Metodološki zvezki, Vol. 17, No. 1, 2020, 1–17

Estimating Bayes factors from minimal summary statistics in repeated measures analysis of variance designs

Thomas J. Faulkenberry[1]

Abstract

In this paper, I develop a formula for estimating Bayes factors directly from minimal summary statistics produced in repeated measures analysis of variance designs. The formula, which requires knowing only the F-statistic, the number of subjects, and the number of repeated measurements per subject, is based on the BIC approximation of the Bayes factor, a common default method for Bayesian computation with linear models. In addition to providing computational examples, I report a simulation study in which I demonstrate that the formula compares favorably to a recently developed, more complex method that accounts for correlation between repeated measurements. The minimal BIC method provides a simple way for researchers to estimate Bayes factors from a minimal set of summary statistics, giving users a powerful index for estimating the evidential value of not only their own data, but also the data reported in published studies.

1 Introduction

In this paper, I discuss how to apply the BIC approximation (Kass and Raftery, 1995; Wagenmakers, 2007; Masson, 2011; Nathoo and Masson, 2016) to compute Bayes factors for repeated measures experiments using only minimal summary statistics from the analysis of variance (e.g., Ly et al., 2018; Faulkenberry, 2018). Critically, I develop a formula (Equation 3.1) that works for repeated measures experiments. Further, I investigate its performance against a method of Nathoo and Masson (2016) which accounts for varying levels of correlation between repeated measurements. Among several "default prior" solutions to computing Bayes factors for common experimental designs (Rouder et al., 2009, 2012), each of which requires raw data for computation, the proposed formula stands out for providing the user with a simple expression for the Bayes factor that can be computed even when only the summary statistics are known. Thus, equipped with only a hand calculator, one can immediately estimate a Bayes factor for many results reported in published paper (even null effects), providing a meta-analytic tool that can be quite useful when trying to establish the evidential value of a collection of published results.

[1]Department of Psychological Sciences, Tarleton State University, Stephenville, TX, USA; faulkenberry@tarleton.edu

2 Background

To begin, let us consider the elementary case of a one-factor independent groups design. Consider a set of data y_{ij}, on which we impose the linear model

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}; \quad i = 1, \cdots, n;\ j = 1, \ldots, k$$

where μ represents the grand mean, α_j represents the treatment effect associated with group j, and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. In all, we have $N = nk$ independent observations. To proceed with hypothesis testing, we define two competing models:

$$\mathcal{H}_0 : \alpha_j = 0 \text{ for } j = 1, \ldots, k$$
$$\mathcal{H}_1 : \alpha_j \neq 0 \text{ for some } j$$

Classically, model selection is performed using the analysis of variance (ANOVA), introduced in the 1920s by Sir Ronald Fisher (Fisher, 1925). Roughly, ANOVA works by partitioning the total variance in the data $\boldsymbol{y}$ into two sources – the variance between the treatment groups, and the residual variance that is left over after accounting for this treatment variability. Then, one calculates an F statistic, defined as the ratio of the between-groups variance to the residual variance. Inference is then performed by quantifying the likelihood of the observed data $\boldsymbol{y}$ under the null hypothesis $\mathcal{H}_0$. Specifically, this is done by computing the probability of obtaining the observed F statistic (or greater) under $\mathcal{H}_0$. If this probability, called the p-value, is small, this indicates that the data $\boldsymbol{y}$ are *rare* under $\mathcal{H}_0$, so the researcher may reject $\mathcal{H}_0$ in favor of the alternative hypothesis $\mathcal{H}_1$. Though it is a classic procedure, some issues arise that make it problematic. First, the p-value is not equivalent to the posterior probability $p(\mathcal{H}_0 \mid \boldsymbol{y})$. Despite this distinction, many researchers incorrectly believe that a p-value directly indexes the probability that $\mathcal{H}_0$ is true (Gigerenzer, 2004), and thus take a small p-value to represent evidence for $\mathcal{H}_1$. However, Berger and Sellke (1987) demonstrated that p-values classically overestimate this evidence. For example, with a t-test performed on a sample size of 100, a p-value of 0.05 transforms to $p(\mathcal{H}_0 \mid \boldsymbol{y}) = 0.52$ – rather than reflecting evidence for $\mathcal{H}_1$, this small p-value reflects data that slightly prefers $\mathcal{H}_0$. Second, the "evidence" provided for $\mathcal{H}_1$ via the p-value is only indirect, as the p-value only measures the predictive adequacy of $\mathcal{H}_0$; the p-value procedure makes no such measurement of predictive adequacy for $\mathcal{H}_1$.

For these reasons, I will consider a Bayesian approach to the problem of model selection. The approach I will describe in this paper is to compute the Bayes factor (Kass and Raftery, 1995), denoted BF_{01}, for $\mathcal{H}_0$ over $\mathcal{H}_1$. In general, the Bayes factor is defined as the ratio of marginal likelihoods for $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. That is,

$$\text{BF}_{01} = \frac{p(\boldsymbol{y} \mid \mathcal{H}_0)}{p(\boldsymbol{y} \mid \mathcal{H}_1)}. \tag{2.1}$$

This ratio is immediately useful in two ways. First, it indexes the relative likelihood of observing data $\boldsymbol{y}$ under $\mathcal{H}_0$ compared to $\mathcal{H}_1$, so $\text{BF}_{01} > 1$ is taken as evidence for $\mathcal{H}_0$ over $\mathcal{H}_1$. Similarly, $\text{BF}_{01} < 1$ is taken as evidence for $\mathcal{H}_1$. Second, the Bayes factor indicates the extent to which the prior odds for $\mathcal{H}_0$ over $\mathcal{H}_1$ are updated after observing data. Said

differently, the ratio of posterior probabilities for $\mathcal{H}_0$ and $\mathcal{H}_1$ can be found by multiplying the ratio of prior probabilities by BF_{01} (a fact which follows easily from Bayes' theorem):

$$\frac{p(\mathcal{H}_0 \mid \boldsymbol{y})}{p(\mathcal{H}_1 \mid \boldsymbol{y})} = \text{BF}_{01} \cdot \frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}. \tag{2.2}$$

