Estimating Bayes factors from minimal ANOVA summaries for repeated-measures designs

Thomas J. Faulkenberry

Tarleton State University

(Dated: May 15, 2019)

Abstract

In this paper, we develop a formula for estimating Bayes factors from repeated measures ANOVA designs. The formula, which requires knowing only minimal information about the ANOVA (e.g., the $F$-statistic), is based on the BIC approximation of the Bayes factor, a common default method for Bayesian computation with linear models. In addition to several computational examples, we report a simulation study in which we demonstrate that despite its simplicity, our formula compares favorably to a recently developed, more complex method that accounts for correlation between repeated measurements. Our 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.

Keywords: Analysis of variance; Bayesian inference; Bayes factor; BIC; repeated measures
In this paper, we discuss how to apply the BIC approximation \cite{7, 10, 11, 17} to compute Bayes factors for repeated measures experiments using only minimal summary statistics \cite[e.g., 9]{9}. Critically, we modify the formula derived in Faulkenberry \cite{3} to work for repeated measures experiments, and investigate its performance against a more complex method of Nathoo and Masson \cite{11} which accounts for varying levels of correlation between repeated measurements. Among several “default prior” solutions to computing Bayes factors for common experimental designs \cite{12, 13}, 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.

I. BACKGROUND

Let us begin with a simple case of a one-factor independent groups design. Consider a set of data \( Y_{ij} \), on which we impose a linear model as follows:

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

where \( \mu \) represents the grand mean, \( \alpha_j \) represents the treatment effect associated with group \( j \), and \( \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{\varepsilon}) \). In all, we have \( N = nk \) independent observations. We define two hypotheses:

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

Recall that for \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \), the Bayes factor \cite{7}, denoted \( B_{01} \), is defined as the ratio of marginal likelihoods for \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \), respectively. That is,

\[
B_{01} = \frac{p(\text{data} | \mathcal{H}_0)}{p(\text{data} | \mathcal{H}_1)}.
\]

This ratio indicates the extent to which the prior odds for \( \mathcal{H}_0 \) over \( \mathcal{H}_1 \) are updated after observing data.
In Faulkenberry [3], it was shown that for any independent-groups design, one can use the results of an analysis of variance to compute an approximation of $B_{01}$ that is based on a unit information prior [10, 17]. Specifically

$$B_{01} \approx \sqrt{N^{df_1} \left(1 + \frac{F_{df_1}}{df_2}\right)^{-N}},$$

(1)

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 “significant” by conventional standards, and traditional practice would dictate that we reject $H_0$ in favor of $H_1$. But is this result really evidential for $H_1$? We can apply Equation (1) as follows:

$$B_{01} \approx \sqrt{100^{3} \left(1 + \frac{0.76 \cdot 3}{96}\right)^{-100}}$$

$$= 100^{3} \left(1 + \frac{0.76 \cdot 3}{96}\right)^{-100}$$

$$= 15.98.$$

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 $H_0$ than $H_1$. Note that the appearance of such contradictory conclusions from two different testing frameworks is actually a classic result known as Lindley’s paradox [8].

II. THE BIC APPROXIMATION FOR REPEATED MEASURES

Our goal now is to modify Equation (1) 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 $n$ subjects. We then have a total of $N$ observations, but they are no longer independent measurements. Assume a linear mixed model structure on the observations:

$$Y_{ij} = \mu + \alpha_j + \pi_i + \varepsilon_{ij}; \quad i = 1, \ldots, n; \quad j = 1 \cdots, k,$$
where $\mu$ represents the grand mean, $\alpha_j$ represents the treatment effect associated with group $j$, $\pi_i$ represents the effect of subject $i$, and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Due to the correlated structure of these data, we have $n(k-1)$ independent observations. We will define $\mathcal{H}_0$ and $\mathcal{H}_1$ as above.

