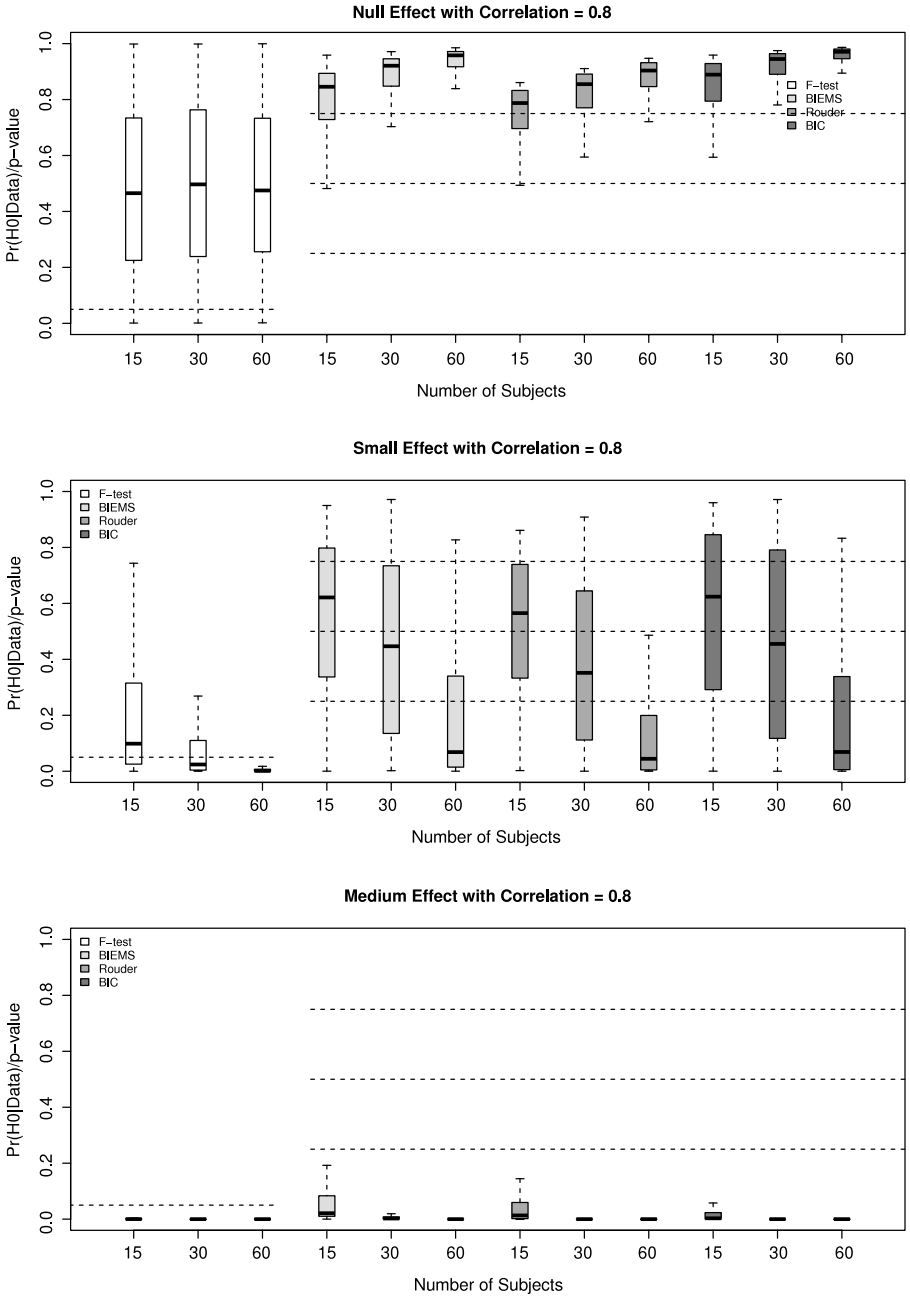

Fig. 4. Results from the simulation study with $\rho = 0.8$. Each boxplot depicts the sampling distribution of the p-value/posterior probability based on 1000 Monte Carlo replications. Horizontal dashed lines correspond to values 0.05 for the F test and 0.75, 0.5 and 0.25 for the Bayesian posterior probability.

example, when the effect size is zero (null hypothesis is true), for the researcher to avoid a decision error the Bayesian methods must produce a posterior probability value greater than 0.5, and the F test must yield a p value greater than 0.05. In Figs. 3 and 4, we show box and whisker plots for the three methods as a function of sample size, effect size, and strength of correlation between conditions. The median is represented by the thick black line inside each bar, and the bar captures the middle 50% of the distribution of outcomes. The arms extending above and below each bar include the upper and lower 25% of the distribution, respectively. The 0.05 threshold for the F test and the 0.25, 0.50, and 0.75 thresholds for the Bayesian tests are shown in the figures as horizontal lines.