One interesting consequence of Equation 2.2 is that we can use the Bayes factor to compute the posterior probability of $\mathcal{H}_0$ as a function of the prior model probabilities. To see this, consider the following. If we solve Equation 2.2 for the posterior probability $p(\mathcal{H}_0 \mid \boldsymbol{y})$ and then use Bayes' theorem, we see

$$\begin{aligned} p(\mathcal{H}_0 \mid \boldsymbol{y}) &= \text{BF}_{01} \cdot \frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)} \cdot p(\mathcal{H}_1 \mid \boldsymbol{y}) \\ &= \frac{\text{BF}_{01} \cdot p(\mathcal{H}_0) \cdot p(\boldsymbol{y} \mid \mathcal{H}_1) \cdot p(\mathcal{H}_1)}{p(\mathcal{H}_1) \cdot p(\boldsymbol{y})} \\ &= \frac{\text{BF}_{01} \cdot p(\mathcal{H}_0) \cdot p(\boldsymbol{y} \mid \mathcal{H}_1)}{p(\boldsymbol{y} \mid \mathcal{H}_0) \cdot p(\mathcal{H}_0) + p(\boldsymbol{y} \mid \mathcal{H}_1) \cdot p(\mathcal{H}_1)}. \end{aligned}$$

Dividing both numerator and denominator by the marginal likelihood $p(\boldsymbol{y} \mid \mathcal{H}_1)$ gives us

$$p(\mathcal{H}_0 \mid \boldsymbol{y}) = \frac{\text{BF}_{01} \cdot p(\mathcal{H}_0)}{\text{BF}_{01} \cdot p(\mathcal{H}_0) + p(\mathcal{H}_1)}.$$

By Equation 2.1, we have $\text{BF}_{10} = 1/\text{BF}_{01}$. It can then be shown similarly that

$$p(\mathcal{H}_1 \mid \boldsymbol{y}) = \frac{\text{BF}_{10} \cdot p(\mathcal{H}_1)}{\text{BF}_{10} \cdot p(\mathcal{H}_1) + p(\mathcal{H}_0)}.$$

In practice, researchers often assume both models are *a priori* equally likely, and thus set both $p(\mathcal{H}_0) = p(\mathcal{H}_1) = 0.5$. In this case, we obtain the simplified forms

$$p(\mathcal{H}_0 \mid \boldsymbol{y}) = \frac{\text{BF}_{01}}{\text{BF}_{01} + 1}, \qquad p(\mathcal{H}_1 \mid \boldsymbol{y}) = \frac{\text{BF}_{10}}{\text{BF}_{10} + 1}. \tag{2.3}$$

Though there are many simple quantities that can be derived from the Bayes factor, the actual computation of BF_{01} can be quite difficult, as the marginal likelihoods in Equation 2.1 each require integrating over a prior distribution of model parameters. This often results in integrals that do not admit closed form solutions, requiring approximate techniques to estimate the Bayes factor. In Faulkenberry (2018), it was shown that for an independent groups design, one can use the F-ratio and degrees of freedom from an analysis of variance to compute an approximation of BF_{01} that is based on a unit information prior (Wagenmakers, 2007; Masson, 2011). Specifically

$$\text{BF}_{01} \approx \sqrt{N^{df_1}\left(1 + \frac{F df_1}{df_2}\right)^{-N}}, \tag{2.4}$$

where $F(df_1, df_2)$ is the F-ratio from a standard analysis of variance applied to these data.

As an example, consider a hypothetical dataset containing $k = 4$ groups of $n = 25$ observations each (for a total of $N = 100$ independent observations). Suppose that

an ANOVA produces $F(3, 96) = 2.76$, $p = 0.046$. This result would be considered as "statistically significant" by conventional null hypothesis standards, and traditional practice would dictate that we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$. But is this result really evidential for $\mathcal{H}_1$? Applying Equation 2.4 shows:

$$\begin{aligned} \text{BF}_{01} &\approx \sqrt{N^{df_1}\Big(1 + \frac{Fdf_1}{df_2}\Big)^{-N}} \\ &= \sqrt{100^3\Big(1 + \frac{0.76 \cdot 3}{96}\Big)^{-100}} \\ &= 15.98. \end{aligned}$$

This result indicates quite the opposite: by definition of the Bayes factor, this implies that the observed data are almost 16 times more likely under $\mathcal{H}_0$ than $\mathcal{H}_1$. Note that the appearance of such contradictory conclusions from two different testing frameworks is actually a classic result known as Lindley's paradox (Lindley, 1957).