Prior work of Wagenmakers [17] has demonstrated that $B_{01}$ can be approximated as $\exp(\Delta BIC_{10}/2)$, 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 $n(k-1)$. $SSE_1$ represents the variability left unexplained by $\mathcal{H}_1$; for an ANOVA, this is equal to $SS_{\text{residual}}$. $SSE_0$ represents the variability left unexplained by $\mathcal{H}_0$; for an ANOVA, this is equal to the sum of $SS_{\text{treatment}}$ and $SS_{\text{residual}}$. 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 $k - 1$.

We are now ready to derive a formula for $B_{01}$. First, we will re-express $\Delta BIC_{10}$ in terms of $F$:

$$\Delta BIC_{10} = \frac{n(k-1) \ln \left( \frac{SS_{\text{residual}}}{SS_{\text{residual}} + SS_{\text{treatment}}} \right)}{n(k-1)} + (k-1) \ln \left( \frac{n}{n-1} \right)$$

$$= \frac{n(k-1) \ln \left( \frac{1}{1 + \frac{SS_{\text{treatment}}}{SS_{\text{residual}}}} \right)}{n(k-1)} + (k-1) \ln \left( \frac{n}{n-1} \right)$$

$$= n(k-1) \ln \left( \frac{\frac{SS_{\text{residual}}}{SS_{\text{residual}} + SS_{\text{treatment}}}}{\frac{SS_{\text{treatment}}}{SS_{\text{residual}}}} \cdot \frac{1}{1 + F} \right) + (k-1) \ln \left( \frac{n}{n-1} \right)$$

$$= n(k-1) \ln \left( \frac{\frac{SS_{\text{residual}}}{SS_{\text{residual}} + SS_{\text{treatment}}}}{\frac{SS_{\text{treatment}}}{SS_{\text{residual}}} \cdot \frac{1}{1 + F}} \right) + (k-1) \ln \left( \frac{n}{n-1} \right)$$

$$= n(k-1) \ln \left( \frac{(n-1)(k-1)}{(n-1)(k-1) + F(k-1)} \right) + (k-1) \ln \left( \frac{n}{n-1} \right)$$

$$= n(k-1) \ln \left( \frac{n-1}{n-1 + F} \right) + (k-1) \ln \left( \frac{n}{n-1} \right)$$
Thus, we can write

\[ B_{01} \approx \exp(\Delta BIC_{10}/2) \]

\[ = \exp \left[ \frac{n(k-1)}{2} \ln \left( \frac{n-1}{n-1+F} \right) + \frac{k-1}{2} \ln (n(k-1)) \right] \]

\[ = \left( \frac{n-1}{n-1+F} \right)^{\frac{n(k-1)}{2}} \cdot (n(k-1))^{\frac{k-1}{2}} \]

\[ = \sqrt{ \left( n(k-1) \right)^{k-1} \cdot \left( \frac{n-1}{n-1+F} \right)^{n(k-1)}} \]

\[ = \sqrt{ (nk-n)^{k-1} \cdot \left( \frac{n-1}{n-1+F} \right)^{nk-n}} \]

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

\[ B_{01} \approx \sqrt{ (nk-n)^{k-1} \cdot \left( 1 + \frac{F}{n-1} \right)^{n-nk}} \]

(2)

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

A. Some examples

We can now apply Equation (2) to compute Bayes factors for a couple of examples. The examples below are based on data from Faulkenberry et al. [5]. 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, median response times were calculated for each of \( k = 2 \) 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 – s). The resulting ANOVA summary table is depicted in Table [I].
TABLE I. ANOVA summary table for response time data of Faulkenberry et al. [5]

| Source   | SS     | df  | MS     | F       | p      |
|----------|--------|-----|--------|---------|--------|
| Subjects | 285639 | 22  | 12984  |         |        |
| Treatment| 45360  | 1   | 45360  | 39.63   | < 0.001|
| Residual | 25182  | 22  | 1145   |         |        |
| Total    | 356181 | 45  |        |         |        |