When the null hypothesis was true, the probability associated with the standard F test was uniformly distributed over the range 0–1, no matter the sample size or strength of correlation. This aspect of p values generated by NHST methods is of course true by construction, but this has only recently been appreciated by researchers in behavioral sciences (Rouder, Speckman, Sun, Morey, & Iverson, 2009). In contrast to the highly variable NHST p values that are produced when the null hypothesis is true, the Bayesian approaches provided a more restricted range of posterior probability values, particularly with larger sample sizes. For all three Bayesian methods, the distributions of probabilities were negatively skewed. In all situations, the BIC method produced a posterior probability greater than 0.5 for all datasets, whereas for the Rouder et al. (2012) and BIEMS methods, a few data sets yielded posterior probabilities slightly smaller than 0.5 when the sample size was small or moderate (15 or 30). Unlike the F test p values, which held to a uniform distribution regardless of sample size or degree of correlation between conditions, in the Bayesian methods the posterior probability converged with increasing sample size to the limiting value of 1. With this approach

to 1, there was also a corresponding decrease in the variability of the probability values. Thus, with the increasing information brought about by larger samples, the Bayesian methods yielded increasingly stronger evidence for the correct decision. The NHST method cannot achieve this convergence, and regardless of the sample size it will always be liable for a type I error rate equivalent to the stated criterion for significance.

For cases in which a real effect is present, the Bayesian methods produced more variable results than when there was no effect. In comparing the behavior of the Bayesian posterior probabilities to the p values generated by the F test, we need to consider carefully how decision making may be guided in each context. We will discuss two possibilities here. In one approach, we will set a threshold for determining whether the obtained results provide clear support for one hypothesis over the other. In the second approach, we will consider the power of each method to detect an effect, but will do so by first establishing a way to equate the methods with respect to type I error (concluding that an effect is present when it is not). Note, for example, that the erstwhile 0.05 value for the NHST was established with an eye toward maintaining reasonable levels of statistical power. In our second approach, we take account of a similar trade-off for the Bayesian methods.

Using Raftery's (1995) categories of evidence strength as a guide, we can reasonably assert that a posterior probability of 0.75 stands as positive evidence in favor of one model over another. In Figs. 3 and 4, we have plotted horizontal lines corresponding to Bayesian posterior probabilities of the null hypothesis equal to 0.25 and 0.75. Results more extreme than the thresholds of 0.25 and 0.75 indicate at least positive evidence in favor of the alternative hypothesis or the null hypothesis, respectively. Probability values within these bounds indicate inconclusive evidence. For the F test, we will use the standard type I error rate of 0.05, and that threshold is indicated by a horizontal line in the plots shown in Figs. 3 and 4. Working with these thresholds, consider first the scenario in which the null hypothesis is true (no effect exists). By definition, any outcome that fails to reject the null hypothesis under the F test constitutes inconclusive evidence (95% of the cases). All three Bayesian methods provide much superior performance assuming a threshold posterior probability of the null equal to 0.75. For the BIC method, for example, inclusive outcomes hardly ever occurred once sample size reached 30. Even in the least successful case, the BIEMS method with $n = 15$, just over half of the simulated studies produced an inconclusive outcome. Moreover, whereas 5% of the studies yielded an incorrect conclusion under the F test (again, by definition), none of the results led any of the Bayesian methods to support the wrong conclusion.

When an effect is present, it is clear that the Bayesian methods were not as likely to provide clear evidence in favor of an effect as was the F test. With small effect size, particularly when the correlation between conditions was weak, the BIC method failed to detect the effect in any of the simulated studies and often resulted in evidence supporting the null hypothesis. The outcome reflects the fact that the prior used in the BIC approximation (the unit information prior) is relatively uninformative, covering a wide range of possible effect sizes. Consequently, the posterior probabilities favoring the alternative hypothesis are rather conservative in the BIC method (Wagenmakers, 2007). The Rouder and the BIEMS methods were more successful, but even they often produced evidence clearly supporting the null hypothesis when the effect size and correlation between conditions were small. A stronger correlation between conditions ameliorated this problem, although the F test was still more likely to detect an effect than were the Bayesian methods. With a medium effect, the Bayesian methods improved substantially, though still did not do as well as the F test unless sample size was at 60, in which case all methods were guaranteed to detect an effect. When a strong correlation between conditions was in place, all methods were at ceiling.

One could describe the results from the threshold approach as indicating that the F test has more power to detect effects, particularly small ones, than do Bayesian methods. That conclusion, however, overlooks the great advantage that the Bayesian methods have with respect to controlling the equivalent of type I error. Consider again the simulation results obtained when no effect was present. Whereas the type I error rate for the F test was fixed at 0.05, for the Bayesian methods, the likelihood of obtaining substantial evidence in favor of the alternative hypothesis was zero (assuming a threshold of 0.25 for concluding that results support the presence of an effect). Indeed, the threshold for the Bayesian tests could have been shifted to 0.5, and still the probability of erroneously concluding that there was evidence in favor of the alternative hypothesis would have been zero, except for the smallest sample size. If we take as a priority making the F test and Bayesian approaches comparable with respect to type I error, then there is some justification for changing the decision threshold in the Bayesian tests to 0.5. If we do that, nothing is lost with respect to the correct handling of null effect cases by the Bayesian methods (type I error rates remain nearly zero). The power of these methods to detect real effects, however, is substantially enhanced. For example, the Rouder method is now at least as powerful as NHST, even with small effect sizes.