3 The BIC approximation for repeated measures

Against this background, the goal now is to extend Equation 2.4 to the case where we have an experimental design with repeated measurements. For context, consider an experiment where k measurements are taken from each of n experimental subjects. We then have a total of $N = nk$ observations, but they are no longer independent measurements. Assume a linear mixed model structure on the observations:

$$y_{ij} = \mu + \alpha_j + \pi_i + \varepsilon_{ij}; \quad i = 1, \ldots, n; \; j = 1 \cdots, k,$$

where μ represents the grand mean, α_j represents the treatment effect associated with group j, π_i represents the effect of subject i, and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_\varepsilon)$. Due to the correlated structure of these data, we have $n(k-1)$ independent observations. We will define models $\mathcal{H}_0$ and $\mathcal{H}_1$ as above. Also, we will denote the sums of squares terms in the model in the usual way, where

$$SSA = n\sum_{j=1}^{k}(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot})^2, \qquad SSB = k\sum_{i=1}^{n}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$$

represent the sums of squares corresponding to the treatment effect and the subject effect, respectively,

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij} - \overline{y}_{\cdot\cdot})^2$$

represents the total sum of squares, and

$$SSR = SST - SSA - SSB$$

represents the residual sum of squares left over after accounting for both treatment and subject effects. From here, we can compute the F-statistic for the treatment effect in our

design as

$$F = \frac{SSA}{SSR} \cdot \frac{df_{\text{residual}}}{df_{\text{treatment}}} = \frac{SSA}{SSR} \cdot \frac{(n-1)(k-1)}{k-1} = \frac{SSA}{SSR} \cdot (n-1).$$

We will now show that this F statistic can be used to estimate BF_{01}.

To this end, note the following. Prior work of Wagenmakers (2007) has shown that BF_{01} can be approximated as

$$\text{BF}_{01} \approx \exp(\Delta BIC_{10}/2),$$

where

$$\Delta BIC_{10} = N \ln\left(\frac{SSE_1}{SSE_0}\right) + (\kappa_1 - \kappa_0)\ln(N).$$

Here, N is equal to the number of independent observations; as noted above, this is equal to $n(k-1)$ for our repeated measures design. SSE_1 represents the variability left unexplained by $\mathcal{H}_1$; for our design, this is equal to the residual sum of squares, SSR. SSE_0 represents the variability left unexplained by $\mathcal{H}_0$; for our design, this is equal to the sum of the treatment sum of squares and the residual sum of squares, $SSA + SSR$. Finally, $\kappa_1 - \kappa_0$ is equal to the difference in the number of parameters between $\mathcal{H}_1$ and $\mathcal{H}_0$; this is equal to $k-1$ for our design.

We are now ready to derive a formula for BF_{01}. First, we will re-express ΔBIC_{10} in terms of F:

$$\begin{aligned}
\Delta BIC_{10} &= N \ln\left(\frac{SSE_1}{SSE_0}\right) + (\kappa_1 - \kappa_0)\ln(N) \\
&= n(k-1)\ln\left(\frac{SSR}{SSR + SSA}\right) + (k-1)\ln\Big(n(k-1)\Big) \\
&= n(k-1)\ln\left(\frac{1}{1 + \frac{SSA}{SSR}}\right) + (k-1)\ln\Big(n(k-1)\Big) \\
&= n(k-1)\ln\left(\frac{n-1}{n-1 + \frac{SSA}{SSR}\cdot(n-1)}\right) + (k-1)\ln\Big(n(k-1)\Big) \\
&= n(k-1)\ln\left(\frac{n-1}{n-1+F}\right) + (k-1)\ln\Big(n(k-1)\Big)
\end{aligned}$$

Thus, we can write

$$\begin{aligned}
\text{BF}_{01} &\approx \exp(\Delta BIC_{10}/2) \\
&= \exp\left[\frac{n(k-1)}{2}\ln\left(\frac{n-1}{n-1+F}\right) + \frac{k-1}{2}\ln\Big(n(k-1)\Big)\right] \\
&= \left(\frac{n-1}{n-1+F}\right)^{\frac{n(k-1)}{2}} \cdot \Big(n(k-1)\Big)^{\frac{k-1}{2}} \\
&= \sqrt{\Big(n(k-1)\Big)^{k-1} \cdot \left(\frac{n-1}{n-1+F}\right)^{n(k-1)}} \\
&= \sqrt{(nk-n)^{k-1} \cdot \left(\frac{n-1}{n-1+F}\right)^{nk-n}}
\end{aligned}$$

If we invert the term containing F and divide $n-1$ into the resulting numerator, we get the following formula:

$$\text{BF}_{01} \approx \sqrt{(nk-n)^{k-1} \cdot \left(1+\frac{F}{n-1}\right)^{n-nk}}, \tag{3.1}$$

where n equals the number of subjects and k equals the number of repeated measurements per subject.

I will now give an example of using Equation 3.1 to compute a Bayes factor. The example below is based on data from Faulkenberry et al. (2018). In this experiment, subjects were presented with pairs of single digit numerals and asked to choose the numeral that was presented in the larger font size. For each of $n = 23$ subjects, response times were recorded in $k = 2$ conditions – congruent trials and incongruent trials. Congruent trials were defined as those in which the physically larger digit was also the numerically larger digit (e.g., 2 – 8). Incongruent trials were defined such that the physically larger digit was numerically smaller (e.g., 2 – 8). Faulkenberry et al. (2018) then fit each subjects' *distribution* of response times to a parametric model (a shifted Wald model; see Anders et al., 2016; Faulkenberry, 2017, for details), allowing them to investigate the effects of congruity on shape, scale, and location of the response time distributions. Specifically, they predicted that the leading edge, or *shift*, of the distributions would not differ between congruent and incongruent trials, thus providing support against an early encoding-based explanation of the observed size-congruity effect (Santens and Verguts, 2011; Faulkenberry et al., 2016; Sobel et al., 2016, 2017). The shift parameter was calculated for both of the $k = 2$ congruity conditions for each of the $n = 23$ subjects. The resulting ANOVA summary table is presented in Table 1.