Applying Equation 1 gives us the following:

\[ B_{01} \approx \sqrt{\frac{(nk - n)^{k-1} \cdot \left(1 + \frac{F}{n - 1}\right)^{n-nk}}{(23 \cdot 2 - 23)^{2-1} \left(1 + \frac{39.63}{23 - 1}\right)^{(23-23-2)}}} \]

\[ = \sqrt{\frac{23^1 \left(1 + \frac{39.63}{22}\right)^{-23}}}{23^1 \left(1 + \frac{39.63}{22}\right)^{-23}} \]

\[ = 0.00003436 \]

The resulting Bayes factor displays quite powerful evidence against \( H_0 \); if we cast the Bayes factor in favor of \( H_1 \), we get \( B_{10} = 1/B_{01} = 1/0.00003436 = 29104 \), indicating that the observed data are approximately 30,000 times more likely under \( H_1 \) than \( H_0 \). This provides overwhelming support for the presence of an effect of physical/numerical congruity on median response times. Converting the Bayes factor to a posterior model probability, we also see incredible evidence for \( H_1 \):

\[ p(H_1 \mid \text{data}) = \frac{B_{10}}{1 + B_{10}} = \frac{29104}{1 + 29104} = 0.99997. \]

Now let us consider our second example. In addition to analyzing median response times, Faulkenberry et al. [5] also fit each subjects’ distribution of response times to a parametric model [i.e., the shifted Wald distribution; see [1, 2] 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 [4, 14–16]. 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 II.

| Source      | SS    | df | MS      | F       | p     |
|-------------|-------|----|---------|---------|-------|
| Subjects    | 103984| 22 | 4727    |         |       |
| Treatment   | 739   | 1  | 739     | 1.336   | 0.26  |
| Residual    | 12176 | 22 | 553     |         |       |
| Total       | 116399| 45 |         |         |       |

**TABLE II. ANOVA summary table for shift parameter data of Faulkenberry et al. [5]**

Applying Equation 1 gives us the following:

\[
B_{01} \approx \sqrt{(nk - n)^{k-1} \cdot \left(1 + \frac{F}{n - 1}\right)^{n-nk}}
\]

\[
= \sqrt{(23 \cdot 2 - 23)^{2-1} \left(1 + \frac{1.336}{23 - 1}\right)^{23-23}}
\]

\[
= \sqrt{23^1 \left(1 + \frac{1.336}{22}\right)^{-23}}
\]

\[
= 2.435
\]

This Bayes factor tells us that the observed data are approximately 2.4 times more likely under $H_0$ than $H_1$. Converting the Bayes factor to a posterior model probability, we also see positive evidence for $H_0$:

\[
p(H_0 | \text{data}) = \frac{B_{01}}{1 + B_{01}}
\]

\[
= \frac{2.435}{1 + 2.435}
\]

\[
= 0.709.
\]
III. ACCOUNTING FOR CORRELATION BETWEEN REPEATED MEASUREMENTS

In a recent paper, Nathoo and Masson [11] took a slightly different approach to the problem we have, investigating the role of *effective sample size* in repeated measures designs [6]. For single-factor repeated measures designs, effective sample size can be computed as

$$n_{eff} = \frac{n k}{1 + \rho (k-1)}$$

where \(\rho\) is the intraclass correlation. When \(\rho = 0\), \(n_{eff} = nk\), and when \(\rho = 1\), \(n_{eff} = n\). Though \(\rho\) is unknown, Nathoo and Masson [11] developed a method to estimate it from SS values in the ANOVA, leading to the following refined estimate:

$$\Delta BIC_{10} = n(k - 1) \ln \left( \frac{SS_{total} - SS_{treatment} - SS_{subject}}{SS_{total} - SS_{subject}} \right)$$

$$+ (k + 2) \ln \left( \frac{n(SS_{total} - SS_{treatment})}{SS_{subject}} \right)$$

$$- 3 \ln \left( \frac{nSS_{total}}{SS_{subject}} \right)$$

Though this estimate certainly provides a better account of the correlation between repeated measurements, the benefit comes at a price of added complexity, and certainly one cannot reduce this formula easily to a simple expression involving only \(F\) as we do with Equation 2. This leads to the natural question: how well does our Equation 2 match up with the more complex approach of Nathoo and Masson [11]?

