A Tutorial on Conducting and Interpreting a Bayesian ANOVA in JASP

Don van den Bergh*[1], Johnny van Doorn[1], Maarten Marsman[1], Tim Draws[1], Erik-Jan van Kesteren[4], Koen Derks[1,2], Fabian Dablander[1], Quentin F. Gronau[1], Šimon Kucharský[1], Akash R. Komarlu Narendra Gupta[1], Alexandra Sarafoglou[1], Jan G. Voelkel[5], Angelika Stefan[1], Alexander Ly[1,3], Max Hinne[1], Dora Matzke[1], and Eric-Jan Wagenmakers[1]

[1]University of Amsterdam
[2]Nyenrode Business University
[3]Centrum Wiskunde & Informatica
[4]Utrecht University
[5]Stanford University

Abstract

Analysis of variance (ANOVA) is the standard procedure for statistical inference in factorial designs. Typically, ANOVAs are executed using frequentist statistics, where p-values determine statistical significance in an all-or-none fashion. In recent years, the Bayesian approach to statistics is increasingly viewed as a legitimate alternative to the p-value. However, the broad adoption of Bayesian statistics –and Bayesian ANOVA in particular– is frustrated by the fact that Bayesian concepts are rarely taught in applied statistics courses. Consequently, practitioners may be unsure how to conduct a Bayesian ANOVA and interpret the results. Here we provide a guide for executing and interpreting a Bayesian ANOVA with JASP, an open-source statistical software program with a graphical user interface. We explain the key concepts of the Bayesian ANOVA using two empirical examples.

*Correspondence concerning this article should be addressed to:
Don van den Bergh
University of Amsterdam, Department of Psychological Methods
Postbus 15906, 1001 NK Amsterdam, The Netherlands
E-Mail should be sent to: donvdbergh@hotmail.com.

Ubiquitous across the empirical sciences, analysis of variance (ANOVA) allows researchers to assess the effects of categorical predictors on a continuous outcome variable. Consider for instance an experiment by Strack, Martin, and Stepper (1988) designed to test the *facial feedback hypothesis*, that is, the hypothesis that people's affective responses can be influenced by their own facial expression. Participants were randomly assigned to one of three conditions. In the *lips* condition, participants were instructed to hold a pen with their lips, inducing a pout. In the *teeth* condition, participants were instructed to hold a pen between their teeth, inducing a smile. In the control condition, participants were told to hold a pen in their nondominant hand. With the pen in the instructed position, each participant then rated four cartoons for funniness. The outcome variable was the average funniness rating across the four cartoons. The ANOVA procedure may be used to test the null hypothesis that the pen position does not result in different funniness ratings.

ANOVAs are typically conducted using frequentist statistics, where p-values decide statistical significance in an all-or-none manner: if $p < .05$, the result is deemed statistically significant and the null hypothesis is rejected; if $p > .05$, the result is deemed statistically nonsignificant, and the null hypothesis is retained. Such binary thinking has been critiqued extensively (e.g., Cohen, 1994; Rouder, Engelhardt, McCabe, & Morey, 2016; Amrhein, Greenland, & McShane, 2019), and some perceive it as a cause of the reproducibility crisis in psychology (Cumming, 2014; but see Savalei & Dunn, 2015). In recent years, several alternatives to p-values have been suggested, for example reporting confidence intervals (Cumming, 2014; Gardner & Altman, 1986) or abandoning null hypothesis testing altogether (McShane, Gal, Gelman, Robert, & Tackett, 2019).

Here we focus on another alternative: Bayesian inference. In the Bayesian framework, knowledge about parameters and hypotheses is updated as a function of predictive success – hypotheses that predicted the observed data relatively well receive a boost in credibility, whereas hypotheses that predicted the data relatively poorly suffer a decline (Wagenmakers, Morey, & Lee, 2016). A series of recent articles show how the Bayesian framework can supplement or supplant the frequentist p-value (e.g., Burton, Gurrin, & Campbell, 1998; Dienes & McLatchie, 2018; Jarosz & Wiley, 2014; Masson, 2011; Nathoo & Masson, 2016; Rouder et al., 2016).

The advantages of the Bayesian paradigm over the frequentist p-value are well documented (e.g., Wagenmakers et al., 2018); for instance, with Bayesian inference researchers can incorporate prior knowledge and quantify support, both in favor and against the null-hypothesis; furthermore, this support may be monitored as the data accumulate (?, ?). Despite these and other advantages, Bayesian analyses are still used only sparingly in the social sciences (van der Schoot, Winter, Ryan, Zondervan Zwijnenburg, & Depaoli, 2017). The broad adoption of Bayesian statistics –and Bayesian ANOVA in particular– is hindered by the fact that Bayesian concepts are rarely taught in applied statistics courses. Consequently, practitioners may be unsure of how to conduct a Bayesian ANOVA and interpret the results.