Table 1: ANOVA summary table for shift parameter data of Faulkenberry et al. (2018)

Source	SS	df	MS	F	p
Subjects	103 984	22	4727		
Treatment	739	1	739	1.336	0.260
Residual	12 176	22	553		
Total	116 399	45			

Applying the minimal BIC method from Equation 2.4 gives us the following:

$$\begin{aligned}
\text{BF}_{01} &\approx \sqrt{(nk-n)^{k-1} \cdot \left(1 + \frac{F}{n-1}\right)^{n-nk}} \\
&= \sqrt{(23 \cdot 2 - 23)^{2-1} \left(1 + \frac{1.336}{23-1}\right)^{(23-23\cdot 2)}} \\
&= \sqrt{23^1 \left(1 + \frac{1.336}{22}\right)^{-23}} \\
&= 2.435
\end{aligned}$$

This Bayes factor tells us that the observed data are approximately 2.4 times more likely under $\mathcal{H}_0$ than $\mathcal{H}_1$. Assuming equal prior model odds, we use Equation 2.3 to convert the Bayes factor to a posterior model probability, giving positive evidence for $\mathcal{H}_0$:

$$\begin{aligned}
p(\mathcal{H}_0 \mid \boldsymbol{y}) &= \frac{\text{BF}_{01}}{\text{BF}_{01} + 1} \\
&= \frac{2.435}{2.435 + 1} \\
&= 0.709.
\end{aligned}$$

4 Accounting for correlation between repeated measurements

In a recent paper, Nathoo and Masson (2016) took a slightly different approach to calculating Bayes factors for repeated measures designs, investigating the role of *effective sample size* in repeated measures designs (Jones, 2011). For single-factor repeated measures designs, effective sample size is defined as

$$n_{\text{eff}} = \frac{nk}{1 + \rho(k-1)},$$

where ρ is the intraclass correlation,

$$\rho = \frac{\sigma_\pi^2}{\sigma_\pi^2 + \sigma_\varepsilon^2}.$$

Thus, $\rho = 0$ implies $n_{\text{eff}} = nk$, whereas $\rho = 1$ implies $n_{\text{eff}} = n$. Though ρ is unknown, Nathoo and Masson (2016) developed a method to estimate it from SS values in the ANOVA, leading to the following:

$$\begin{aligned}\Delta BIC_{10} = n(k-1)\ln\left(\frac{SST - SSA - SSB}{SST - SSB}\right) \\ + (k+2)\ln\left(\frac{n(SST - SSA)}{SSB}\right) \\ - 3\ln\left(\frac{nSST}{SSB}\right)\end{aligned}$$

This estimate provides a better account of the correlation between repeated measurements, but the benefit comes at a price of added complexity, and it is not clear how to reduce this formula to a simple expression involving only F as we do with Equation 3.1. This leads to the natural question: how well does the minimal BIC method from Equation 3.1 match up with the more complex approach of Nathoo and Masson (2016)?

As a first step toward answering this question, let us revisit the example presented above. We can apply the Nathoo and Masson formula to the ANOVA summary in Table 1:

$$\begin{aligned}\Delta BIC_{10} &= 23(2-1)\ln\left(\frac{116399 - 739 - 103984}{116399 - 103984}\right) \\ &\quad + (2+2)\ln\left(\frac{23(116399 - 739)}{103984}\right) \\ &\quad - 3\ln\left(\frac{23(116399)}{103984}\right) \\ &= 23\ln(0.9405) + 4\ln(25.583) - 3\ln(25.746) \\ &= 1.812.\end{aligned}$$

This equates to a Bayes factor of

$$\begin{aligned}\text{BF}_{01} &= \exp(\Delta BIC_{10}/2) \\ &= \exp(1.812/2) \\ &= 2.474\end{aligned}$$

and a posterior model probability of $p(\mathcal{H}_0 \mid \boldsymbol{y}) = 2.474/(2.474 + 1) = 0.712$. Clearly, these computations are quite similar to the ones we performed with Equation 3.1, with both methods indicating positive evidence for $\mathcal{H}_0$ over $\mathcal{H}_1$.

5 Simulation study

The computations in the previous section reflect two preliminary facts. First, the method of Nathoo and Masson (2016) yields Bayes factors and posterior model probabilities that

take into account an estimate of the correlation between repeated measurements. This is a highly principled approach which the minimal BIC method of Equation 3.1 does not take. However, as we can see with both computations, the general conclusion remains the same regardless of whether we use the minimal BIC method or the method of Nathoo and Masson. Given that our Equation 3.1 is (1) easy to use, and (2) requires only three inputs (the number of subjects n, the number of repeated measurement conditions k, and the F statistic), we wonder if the minimal BIC method produces results that are sufficient for day-to-day work, with the risk of being conservative being outweighed by its simplicity. To answer this question, I conducted a Monte Carlo simulation[3] to systematically investigate the relationship between Equation 3.1 and the Nathoo and Masson method across a wide variety of randomly generated datasets.

In this simulation, I randomly generated datasets that reflected the repeated measures designs that we have discussed throughout this paper. Specifically, data were generated from the linear mixed model

$$Y_{ij} = \mu + \alpha_j + \pi_i + \varepsilon_{ij}; \quad i = 1, \ldots, n; \quad j = 1, \ldots, k,$$

where μ represents a grand mean, α_j represents a treatment effect, and π_i represents a subject effect. For convenience, I set $k = 3$, though similar results were obtained with other values of k (not reported here). Also, I assumed $\pi_i \sim \mathcal{N}(0, \sigma_\pi^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. I then systematically varied three components of the model:

1. The number of subjects n was set to either $n = 20$, $n = 50$, or $n = 80$;

2. The intraclass correlation ρ between treatment conditions was set to be either $\rho = 0.2$ or $\rho = 0.8$;

3. The size of the treatment effect was manipulated to be either null, small, or medium. Specifically, these effects were defined as follows. Let $\mu_j = \mu + \alpha_j$ (i.e., the condition mean for treatment j). Then we define effect size as

$$\delta = \frac{\max(\mu_j) - \min(\mu_j)}{\sqrt{\sigma_\pi^2 + \sigma_\varepsilon^2}},$$

and correspondingly, we set δ to one of three values: $\delta = 0$ (null effect), $\delta = 0.2$ (small effect), and $\delta = 0.5$ (medium effect). Also note that since we can write the intraclass correlation as

$$\rho = \frac{\sigma_\pi^2}{\sigma_\pi^2 + \sigma_\varepsilon^2},$$

it follows directly that we can alternatively parameterize effect size as

$$\delta = \frac{\sqrt{\rho}\big(\max(\mu_j) - \min(\mu_j)\big)}{\sigma_\pi}.$$

Using this expression, I was able to set the marginal variance $\sigma_\pi^2 + \sigma_\varepsilon^2$ to be constant across the varying values of our simulation parameters.