As a first step toward answering this question, let us revisit the two examples presented above. If we apply the Nathoo and Masson formula to the ANOVA summary in Table [11] we
\[
\Delta BIC_{10} = 23(2 - 1) \ln \left( \frac{356181 - 45360 - 285639}{356181 - 285639} \right) \\
+ (2 + 2) \ln \left( \frac{23(356181 - 45360)}{285639} \right) \\
- 3 \ln \left( \frac{23(356181)}{285639} \right) \\
= 23 \ln(0.3570) + 4 \ln(25.028) - 3 \ln(28.680) \\
= -20.879
\]

We can convert \( \Delta BIC_{10} \) to a Bayes factor, giving us \( B_{01} \approx \exp(\Delta BIC_{10}/2) = \exp(-20.879/2) = 0.00002925 \). As above, we cast this Bayes factor in favor of \( \mathcal{H}_1 \) by inverting, so \( B_{10} = 1/0.0002925 = 34188.03 \). This implies \( p(\mathcal{H}_1 \mid \text{data}) = 34188.03/(1 + 34188.03) = 0.99997 \).

Note that the general interpretation of these results is on par with our earlier method; both indicate overwhelming support for \( \mathcal{H}_1 \). If anything, the approximation we obtained with Equation 2 is slightly conservative regarding support for \( \mathcal{H}_1 \); this is because the method of Nathoo and Masson was designed to reduce the BIC penalty for \( \mathcal{H}_1 \) when repeated measures conditions are highly correlated; compared to the formulation upon which Equation 2 is based, this will tend to increase the support for \( \mathcal{H}_1 \).

We can do a similar computation with the data from Table II:

\[
\Delta BIC_{10} = 23(2 - 1) \ln \left( \frac{116399 - 739 - 103984}{116399 - 103984} \right) \\
+ (2 + 2) \ln \left( \frac{23(116399 - 739)}{103984} \right) \\
- 3 \ln \left( \frac{23(116399)}{103984} \right) \\
= 23 \ln(0.9405) + 4 \ln(25.583) - 3 \ln(25.746) \\
= 1.812
\]

This equates to a Bayes factor of \( B_{01} = \exp(\Delta BIC_{10}/2) = \exp(1.812/2) = 2.474 \) and a posterior model probability of \( p(\mathcal{H}_0 \mid \text{data}) = 2.474/(1 + 2.474) = 0.712 \). Clearly, these
computations are quite similar to the ones we performed with Equation 2; both indicate positive evidence for $H_0$ over $H_1$.

IV. SIMULATION STUDY

The computations in the previous section reflect two preliminary findings. First, the revised BIC formula of Nathoo and Masson [11] 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 our Equation 2 does not take. However, as we can see with both computations, the general conclusion remains the same regardless of whether we used Equation 2 or the Nathoo and Masson method. Given that our Equation 2 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), could it be that Equation 2 produces results that are sufficient for day-to-day work, with the risk of being conservative being outweighed by the simplicity of our formula? To answer this question, we conducted a Monte Carlo simulation to systematically investigate the relationship between Equation 2 and the Nathoo and Masson method across a wide variety of randomly generated datasets.

In this simulation, we 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 $\mu$ represents a grand mean, $\alpha_j$ represents a treatment effect, and $\pi_i$ represents a subject effect. For convenience, we set $k = 3$, though similar results were obtained with other values of $k$ (not reported here). Also, we assume $\pi_i \sim \mathcal{N}(0, \sigma^2\pi)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2\varepsilon)$. We systematically varied three components of the model:

1. The number of observations $n$ for each subject was set to either $n = 20$, $n = 50$, or $n = 80$;

2. The intraclass correlation $\rho$ 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

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

and correspondingly, we set $\delta$ 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

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

it follows directly that we can alternatively parameterize effect size as

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

Using this expression, we were able to set our marginal variance $\sigma^2_\pi + \sigma^2_\varepsilon$ to be constant across the varying values of our simulation parameters.