We recognize that Bayesian methods are intended to provide a continuous measure of evidence strength rather than a thresholded test. Our purpose in working with hypothetical thresholds was to provide a basis for comparing the implications of results from NHST and Bayesian methods. Many researchers considering a shift to using Bayesian methods will be keenly interested in the prospects of finding evidence for effects. The simulation results we have presented indicate that the Bayesian approach is likely to make it more challenging to generate convincing evidence in favor of effects being present. We return to this issue and provide some suggestions regarding how to deal with it in the final section.

We have conducted additional simulation studies to examine the sensitivity of the F test and the BIC and Rouder et al. (2012) methods to model misspecification, specifically to violation of the compound symmetry assumption in the repeated-measures design. The results demonstrated that all three approaches are fairly robust to violations of this assumption.

7. Conclusions and recommendations

We have shown (see Appendix A) that the formulation of the BIC suggested by Masson (2011) for the repeated-measures design is valid, although the penalty term should be modified. The modification, recommended by Jones (2011), applies a potentially less severe penalty than did the method presented by Masson, depending on the degree of correlation between conditions. We have also supplied as an online supplement an easy-to-use routine that can be run in the freely available R statistical package. This routine takes as input sum of squares values and means from a standard ANOVA and yields a posterior probability for the null hypothesis as well as posterior probability density distributions, and 95% highest posterior density intervals for condition means.[1] The BIC method provides a full Bayesian analysis for the one-factor repeated-measures case without the need to specify a prior distribution (this is done automatically within the BIC method)

[1] The posterior density distributions and highest density intervals that our routine computes are influenced by between-subject variability. In that sense, they are analogous to standard confidence intervals rather than within-subject confidence intervals (Loftus & Masson, 1994).

or the need to fit models numerically. These features make the routine highly accessible and relatively easy to use and interpret, even for newcomers to Bayesian analyses.

A potential disadvantage of the BIC method relative to Rouder et al. (2012) is that it is somewhat conservative with respect to providing evidence in support of the alternative hypothesis. Our simulation results showed that the Rouder et al. method performed about on par with the F test, whereas the BIC method was somewhat less likely to detect effects when effects were small or correlations between conditions were weak. Wagenmakers (2007) noted that the BIC method may fail to support the alternative hypothesis in situations where a standard subjective Bayesian method using a more restrictive prior distribution than that assumed by the BIC method would do so. Wagenmakers attributed this aspect of the BIC to its reliance on the unit information prior, which is relatively uninformative and therefore tends to decrease the prior probability of the alternative hypothesis. He suggested that as a result, the BIC method could be viewed as an objective baseline reference for Bayesian testing. Although that is a reasonable option, we believe it is important to also point out that the BIC method was more decisive than the Rouder et al. method with respect to providing evidence in favor of the null hypothesis when no effects were present in the populations. This aspect of the BIC is especially noteworthy in a context in which one of the major benefits of adopting a Bayesian approach is that it provides a means of gauging the strength of evidence in favor of the null hypothesis.

The BIC has the advantage of being easy to use and for the mixed model considered here we have shown that it can be expressed entirely in terms of standard summary statistics taken from the repeated measures ANOVA table. It is thus recommended for users transitioning to the use of Bayesian methods for hypothesis testing who may not be familiar or comfortable using techniques based on MCMC sampling algorithms. More experienced users may have a preference for a particular prior and use the method corresponding to that prior. In this regard we note that Kass and Raftery (1995) stress the importance of assessing sensitivity of conclusions to the prior distribution when Bayes factors are used. Given this, we recommend that all three approaches can be used as part of a standard prior sensitivity assessment in any repeated-measures analysis.

One general observation we offer based on our simulation results pertains to the impact of sample size on the behavior of the Bayesian tests. Namely, in cases where effects are difficult to detect (i.e., small effect size and/or small correlation between conditions), increasing the sample size from 30 to 60 had a substantial impact on the consistency in the results produced by both the BIC method and the Rouder et al. (2012) method. This effect suggests that researchers should give careful consideration to building sample sizes in excess of 30 when it is anticipated that effects will be elusive. In that regard, we note that in Bayesian analysis, in contrast to NHST, it is quite legitimate to engage in optional stopping during data collection.