[3]The simulation script (in R) and resulting simulated datasets can be downloaded from `https://git.io/Jfekh`.

For each combination of number of observations ($n = 20, 50, 80$), effect size ($\delta = 0, 0.2, 0.5$), and intraclass correlation ($\rho = 0.2, 0.8$), I generated 1000 simulated datasets. For each of the datasets, I performed a repeated measures analysis of variance and, using the F statistic and relevant values of n and k, extracted two Bayes factors for $\mathcal{H}_0$; one based on the minimal BIC method of Equation 3.1 and one based on the method of Nathoo and Masson (2016) which accounts for correlation between repeated measurements. These Bayes factors were then converted to posterior probabilities via Equation 2.3. To compare the performance of both methods in the simulation, I considered four analyses for each simulated dataset: (1) a visualization of the distribution of posterior probabilities $p(\mathcal{H}_0 \mid \boldsymbol{y})$; (2) a calculation of the proportion of simulated trials for which the correct model was chosen (i.e., model choice accuracy); (3) a calculation of the proportion of simulated trials for which both methods chose the same model (i.e., model choice consistency); and (4) a calculation of the correlation between posterior probabilities from both methods.

First, let us visualize the distribution of posterior probabilities $p(\mathcal{H}_0 \mid \boldsymbol{y})$. To this end, I constructed boxplots of the posterior probabilities, which can be seen in Figure 1. The primary message of Figure 1 is clear. Our Equation 3.1, which was derived from minimal BIC method developed in this paper appears to produce a distribution of posterior probabilities which is similar to those produced by the method of Nathoo and Masson (2016). Moreover, this consistency extends across a variety of reasonably common empirical situations. In the cases where $\mathcal{H}_0$ was true (the first row of Figure 1, both Equation 3.1 and the Nathoo and Masson (2016) method produce posterior probabilities for $\mathcal{H}_0$ that are reasonably large. For both methods, the variation of these estimates decreases as the number of observations increases. When the intraclass correlation is small ($\rho = 0.2$), the estimates from Equation 3.1 and the Nathoo and Masson (2016) method are virtually identical. When the intraclass correlation is large ($\rho = 0.8$), the Nathoo and Masson (2016) method introduces slightly more variability in the posterior probability estimates. In all, these results indicate that Equation 3.1 is slightly more favorable when $\mathcal{H}_0$ is true.

For small effects (row 2 of Figure 1), the performance of both methods depended heavily on the correlation between repeated measurements. For small intraclass correlation ($\rho = 0.2$), both methods were quite supportive of $\mathcal{H}_0$, even though $\mathcal{H}_1$ was the true model. This reflects the conservative nature of the BIC approximation (Wagenmakers, 2007); since the unit information prior is uninformative and puts reasonable mass on a large range of possible effect sizes, the predictive updating value for any positive effect (i.e., BF_{10}) will be smaller than would be the case if the prior was more concentrated on smaller effects. As a result, the posterior probability for $\mathcal{H}_1$ is smaller as well. Regardless, the minimal BIC method (Equation 3.1) and the Nathoo and Masson (2016) method produce a similar range of posterior probabilities. The picture is different when the intraclass correlation is large ($\rho = 0.8$); both methods produce a wide range of posterior probabilities, though they are again highly comparable. It is worth pointing out that the posterior probability estimates all improve with increasing numbers of observations; but this should not be surprising, given that the BIC approximation underlying both the minimal BIC method and the Nathoo and Masson (2016) method is a large sample approximation technique. For medium effects (row 3 of Figure 1), we see much of the same message that we've already discussed previously. Both Equation 3.1 and the Nathoo and Masson (2016) method produce similar posterior probability values for $\mathcal{H}_0$. Both methods im-