For each combination of number of observations ($n = 20, 50, 80$), effect size ($\delta = 0, 0.2, 0.5$), and intraclass correlation ($\rho = 0.2, 0.8$), we generated 1000 simulated datasets. For each of the datasets, we applied a repeated-measures ANOVA model and extracted two posterior probabilities for $H_0$; one based on Equation 2 and one based on the refined estimate of Nathoo and Masson [11]. The results are depicted in Figure 1.

The primary message of Figure 1 is clear; our Equation 2, which was derived from the original BIC method [3, 10, 17] performs comparably to the refined BIC method of Nathoo and Masson [11] across a variety of empirical situations. In the cases where $H_0$ was true (the first row of Figure 1, both Equation 2 and the Nathoo and Masson [11] method produce posterior probabilities for $H_0$ 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 2 and the Nathoo and Masson [11] method are virtually identical. When the intraclass correlation is large ($\rho = 0.8$), the Nathoo and Masson [11] method introduces slightly more variability in the posterior probability estimates. In all, these results indicate that Equation 2 is slightly more favorable when $H_0$ 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 \([17]\); 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., \( B_{10} \) 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 original BIC method (Equation 2 and the Nathoo and Masson \([11]\) method produce similar results. 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 Equation 2 and the Nathoo and Masson \([11]\) method is 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 2 and the Nathoo and Masson \([11]\) method produce similar posterior probability values for \( \mathcal{H}_0 \). Both methods improve with increasing sample size, and at least for medium-size effects, the computations are quite reliable for high values of correlation between repeated measurements.

V. CONCLUSION

In this paper, we have proposed a formula for estimating Bayes factors from repeated measures ANOVA designs. These ideas extend previous work of Faulkenberry \([3]\), 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. We think this provides a welcome addition to the collection of tools for doing Bayesian computation with summary statistics \([e.g., 9]\).

Further, we demonstrated that our formula performs similarly to a more refined, yet more complex formula of Nathoo and Masson \([11]\), who were able to explicitly estimate and account for the correlation between repeated measurements. Though the Nathoo and
FIG. 1. Results from our simulation. Each boxplot depicts the distribution of the posterior probability $p(H_0 \mid \text{data})$ for 1000 Monte Carlo simulations. White boxes represent posterior probabilities derived from Bayes factors that were computed using Equation 2. Gray boxes represent posterior probabilities that come from the refined Bayes factor of Nathoo and Masson [11].
Masson [11] 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 to our knowledge, there is not yet any obvious way to recover the Nathoo and Masson [11] estimates from the $F$ statistic alone. Thus, given the similar performance between our method compared to the Nathoo and Masson [11] method, we think our method stands at a slight advantage, not only for its simplicity, but also its power in light of minimal available information.

[1] Anders, R., Alario, F.-X., and Van Maanen, L. (2016). The shifted wald distribution for response time data analysis. *Psychological Methods*, 21(3):309327.

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

[3] Faulkenberry, T. J. (2018). Computing bayes factors to measure evidence from experiments: An extension of the bic approximation. *Biometrical Letters*, 55(1):3143.

[4] Faulkenberry, T. J., Cruise, A., Lavro, D., and Shaki, S. (2016). Response trajectories capture the continuous dynamics of the size congruity effect. *Acta Psychologica*, 163:114123.

[5] 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.

[6] Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, 30(25):30503056.

[7] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773795.

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

[9] 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.

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

[11] Nathoo, F. S. and Masson, M. E. (2016). Bayesian alternatives to null-hypothesis significance
testing for repeated-measures designs. *Journal of Mathematical Psychology*, 72:144157.

[12] 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.

[13] 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.

[14] Santens, S. and Verguts, T. (2011). The size congruity effect: Is bigger always more? *Cognition*, 118(1):94110.

[15] 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):13241336.

[16] 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):444453.

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