We are mindful of the potential reluctance of researchers to adopt a tool that has a tendency toward conservatism with respect to strength of evidence for the alternative hypothesis. There are two points we wish to make in this regard. First, Wagenmakers (2007) provided a very insightful warning about a crucial oversight associated with NHST, which emphasizes strength of evidence against the null hypothesis. He observed that "the NHST procedure is oblivious to the very real possibility that although the data may be unlikely under H_0, they are even less likely under H_1" (p. 793). For example, effects that just barely reach significance under NHST (e.g., $p = 0.04$) often are associated with Bayesian outcomes that actually favor the null hypothesis. Second, given the caution that Bayesian analysis suggests when data outcomes would barely meet the NHST threshold for rejecting the null hypothesis, we suggest that authors who are reluctant to abandon NHST p values, report both the outcome of the significance test along with the posterior probability or Bayes factor associated with the effect (e.g., $F(1, 29) = 10.77$, $MSE = 2574$, $p < 0.01$, $BF = 4.5$). Including information from a Bayesian analysis, particularly when a p value hovers near 0.05 and the Bayesian result does not clearly support the alternative hypothesis, will at least alert readers to the possibility that the reported effect may not be likely to replicate.

At this stage in the development of the BIC method, it is not clear how multifactor designs that include at least one repeated-measures factor should be handled. Masson (2011) suggested a general approach with respect to how the value of n should be interpreted in the two terms of the equation used to define ΔBIC_{10} when a repeated-measures factor is involved, namely, $n = s(k-1)$, where s is the number of subjects and k is the number of levels of the repeated-measures variable. For hypotheses involving the interaction between multiple repeated-measures factors, the term $(k-1)$ would be replaced by the degrees of freedom for the interaction. This suggestion has not been verified and we plan to extend our examination of the BIC method to determine whether a closed form solution for such designs is possible and, if so, what the correct definition for n would be in such cases. In the interim, it likely is advisable to consider using the Rouder et al. (2012) method at least for factorial designs. Its implementation in R makes it readily accessible, and the associated manual should make this an attractive option even for novice users of Bayesian methods (http://cran.r-project.org/web/packages/BayesFactor/index.html). For simple t test designs, the method described by Kruschke (2013) continues to be a very attractive option (http://www.indiana.edu/kruschke/BEST/) as it provides a complete Bayesian analysis including posterior density distributions and highest density intervals.

In this paper we have focused on a standard fairly simple model for the analysis of repeated-measures data that assumes compound symmetry and multivariate normality. Extensions to alternative more flexible models for repeated measures can be considered by relaxing these assumptions. In such settings the utility of this additionally flexibility may be evaluated through the use of Bayes factors. Finally, in our view, Bayesian analysis of ANOVA designs can be profitably supplemented by graphical presentation of data that includes condition means and associated confidence intervals. There is a growing literature on computing confidence intervals for designs that include repeated-measures factors that provides good support for this enterprise (e.g., Franz & Loftus, 2012; Hollands & Jarmasz, 2010; Loftus & Masson, 1994 and Masson & Loftus, 2003). In keeping with this tool, we are currently working on a supplement to our R routine that will generate a set of modified posterior probability distributions for condition means in a one-factor repeated-measures design that is conditionalized on between-subject variability. These modified probability density distributions will be sensitive to the consistency of effects across subjects and will ignore between-subject differences, in much the same way that within-subjects confidence intervals do (Loftus & Masson, 1994). If this enterprise is successful, it would offer a further alternative for the graphical presentation of empirical results.

Acknowledgments

This work was supported by discovery grants to Farouk Nathoo and Michael Masson from the Natural Sciences and Engineering Research Council of Canada (grant numbers 6542 and 7910). Farouk Nathoo holds a Tier II Canada Research Chair in Biostatistics.

Appendix A. ΔBIC for single factor repeated-measures ANOVA

Here we derive an analytic form for $\Delta BIC = BIC(H_1) - BIC(H_0)$ for the single factor repeated-measures ANOVA based on the linear mixed model

$$Y_{ij} = \mu_j + b_i + \epsilon_{ij}, \quad i = 1, \ldots, n;\ j = 1, \ldots, k, \tag{A.1}$$

where $b_i \overset{i.i.d.}{\sim} N(0, \sigma_b^2)$ independent of $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$, and where interest lies in the test of hypotheses $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$ versus $H_1 : \mu_j$ *not all equal*. For each of the models corresponding to the null and alternative hypothesis we derive the maximum likelihood estimators and use these to derive an analytic expression for ΔBIC in terms of standard output arising from a repeated measures ANOVA table. Under the full model (A.1) we have $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ik})' \overset{i.i.d.}{\sim} MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ and $\boldsymbol{\Sigma}$ exhibits compound symmetry with

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\epsilon^2 + \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \ddots & & \vdots \\ \vdots & & & \\ \sigma_b^2 & \cdots & & \sigma_\epsilon^2 + \sigma_b^2 \end{pmatrix}. \tag{A.2}$$