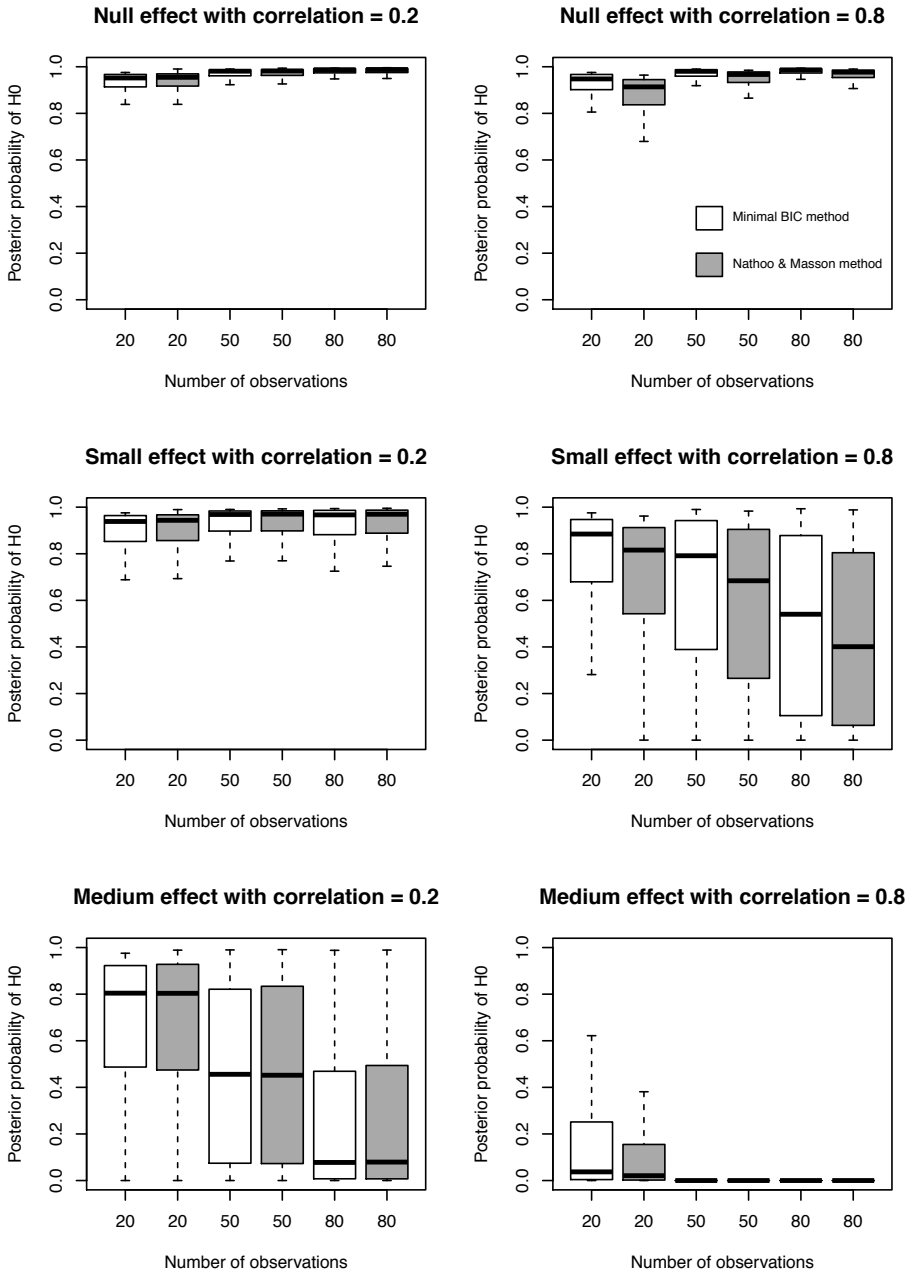

Figure 1: Results from our simulation. Each boxplot depicts the distribution of the posterior probability $p(\mathcal{H}_0 \mid \boldsymbol{y})$ for 1000 Monte Carlo simulations. White boxes represent posterior probabilities derived from Bayes factors that were computed using the minimal BIC method of Equation 3.1. Gray boxes represent posterior probabilities that come from the method of Nathoo and Masson (2016) which accounts for correlation between repeated measurements.

Table 2: Model choice accuracy for the minimal BIC method and the Nathoo and Masson (2016) method, calculated as the proportion of simulated datasets for which the correct model was chosen

	Correlation = 0.2		Correlation = 0.8	
	Minimal BIC	Nathoo & Masson	Minimal BIC	Nathoo & Masson
Null effect				
$n = 20$	0.969	0.968	0.979	0.954
$n = 50$	0.989	0.988	0.991	0.981
$n = 80$	0.992	0.992	0.992	0.985
Small effect				
$n = 20$	0.068	0.072	0.148	0.218
$n = 50$	0.058	0.056	0.307	0.374
$n = 80$	0.062	0.062	0.485	0.550
Medium effect				
$n = 20$	0.259	0.266	0.867	0.910
$n = 50$	0.526	0.530	0.997	0.999
$n = 80$	0.760	0.756	1.000	1.000

prove with increasing sample size, and at least for medium-size effects, the computations are quite reliable for high values of correlation between repeated measurements.

Though the distributions of posterior probabilities appear largely the same, it is not clear to what extent the two methods provide the user with an accurate inference. Since the data are simulated, it is possible to define a "correct" model in each case – for simulated datasets where $\delta = 0$, the correct model is $\mathcal{H}_0$, whereas when $\delta = 0.2$ or $\delta = 0.5$, the correct model is $\mathcal{H}_1$. To compare the performance of both methods, I calculated *model choice accuracy*, defined as the proportion of simulated datasets for which the correct model was chosen. Model choice was defined by considering $\mathcal{H}_0$ to be chosen whenever $\text{BF}_{01} > 1$ and $\mathcal{H}_1$ to be chosen whenever $\text{BF}_{01} < 1$. The results are displayed in Table 2.

Let us consider Table 2 in three sections. First, for data that were simulated from a null model, it is clear that the accuracy of both methods is excellent, with model choice accuracies all above 95%. Further, the minimal BIC method outperforms the Nathoo and Masson (2016) method across all possible sample sizes as well as correlation conditions. However, the overall performance of both methods becomes more questionable for small effects. Model choice accuracies are no better than 5-7% (regardless of sample size) for datasets with small correlation ($\rho = 0.2$) between repeated measurements. The situation improves a bit when this correlation increases to 0.8, though never gets better than 55%. Across all the small-effect datasets, the Nathoo and Masson method is slightly more accurate in choosing the correct model. This pattern continues for datasets which are simulated to have a large effect, though overall accuracy is much better in this case.