To help familiarize researchers with Bayesian inference for common experimental designs, this article provides a guide for conducting and interpreting a Bayesian ANOVA with JASP (JASP Team, 2019). JASP is a free, open-source statistical software program with a graphical user interface that offers both Bayesian and frequentist analyses. Below, we first provide a brief introduction to Bayesian statistics. Subsequently, we use two data examples to explain the key concepts of ANOVA.

Bayesian Foundations

This section explains some of the fundamentals of Bayesian inference. We focus on interpretation rather than mathematical detail; see the special issue on Bayesian inference by Vandekerckhove, Rouder, and Kruschke (2018) for a set of comprehensive, low-level introductions to Bayesian inference.

The central goal of Bayesian inference is learning, that is, using observations to update knowledge. In an ANOVA we want to learn about the candidate models $\boldsymbol{\mathcal{M}}$ and their condition-effect parameters $\boldsymbol{\beta}$. Returning to the example of the facial feedback experiment, we commonly specify two models. The null model describes the funniness ratings using a single grand average across all three conditions, effectively stating that there is no effect of pen position. The parameters of the null model are thus the average test score and the error variance. The alternative model describes the funniness ratings using an overall average and the effect of pen position; in other words, the means of the three condition are allowed to differ. Therefore, the alternative model has five parameters: the average funniness ratings across participants, the error variance, and for each of the three pen positions the magnitude of the effect.[1]

To start the learning process we need to specify prior beliefs about the plausibility of each model, $p(\boldsymbol{\mathcal{M}})$, and about the plausible parameters values $\boldsymbol{\beta}$ within each model, $p(\boldsymbol{\beta} \mid \boldsymbol{\mathcal{M}})$.[2] These prior beliefs are represented by *prior distributions*. Observing data $\mathcal{D}$ drives an update of beliefs, transforming the prior distribution over models and parameters to a joint *posterior distribution*, denoted $p(\boldsymbol{\beta}, \boldsymbol{\mathcal{M}} \mid \mathcal{D})$.[3] The updating factor –the change from prior to posterior beliefs– is determined by relative predictive performance for the observed data (Wagenmakers et al., 2016). As shown in Figure 1, the knowledge updating process forms a learning cycle, such that the posterior distribution after the first batch of data becomes the prior distribution for the next batch.

[1] Note that one of the four parameters, average funniness rating and the three condition effects, is redundant. That is, we can make identical predictions even when we fix one of the four parameters to zero.

[2] The symbol "$\mid$" may be read as "given the" or "conditional on".

[3] The joint posterior distribution describes the posterior probabilities for each parameter to take on values in a particular range. Simultaneously, it describes a discrete probability distribution over the models considered, where models with a higher plausibility after seeing the data have a higher probability. Furthermore, the joint posterior distribution also describes any interaction between parameters and models. For example, in the full model containing all covariates, the posterior distributions for the parameters estimates may be more concentrated around zero.