Under compound symmetry (A.2) it can be shown (see e.g. Jones, 1993) that

$$|\boldsymbol{\Sigma}| = (k\sigma_b^2 + \sigma_\epsilon^2)(\sigma_\epsilon^2)^{k-1} \tag{A.3}$$

and that

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_\epsilon^2}\left[\boldsymbol{I}_k - \frac{\sigma_b^2 \mathbf{1}_k \mathbf{1}_k'}{k\sigma_b^2 + \sigma_\epsilon^2}\right] \tag{A.4}$$

where $\boldsymbol{I}_k$ is the identity matrix of dimension k, and $\mathbf{1}_k$ is a k-vector of 1's. The probability density function of $\boldsymbol{Y}_i$ is given by $f(\boldsymbol{Y}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-k/2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu})\}$ so that the log likelihood is $l(\boldsymbol{\mu}, \sigma_b^2, \sigma_\epsilon^2) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^n (\boldsymbol{Y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu})$ (where log refers to the natural logarithm) and the MLE's are obtained by solving the first-order conditions

$$\frac{\partial l}{\partial \boldsymbol{\mu}} = \mathbf{0}, \qquad \frac{\partial l}{\partial \sigma_b^2} = 0, \qquad \frac{\partial l}{\partial \sigma_\epsilon^2} = 0.$$

Beginning with $\frac{\partial l}{\partial \boldsymbol{\mu}} = \sum_{i=1}^n \boldsymbol{Y}_i'\boldsymbol{\Sigma}^{-1} - n\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1} = \mathbf{0}$ and solving yields $\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{Y}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{Y}_i$. Substituting this along with (A.3) and (A.4) back into the log likelihood and simplifying yields $l = -\frac{n}{2}\log(k\sigma_b^2 + \sigma_\epsilon^2) - \frac{n(k-1)}{2}\log(\sigma_\epsilon^2) - \frac{(SS_T - SS_C)}{2\sigma_\epsilon^2} + \frac{\sigma_b^2 kSS_S}{2\sigma_\epsilon^2(k\sigma_b^2 + \sigma_\epsilon^2)}$ where $SS_T = \sum_{i=1}^n \sum_{j=1}^k (Y_{ij} - \overline{Y}_{..})^2$, $SS_S = k\sum_{i=1}^n (\overline{Y}_{i.} - \overline{Y}_{..})^2$, $SS_C = n\sum_{j=1}^k (\overline{Y}_{.j} - \overline{Y}_{..})^2$, $\overline{Y}_{..} = \frac{1}{nk}\sum_{i=1}^n \sum_{j=1}^k Y_{ij}$, $\overline{Y}_{i.} = \frac{1}{k}\sum_{j=1}^k Y_{ij}$, and $\overline{Y}_{.j} = \frac{1}{n}\sum_{i=1}^n Y_{ij}$. Taking the derivative WRT σ_b^2 yields

$$\frac{\partial l}{\partial \sigma_b^2} = \frac{-nk}{2(k\sigma_b^2 + \sigma_\epsilon^2)} + \frac{kSS_S}{2\sigma_\epsilon^2(k\sigma_b^2 + \sigma_\epsilon^2)} - \frac{k^2\sigma_b^2 SS_S}{2\sigma_\epsilon^2(k\sigma_b^2 + \sigma_\epsilon^2)^2}$$

and simplifying the corresponding first-order condition yields

$$k\sigma_b^2 + \sigma_\epsilon^2 = SS_S/n. \tag{A.5}$$

Taking the derivative WRT σ_ϵ^2 yields

$$\frac{\partial l}{\partial \sigma_\epsilon^2} = -\frac{n}{2(k\sigma_b^2 + \sigma_\epsilon^2)} - \frac{n(k-1)}{2\sigma_\epsilon^2} + \frac{SS_T - SS_C}{2(\sigma_\epsilon^2)^2} - \frac{\sigma_b^2 kSS_s(k\sigma_b^2 + 2\sigma_\epsilon^2)}{2(\sigma_\epsilon^2)^2(k\sigma_b^2 + \sigma_\epsilon^2)^2}$$

and substituting (A.5) along with solving the corresponding first-order condition yields $\hat{\sigma}_\epsilon^2 = (SS_T - SS_C - SS_S)/n(k-1)$ and expression (A.5) then yields $\hat{\sigma}_b^2 = (kSS_S - SS_T + SS_C)/nk(k-1)$ *provided* $kSS_S - SS_T + SS_C > 0$, otherwise $\hat{\sigma}_b^2 = 0$. In the latter case the MLE for σ_ϵ^2 is obtained from solving the first-order condition on the boundary

$$-\frac{n}{2\sigma_\epsilon^2} - \frac{n(k-1)}{2\sigma_\epsilon^2} + \frac{SS_T - SS_C}{2(\sigma_\epsilon^2)^2} = 0$$

which yields $\hat{\sigma}_\epsilon^2 = (SS_T - SS_C)/nk$. The MLE's for the variance components in the mixed model are thus given by

$$\hat{\sigma}_b^2 = \begin{cases} \dfrac{(kSS_S - SS_T + SS_C)}{nk(k-1)}, \\ \quad \text{if } kSS_S - SS_T + SS_C > 0 \\ 0, \quad \text{otherwise} \end{cases}$$

$$\hat{\sigma}_\epsilon^2 = \begin{cases} \dfrac{(SS_T - SS_C - SS_S)}{n(k-1)}, \\ \quad \text{if } kSS_S - SS_T + SS_C > 0 \\ \dfrac{SS_T - SS_C}{nk}, \quad \text{otherwise} \end{cases}$$

and upon substitution into the log likelihood the value of minus twice the maximized log likelihood takes the form

$$-2\hat{l} = \begin{cases} n\log\left(\dfrac{SS_S}{n}\right) + n(k-1)\log\left(\dfrac{SS_T - SS_C - SS_S}{n(k-1)}\right) \\ \quad + nk, \quad \text{if } kSS_S - SS_T + SS_C > 0 \\ nk\log\left(\dfrac{SS_T - SS_C}{nk}\right) + nk, \quad \text{otherwise.} \end{cases}$$