Overall, this pattern of results permits two conclusions. First, the BIC method (upon which both methods are based) tends to be conservative (Wagenmakers, 2007), so the tendency to select the null model in the presence of small effects is unsurprising. Second, though performance was variable in the presence of small and medium effects, the differences in model choice accuracies between the minimal BIC method and the Nathoo and

Table 3: Model choice consistency for the minimal BIC method and the Nathoo and Masson (2016) method, calculated as the proportion of simulated datasets for which both methods chose the same model

	Null effect	Small effect	Medium effect
Correlation = 0.2			
$n = 20$	0.997	0.994	0.977
$n = 50$	0.999	0.994	0.984
$n = 80$	1.000	0.998	0.994
Correlation = 0.8			
$n = 20$	0.975	0.930	0.957
$n = 50$	0.990	0.933	0.998
$n = 80$	0.993	0.935	1.000

Masson (2016) method were small. Thus, any performance penalty that is exhibited for the minimal BIC method is shared by the Nathoo & Masson method as well, reflecting not a limitation of the minimal BIC method, but a limitation of the BIC method in general. To further validate this claim, I calculated model choice *consistency*, defined as the proportion of simulated datasets for which both methods chose the *same* model. As can be seen in Table 3, both the minimal BIC method and the Nathoo and Masson method choose the same model in a large proportion of the simulated datasets, regardless of effect size, sample size, or correlation between repeated measurements.

As a final investigation, I calculated the correlations between the posterior probabilities that were produced by both methods. These correlations can be seen in Table 4 and Figure 2 – note that the figure only shows scatterplots for the $n = 50$ condition, though the $n = 20$ and $n = 80$ conditions produce similar plots. Table 4 shows very high correlations between the posterior probability calculations. As can be seen in Figure 4, the relationship is linear when repeated measurements are assumed to have a small correlation, but nonlinear in the presence of highly correlated repeated measurements. For highly correlated measurements, the curvature of the scatterplot indicates that for a given simulated dataset, the posterior probability (for $\mathcal{H}_0$) calculated by the minimal BIC method will tend to be greater than the posterior probability calculated by the Nathoo and Masson (2016) method. Again, this is hardly surprising, as the Nathoo and Masson method is designed to better take into account the correlation between repeated measurements. One should note that this correction is advantageous for datasets generated from a positive-effects model, but disadvantageous for datasets generated from a null model.

In all, the performance of the minimal BIC method is quite comparable to the Nathoo and Masson (2016) method. Though the Nathoo and Masson method is designed to better account for the correlation between repeated measurements, this advantage comes at a cost of increased complexity. On the other hand, the minimal BIC method introduced in this paper requires the user to only know the F-statistic, the number of subjects, and the number of repeated measures conditions. Thus, the small performance penalties for the minimal BIC method are far outweighed by its computational simplicity.