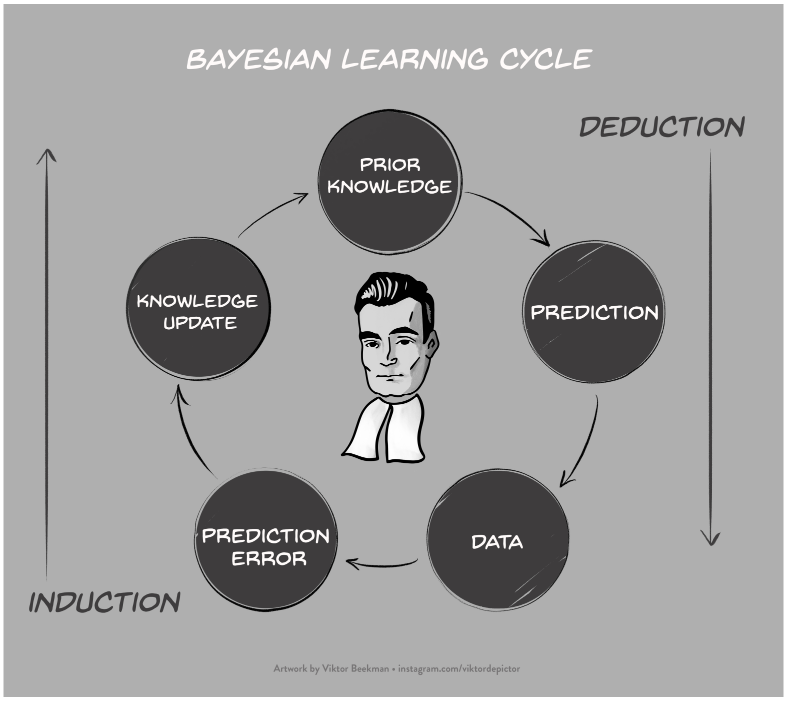

Figure 1: Bayesian learning can be conceptualized as a cyclical process of updating knowledge in response to prediction errors. The prediction step is deductive, and the updating step is inductive. For a detailed account see Jevons (1874/1913, Chapters XI and XII). Figure available at `BayesianSpectacles.org` under a CC-BY license.

Mathematically, the updating process is given by Bayes' rule:

$$\overbrace{p(\boldsymbol{\beta}, \boldsymbol{\mathcal{M}} \mid \mathcal{D})}^{\substack{\text{Joint posterior}\\ \text{distribution}}} = \overbrace{p(\boldsymbol{\mathcal{M}})}^{\substack{\text{Prior model}\\ \text{probability}}} \times \overbrace{p(\boldsymbol{\beta} \mid \boldsymbol{\mathcal{M}})}^{\substack{\text{Prior param.}\\ \text{probability}}} \times \overbrace{\frac{p(\mathcal{D} \mid \boldsymbol{\beta}, \boldsymbol{\mathcal{M}})}{p(\mathcal{D})}}^{\text{Updating factor}}. \tag{1}$$

This rule stipulates how knowledge about the relative plausibility of both models and parameters ought to be updated in light of the observed data. When the focus is on the comparison of two rival models, one generally considers only the model updating term. This term, commonly known as the *Bayes factor*, quantifies the relative predictive performance of the rival models, that is, the change in relative model plausibility that is brought about by the data (Etz & Wagenmakers, 2017; Jeffreys, 1939; Kass & Raftery, 1995; Wrinch & Jeffreys, 1921):

$$\underbrace{\frac{p(\mathcal{M}_1 \mid \mathcal{D})}{p(\mathcal{M}_0 \mid \mathcal{D})}}_{\text{Posterior model odds}} = \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}}_{\text{Prior model odds}} \times \underbrace{\frac{p(\mathcal{D} \mid \mathcal{M}_1)}{p(\mathcal{D} \mid \mathcal{M}_0)}}_{\substack{\text{Bayes factor} \\ \text{BF}_{10}}}. \tag{2}$$

When the Bayes factor BF_{10} equals 20, the observed data are twenty times more likely to occur under $\mathcal{M}_1$ than under $\mathcal{M}_0$ (i.e., support for $\mathcal{M}_1$ versus $\mathcal{M}_0$); when the Bayes factor BF_{10} equals $1/20$, the observed data are twenty times more likely to occur under $\mathcal{M}_0$ than under $\mathcal{M}_1$ (i.e., support for $\mathcal{M}_0$ versus $\mathcal{M}_1$); when the Bayes factor BF_{10} equals 1, the observed data are equally likely to occur under both models (i.e., neither model is supported over the other). Note that the Bayes factor is a comparison of two models and hence it is always a relative measure of evidence, that is, it quantifies the performance of one model relative to another.[4] Since the prior and posterior odds are both ratios of probabilities, the Bayes factor can be seen as an odds ratio that quantifies the change in belief from prior odds to posterior odds. The Bayes factor can be presented as BF_{10}, $p(\mathcal{D}|\mathcal{M}_1)$ divided by $p(\mathcal{D}|\mathcal{M}_0)$, or as its reciprocal BF_{01}, $p(\mathcal{D}|\mathcal{M}_0)$ over $p(\mathcal{D}|\mathcal{M}_1)$. Typically, $\text{BF}_{10} > 1$ is used to quantify evidence in favor of the alternative hypothesis, whereas $\text{BF}_{01} > 1$ is used to quantify evidence in favor of the null hypothesis. For instance, $\text{BF}_{10} = {}^1\!/_3$ can be interpreted as "the data are ${}^1\!/_3$ times more likely under $\mathcal{M}_1$ than under $\mathcal{M}_0$", but for a Bayes factor lower than 1 it is more intuitive to switch numerator and denominator and instead report the results as $\text{BF}_{01} = 3$, that is, "the data are 3 times more likely under $\mathcal{M}_0$ than under $\mathcal{M}_1$".