The expression for $BIC(H_1)$ is then obtained as $BIC(H_1) = -2\hat{l} + (k+2)\log(n_{eff})$, where $k+2$ denotes the number of parameters under H_1 (namely, the k population means, and the two variance parameters, $\sigma_b^2, \sigma_\epsilon^2$) and n_{eff} is the effective sample size accounting for repeated measures correlation. Jones (2011) derives an expression for n_{eff} for the model (A.1) based on the Fisher Information as $n_{eff} = \frac{nk}{1+(k-1)\rho}$ where $\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2}$ and suggests that an estimate of ρ be used to determine n_{eff} in practice. We use the MLE for ρ and obtain

$$n_{eff} = \begin{cases} \dfrac{n(SS_T - SS_C)}{SS_S}, & \text{if } kSS_S - SS_T + SS_C > 0 \\ nk, & \text{otherwise.} \end{cases}$$

The expression for $BIC(H_1)$ is then

$$BIC(H_1) = \begin{cases} n\log\left(\dfrac{SS_S}{n}\right) + n(k-1) \\ \quad \times \log\left(\dfrac{SS_T - SS_C - SS_S}{n(k-1)}\right) + nk \\ \quad + (k+2)\log\left(\dfrac{n(SS_T - SS_C)}{SS_S}\right), \\ \quad \text{if } kSS_S - SS_T + SS_C > 0 \\ nk\log\left(\dfrac{SS_T - SS_C}{nk}\right) + nk \\ \quad + (k+2)\log(nk), \quad \text{otherwise.} \end{cases} \tag{A.6}$$

Under the null model $Y_{ij} = \mu + b_i + \epsilon_{ij}$ which has three unknown parameters μ, σ_b^2, σ_ϵ^2, we follow the same steps and obtain expressions for n_{eff} and $BIC(H_0)$ as

$$n_{eff} = \begin{cases} \dfrac{nSS_T}{SS_S}, & \text{if } kSS_S - SS_T > 0 \\ nk, & \text{otherwise} \end{cases}$$

$$BIC(H_0) = \begin{cases} n\log\left(\frac{SS_S}{n}\right) + n(k-1)\log\left(\frac{SS_T - SS_S}{n(k-1)}\right) \\ \quad + nk + 3\log\left(\frac{nSS_T}{SS_S}\right), \\ \quad \text{if } kSS_S - SS_T > 0 \\ nk\log\left(\frac{SS_T}{nk}\right) + nk + 3\log(nk), \quad \text{otherwise.} \end{cases} \tag{A.7}$$

We thus obtain an expression for $\Delta BIC_{10} = BIC(H_1) - BIC(H_0)$ in terms of the standard statistics used in repeated measures ANOVA as

$$\Delta BIC_{10} = \begin{cases} n(k-1)\log\left(\frac{SS_T - SS_C - SS_S}{SS_T - SS_S}\right) \\ \quad + (k+2)\log\left(\frac{n(SS_T - SS_C)}{SS_S}\right) \\ \quad - 3\log\left(\frac{nSS_T}{SS_S}\right), \quad \text{if } kSS_S > SS_T \\ n\log\left(\frac{SS_S}{n}\right) + n(k-1) \\ \quad \times \log\left(\frac{SS_T - SS_C - SS_S}{n(k-1)}\right) \\ \quad - nk\log\left(\frac{SS_T}{nk}\right) - 3\log(nk) \\ \quad + (k+2)\log\left(\frac{n(SS_T - SS_C)}{SS_S}\right), \\ \quad \text{if } SS_T - SS_C < kSS_S \le SS_T \\ nk\log\left(\frac{SS_T - SS_C}{SS_T}\right) + (k-1)\log(nk), \\ \quad \text{if } kSS_S \le SS_T - SS_C. \end{cases} \tag{A.8}$$

The Bayes factor is then approximated as $BF \approx \exp(\Delta BIC_{10}/2)$ and from this the posterior probability is obtained as $Pr(H_0|\,Data) = \frac{BF}{BF+1}$.