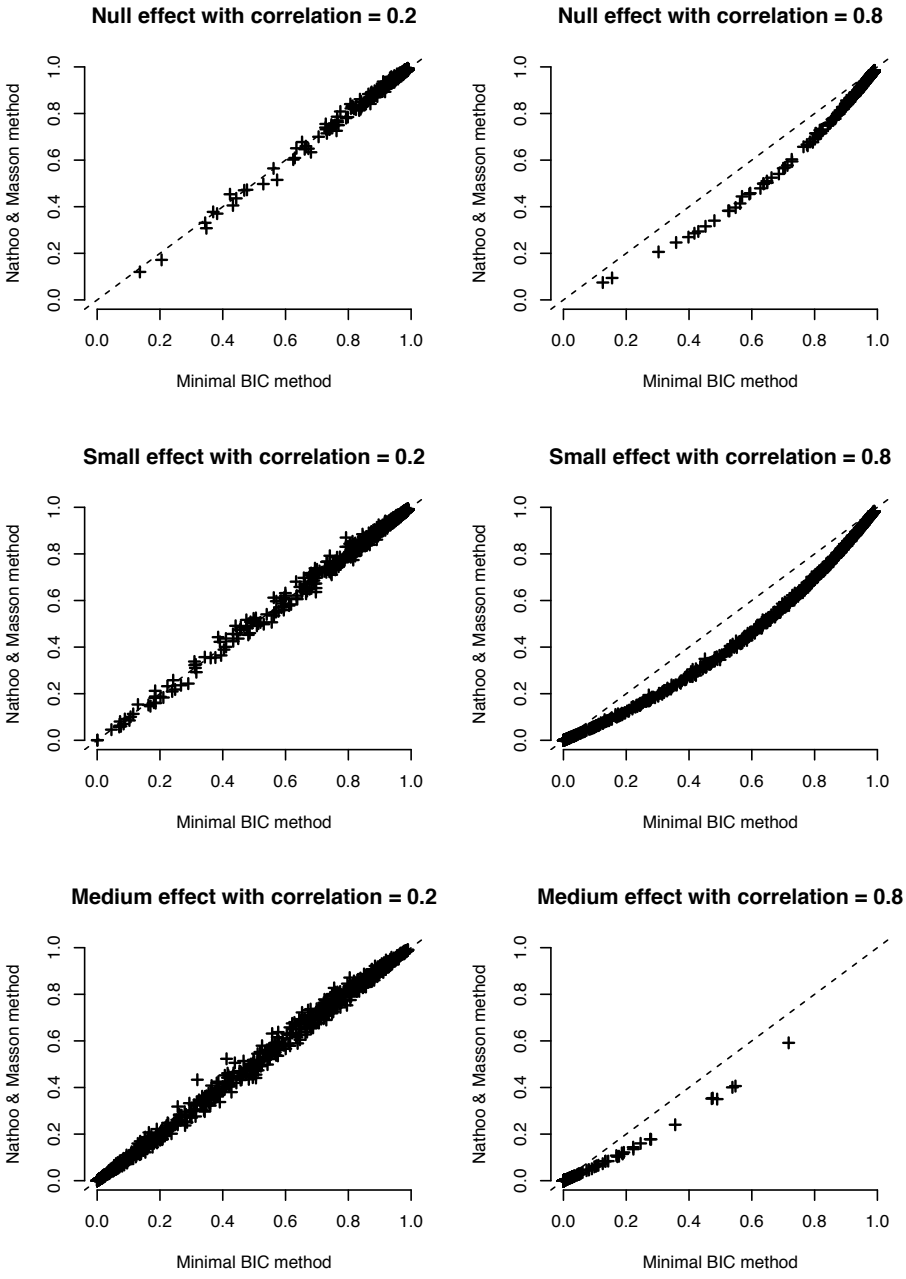

Figure 2: Scatterplot demonstrating the relationship between posterior probabilities calculated by the minimal BIC method (on the horizontal axis) and the Nathoo and Masson (2016) method (on the vertical axis). Sample size is assumed to be $n = 50$ for all plots.

Table 4: Correlations between the posterior probabilities $p(\mathcal{H}_0 \mid \boldsymbol{y})$ calculated by the minimal BIC method and the Nathoo and Masson (2016) method

	Correlation = 0.2	Correlation = 0.8
Null effect		
$n = 20$	0.993	0.987
$n = 50$	0.997	0.990
$n = 80$	0.998	0.988
Small effect		
$n = 20$	0.994	0.989
$n = 50$	0.998	0.991
$n = 80$	0.999	0.991
Medium effect		
$n = 20$	0.995	0.990
$n = 50$	0.999	0.995
$n = 80$	0.999	0.999

6 Conclusion

In this paper, I have proposed a formula for estimating Bayes factors from repeated measures ANOVA designs. These ideas extend previous work of Faulkenberry (2018), who presented such formulas for between-subject designs. Such formulas are advantageous for researchers in a wide variety of empirical disciplines, as they provide an easy-to-use method for estimating Bayes factors from a minimal set of summary statistics. This gives the user a powerful index for estimating evidential value from a set of experiments, even in cases where the only data available are the summary statistics published in a paper. I think this provides a welcome addition to the collection of tools for doing Bayesian computation with summary statistics (e.g., Ly et al., 2018; Faulkenberry, 2019).

Further, I demonstrated that the minimal BIC method performs similarly to a more complex formula of Nathoo and Masson (2016), who were able to explicitly estimate and account for the correlation between repeated measurements. Though the Nathoo and Masson (2016) approach is certainly more principled than a "one-size-fits-all" approach, it does require knowledge of the various sums-of-squares components from the repeated measures ANOVA, and though I have tried, I have not found an obvious way to recover the Nathoo and Masson (2016) estimates from the F statistic alone. As such, the Nathoo and Masson approach is inaccessible without access to the raw data – or at least the various SS components, which are rarely reported in empirical papers. Thus, given the similar performance compared to the Nathoo and Masson (2016) method, the new minimal BIC method stands at an advantage, not only for its computational simplicity, but also its power in producing maximal information given minimal input.

References

[1] Anders, R., Alario, F.X., and Van Maanen, L. (2016): The shifted Wald distribution for response time data analysis. *Psychological Methods*, **21**(3), 309–327.

[2] Berger, J.O. and Sellke, T. (1987): Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, **82**(397), 112.

[3] Faulkenberry, T.J. (2017): A single-boundary accumulator model of response times in an addition verification task. *Frontiers in Psychology*, **8**, 01225.

[4] Faulkenberry, T.J. (2018): Computing Bayes factors to measure evidence from experiments: An extension of the BIC approximation. *Biometrical Letters*, **55**(1), 31–43.

[5] Faulkenberry, T.J. (2019): Estimating evidential value from analysis of variance summaries: A comment on Ly et al. (2018). *Advances in Methods and Practices in Psychological Science* **2**(4), 406–409.

[6] Faulkenberry, T.J., Cruise, A., Lavro, D., and Shaki, S. (2016): Response trajectories capture the continuous dynamics of the size congruity effect. *Acta Psychologica*, **163**, 114–123.

[7] Faulkenberry, T.J., Vick, A. D., and Bowman, K.A. (2018): A shifted Wald decomposition of the numerical size-congruity effect: Support for a late interaction account. *Polish Psychological Bulletin*, **49**(4), 391–397.

[8] Fisher, R.A. (1925): *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

[9] Gigerenzer, G. (2004): Mindless statistics. *The Journal of Socio-Economics*, **33**(5), 587–606.

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

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

[12] Lindley, D.V. (1957): A statistical paradox. *Biometrika*, **44**(1-2), 187–192.

[13] Ly, A., Raj, A., Etz, A., Marsman, M., Gronau, Q.F., and Wagenmakers, E.-J. (2018): Bayesian reanalyses from summary statistics: A guide for academic consumers. *Advances in Methods and Practices in Psychological Science*, **1**(3), 367–374.

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

[15] Nathoo, F.S. and Masson, M.E. (2016): Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. *Journal of Mathematical Psychology*, **72**, 144–157.

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

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

[18] Santens, S. and Verguts, T. (2011): The size congruity effect: Is bigger always more? *Cognition*, **118**(1), 94–110.

[19] Sobel, K. V., Puri, A.M., and Faulkenberry, T.J. (2016): Bottom-up and top-down attentional contributions to the size congruity effect. *Attention, Perception, & Psychophysics*, **78**(5), 1324–1336.

[20] Sobel, K.V., Puri, A.M., Faulkenberry, T.J., and Dague, T.D. (2017): Visual search for conjunctions of physical and numerical size shows that they are processed independently. *Journal of Experimental Psychology: Human Perception and Performance*, **43**(3), 444–453.

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