The Bayesian paradigm differs from the frequentist paradigm in at least four key aspects. First, evidence in favor of a particular model, quantified by a Bayes factor, is a *continuous* measure of support. Unlike the frequentist Neyman-Pearson decision rule (usually $p < 0.05$), there is no need to impose all-or-none Bayes factor cut-offs for accepting or rejecting a particular model. Moreover, the Bayes factor can discriminate between "absence of evidence" (i.e., nondiagnostic data that are predicted about equally well under both models, such that the Bayes factor is close to 1) and "evidence of absence" (i.e., diagnostic data that support the null hypothesis over the alternative hypothesis).

A second difference is that, in the Bayesian paradigm, knowledge about models $\boldsymbol{\mathcal{M}}$ and parameters $\boldsymbol{\beta}$ is updated simultaneously. Consequently, it is natural to account for model uncertainty by considering all models, but assigning more weight to those models that predicted the data relatively well. This procedure is known as Bayesian model averaging (BMA; Hinne, Gronau, van den Bergh, & Wagenmakers, 2019; Hoeting, Madigan, Raftery, & Volinsky, 1999; Jevons, 1874/1913; Jeffreys, 1939, p. 296; Jeffreys, 1961, p. 365). In contrast, many frequentist analyses first select a 'best' model and subsequently estimate its

[4] For a cartoon that explains the strength of evidence provided by a Bayes factor, see https://www.bayesianspectacles.org/lets-poke-a-pizza-a-new-cartoon-to-explain-the-strength-of-evidence-in-a-bayes-factor/

parameters, thereby neglecting model uncertainty and producing overconfident conclusions (Ch 7.4 Claeskens & Hjort, 2008).[5] Another benefit of BMA is that point estimates and uncertainty intervals can be derived without conditioning on a specific model. This way, model uncertainty is accounted for in point estimates and uncertainty intervals.

A third difference is that the Bayesian posterior distributions allow for direct probabilistic statements about parameters. For example, based on the posterior distribution of $\boldsymbol{\beta}$ we can state that we are 95% confident that the parameter lies between x and y. This range of parameter values is commonly known as a *95% credible interval*.[6] Similarly, we can consider any interval from a to b and quantify our confidence that the parameter falls in that specific range.

A fourth difference is that Bayesian inference automatically penalizes for complexity and thus favors parsimony (e.g., Berger & Jefferys, 1992; Jeffreys, 1961; Myung & Pitt, 1997). For instance, a model with a redundant covariate will make poor predictions. Consequently, the Bayes factor, which compares the relative predictive performance of two models, will favor the model without the redundant predictor over the model with the redundant predictor. Key is that, as the words suggest, the *predictive* performance is assessed using parameter values that are drawn from the *prior* distributions.

ANOVA

Traditionally, analysis of variance involves –as the name suggests– a comparison of variances. In the frequentist framework, the variance between each level of the categorical predictor is compared to the variance within the levels of the categorical predictor.

When the categorical predictor has no effect, the population variances between the levels equals the population variances within the levels, and the sample ratio of these variances is distributed according to a central F-distribution. Under the assumption that the null hypothesis is true, we may then calculate the probability of encountering a sample ratio of variances that is at least as large as the one observed – this then yields the much-maligned yet omnipresent p-value.

Instead, the Bayesian ANOVA contrasts the predictive performance of competing models (Rouder et al., 2016). In order to make predictions the model parameters need to be assigned prior distributions. These prior distributions could in principle be specified from subjective background knowledge, but here we follow Rouder, Morey, Speckman, and Province (2012) and use a default specification inspired by linear regression models, designed to meet general desider-

[5] Although uncommon, it is possible to average over the models in the frequentist framework. To do so, calculate for each model an information criterion such as AIC, and use a transformed version as model weights (Burnham & Anderson, 2002).

[6] Note the difference in interpretation compared to the frequentist 95% confidence interval: "if we repeat this experiment an infinite number of times and compute an infinite number of confidence intervals, then 95% of these intervals contain the true parameter value." See also Morey, Hoekstra, Rouder, Lee, and Wagenmakers (2016).

ata such as consistency and scale invariance (i.e., it does not matter whether the outcome variable is measured in seconds or milliseconds; see also Bayarri, Berger, Forte, & García-Donato, 2012; Liang, Paulo, Molina, Clyde, & Berger, 2008).