Appendix B. Approximate posterior distribution for condition means

For the mixed model (A.1) we develop, as an alternative to MCMC sampling, a large sample approximation to the marginal posterior distribution of the condition means $p(\boldsymbol{\mu}|\,Data)$ that can be expressed simply in terms of basic statistics arising in a standard repeated measures ANOVA, and being a large sample approximation, obviates the need to consider priors for parameters $\boldsymbol{\mu}, \sigma_b^2, \sigma_\epsilon^2$. The approximation is based on the Bayesian Central Limit Theorem (see e.g. Carlin & Louis, 1996, pp. 142–145), which adapted to the current setting yields

$$\boldsymbol{\mu}|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \overset{approx}{\sim} MVN(\hat{\boldsymbol{\mu}}, [\hat{\boldsymbol{I}}(\boldsymbol{Y})]^{-1}_{\mu,\mu})$$

where $\hat{\boldsymbol{\mu}}$ is the MLE for $\boldsymbol{\mu}$, $[\hat{\boldsymbol{I}}(\boldsymbol{Y})]^{-1}$ is the inverse of the observed Fisher Information matrix, and $[\hat{\boldsymbol{I}}(\boldsymbol{Y})]^{-1}_{\mu,\mu}$ is the $k \times k$ block of this matrix corresponding to $\boldsymbol{\mu}$. We have shown in Appendix A that $\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{Y}}$, so it remains to determine $[\hat{\boldsymbol{I}}(\boldsymbol{Y})]^{-1}_{\mu,\mu}$. We begin by noting that

$$\hat{\boldsymbol{I}}(\boldsymbol{Y}) = -\begin{pmatrix} \frac{\partial^2 l}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}} & \frac{\partial^2 l}{\partial\boldsymbol{\mu}\partial\boldsymbol{\sigma}^2} \\ \frac{\partial^2 l}{\partial\boldsymbol{\sigma}^2\partial\boldsymbol{\mu}} & \frac{\partial^2 l}{\partial\boldsymbol{\sigma}^2\partial\boldsymbol{\sigma}^2} \end{pmatrix}_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}},\quad \boldsymbol{\sigma}^2=\hat{\boldsymbol{\sigma}}^2}$$

where $\boldsymbol{\sigma}^2 = (\sigma_b^2, \sigma_\epsilon^2)'$ and we write this more conveniently as

$$\hat{\boldsymbol{I}}(\boldsymbol{Y}) = -\begin{pmatrix} \frac{\partial^2 \hat{l}}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}} & \frac{\partial^2 \hat{l}}{\partial\boldsymbol{\mu}\partial\boldsymbol{\sigma}^2} \\ \frac{\partial^2 \hat{l}}{\partial\boldsymbol{\sigma}^2\partial\boldsymbol{\mu}} & \frac{\partial^2 \hat{l}}{\partial\boldsymbol{\sigma}^2\partial\boldsymbol{\sigma}^2} \end{pmatrix}.$$

Beginning with $\frac{\partial l}{\partial\boldsymbol{\mu}} = \sum_{i=1}^n \boldsymbol{Y}_i'\boldsymbol{\Sigma}^{-1} - n\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}$ and differentiating WRT $\boldsymbol{\mu}$ we obtain $\frac{\partial^2 l}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}} = n\boldsymbol{\Sigma}^{-1} \to \frac{\partial^2 \hat{l}}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}} = n\hat{\boldsymbol{\Sigma}}^{-1}$ where $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}_\epsilon^2\boldsymbol{I}_k + \hat{\sigma}_b^2\boldsymbol{1}_k\boldsymbol{1}_k'$. Using (A.4) we can write

$$\frac{\partial l}{\partial\boldsymbol{\mu}} = \left(\sum_{i=1}^n \boldsymbol{Y}_i' - n\boldsymbol{\mu}'\right)\sigma_\epsilon^{-2}\left[\boldsymbol{I}_k - \frac{\sigma_b^2\boldsymbol{1}_k\boldsymbol{1}_k'}{k\sigma_b^2 + \sigma_\epsilon^2}\right]$$

so that $\frac{\partial^2 l}{\partial\sigma_b^2\partial\boldsymbol{\mu}} = -(\sum_{i=1}^n \boldsymbol{Y}_i' - n\boldsymbol{\mu}')(k\sigma_b^2 + \sigma_\epsilon^2)^{-2} \to \frac{\partial^2 \hat{l}}{\partial\sigma_b^2\partial\boldsymbol{\mu}} = -(\sum_{i=1}^n \boldsymbol{Y}_i' - \sum_{i=1}^n \boldsymbol{Y}_i')(k\hat{\sigma}_b^2 + \hat{\sigma}_\epsilon^2)^{-2} = \boldsymbol{0}$, and similarly $\frac{\partial^2 \hat{l}}{\partial\sigma_\epsilon^2\partial\boldsymbol{\mu}} = \boldsymbol{0}$ so that $\frac{\partial^2 \hat{l}}{\partial\boldsymbol{\sigma}^2\partial\boldsymbol{\mu}} = \boldsymbol{0}$ which yields

$$\boldsymbol{I}(\hat{\boldsymbol{Y}}) = \begin{pmatrix} n\hat{\boldsymbol{\Sigma}}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & -\frac{\partial^2 \hat{l}}{\partial\boldsymbol{\sigma}^2\partial\boldsymbol{\sigma}^2} \end{pmatrix} \to \boldsymbol{I}(\hat{\boldsymbol{Y}})^{-1} = \begin{pmatrix} \frac{1}{n}\hat{\boldsymbol{\Sigma}} & \boldsymbol{0} \\ \boldsymbol{0} & -\left[\frac{\partial^2 \hat{l}}{\partial\boldsymbol{\sigma}^2\partial\boldsymbol{\sigma}^2}\right]^{-1} \end{pmatrix}$$

so we then have $[\hat{\boldsymbol{I}}(\boldsymbol{Y})]^{-1}_{\mu,\mu} = \frac{1}{n}\hat{\boldsymbol{\Sigma}}$.

The large sample Gaussian approximation to the posterior distribution of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ is thus $\boldsymbol{\mu}|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \overset{approx}{\sim} MVN(\overline{\boldsymbol{Y}}, \frac{1}{n}\hat{\boldsymbol{\Sigma}})$ where

$$\hat{\boldsymbol{\Sigma}} = \begin{cases} \frac{(SS_T - SS_C - SS_S)}{n(k-1)}\boldsymbol{I}_k + \frac{(kSS_S - SS_T + SS_C)}{nk(k-1)}\boldsymbol{1}_k\boldsymbol{1}_k', \\ \quad \text{if } kSS_S - SS_T + SS_C > 0 \\ \frac{(SS_T - SS_C)}{nk}\boldsymbol{I}_k, \quad \text{otherwise.} \end{cases}$$

Appendix C. Example use of *rmBayes*() R software

```
> #Please insert the correct path for your system depending on where
you have saved the file rmBayes.R
> path<-'/Users/farouknathoo/research13/BIC_study/rmBayes.R'
> #load software
> source(path)
> #Values for sample means, sums-of-squares and n are set as in Section 5
of the paper
> rmBayes(Ymean = c(204, 224, 225), SS.T=22422 , SS.C=3196, SS.S= 16877,
names=c('same-target','new-target','new-scene'),n=12)

BAYESIAN SINGLE FACTOR REPEATED MEASURES ANALYSIS

The effective sample size accounting for repeated measures correlation
is n.eff = 13.67

The Null Hypotheis is H0: mu1=mu2=...= muk (condition means all equal)

The posterior probability of H0 is Pr(H0|Data) = 0.000362424

95% HDI posterior intervals for each condition mean:
  Condition    HDI.Lower95 HDI.Upper95
1 same-target  190.9245    217.0755
2 new-target   210.9245    237.0755
3  new-scene   211.9245    238.0755
```

Appendix D. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmp.2015.03.003.

References

Armitage, P., McPherson, C., & Rowe, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, *132*, 235–244.

Berger, J., Bayarri, M., & Pericchi, L. (2014). The effective sample size. *Econometric Reviews*, *33*(1–4), 197–217.

Berger, J. O., & Berry, D. A. (1988). The relevance of stopping rules in statistical inference. In *Statistical decision theory and related topics IV, Vol. 1* (pp. 29–47).

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*(433), 109–122.

Bortolussi, M., & Dixon, P. (2003). *Psychonarratology: foundations for the empirical study of literary response*. Cambridge University Press.

Campbell, J. I., & Thompson, V. A. (2012). Morepower 6.0 for anova with relational confidence intervals and Bayesian analysis. *Behavior Research Methods*, *44*(4), 1255–1265.

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis*. CRC Press.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, *49*(12), 997–1003.

Cumming, G. (2014). The new statistics why and how. *Psychological Science*, *25*, 7–29.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193.

Franz, V. H., & Loftus, G. R. (2012). Standard errors and confidence intervals in within-subjects designs: generalizing loftus and masson (1994) and avoiding the biases of alternative accounts. *Psychonomic Bulletin & Review*, *19*(3), 395–404.

Glover, S., & Dixon, P. (2004). Likelihood ratios: a simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*(5), 791–806.

Hollands, J. G., & Jarmasz, J. (2010). Revisiting confidence intervals for repeated measures designs. *Psychonomic Bulletin & Review*, *17*(1), 135–138.

Jones, R. H. (1993). *Longitudinal data with serial correlation: a state-space approach. Vol. 47*. CRC Press.

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, *30*(25), 3050–3056.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, *90*(431), 928–934.

Kruschke, J. (2011). *Doing Bayesian data analysis: a tutorial introduction with R*. Academic Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*(2), 573.

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490.

Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*(3), 679–690.

Masson, M. E., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*(3), 203.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science a comment on Cumming (2014). *Psychological Science*, *25*, 1289–1290.

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). Biems: a fortran90 program for calculating bayes factors for inequality and equality constrained models. *Journal of Statistical Software*, *46*(2), 1–39.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–164.

Raftery, A. E. (1999). Bayes factors and BIC. *Sociological Methods & Research*, *27*(3), 411–417.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: applications to developmental research. *Child Development*, *85*(3), 842–860.

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

Wilkinson, L. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, *54*(8), 594–604.