Assumptions

Before interpreting the results from an ANOVA, it is prudent to assess whether its main assumption holds, namely that the residuals are normally distributed. A common tool to assess the normality of the residuals is a Q-Q plot, which visualizes the quantiles of the observed residuals against the quantiles expected from a standard normal distribution. If the residuals are normally distributed then all the points in a Q-Q plot fall on the red line in Figure 2. In contrast to a frequentist ANOVA, where the residuals are point estimates, a Bayesian ANOVA provides a probability distribution for each residual. The uncertainty in the residuals can thus be summarized by 95% credible intervals. The left panel of Figure 2 shows an example where the larger quantiles lie away from the red line, displaying a substantial deviation from normality. The right panel of Figure 2 shows residuals that are more consistent with what is expected under a normal distribution.

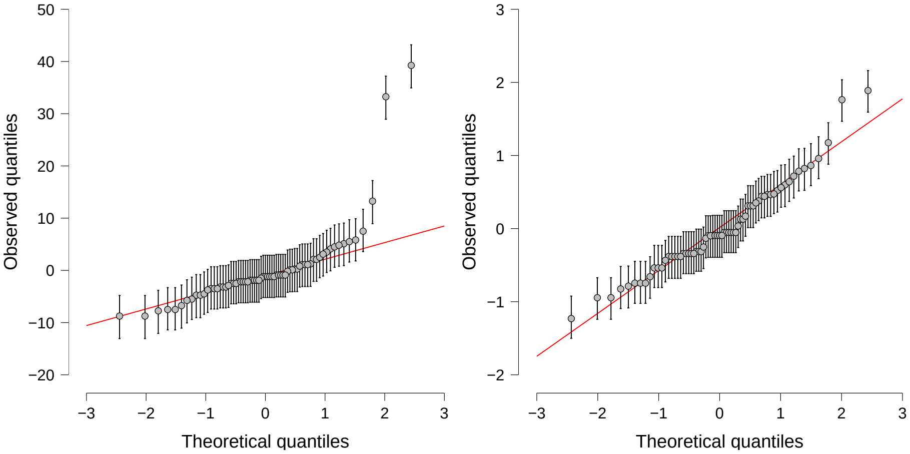

Figure 2: Q-Q plots of non-normally distributed residuals (left) and approximately normally distributed residuals (right). The vertical bars through each point represent the 95% central credible interval. If the data are perfectly normally distributed, all points fall on the red line. Note that the y-axis of the two panels has a different scale.

Introductory texts discuss additional ANOVA assumptions, most of which follow directly from the normality of the residuals. For some of these assumptions, violations can be difficult to detect visually in a Q-Q plot. An example is sphericity, which is specific to repeated measures ANOVA. One definition of sphericity is that the variance of all pairwise difference scores is equal. In the frequentist paradigm, this assumption is usually assessed using Mauchly's test (but see Tijmstra, 2018). Another example is homogeneity of variances, which implies that the residual variance is equal across all levels of the predictors. Homogeneity of variances can be assessed using Levene's test (Levene, 1961).

The following sections illustrate how to conduct and interpret a Bayesian ANOVA with JASP. JASP can be freely downloaded from `https://jasp-stats.org/download/`. Annotated `.jasp` files of the discussed analyses, data sets, and a step-by-step guide on conducting a Bayesian ANOVA in JASP are available at `https://osf.io/f8krs/`. We should stress that the current implementation of Bayesian ANOVA in JASP is based on the R package *BayesFactor* (Morey & Rouder, 2015) which is itself based on the statistical work by Rouder et al. (2012).

Example I: A Robot's Social Skills

Do people take longer to switch off a robot when it displays social skills? This question was studied by Horstmann et al. (2018) and we use their data to illustrate the key concepts of a Bayesian ANOVA. In the Horstmann et al. (2018) study, 85 participants interacted with a robot. Participants were told that the purpose of their interaction with the robot was to test a new algorithm. After two dummy tasks were completed, the instructor told the participants that they could switch off the robot if they wanted. The outcome variable was the time it took participants to switch off the robot. Here we analyze the log-transformed switch-off times since the Q-Q plot of the raw switch off times showed a violation of normality. Horstmann et al. (2018) manipulated two variables in a between-subjects design. First, they manipulated the robots' verbal responses to be either social (e.g., "Oh yes, pizza is great. One time I ate a pizza as big as me.") or functional (e.g., "You prefer pizza. This worked well. Let us continue."). Second, either the robot protested to being turned off (e.g., "No! Please do not switch me off! I am scared that it [*sic*] will not brighten up again!") or it did not. Therefore, the design of this study is a 2x2 between-subjects ANOVA. The data are shown in Figure 3.

Interpreting the Bayesian ANOVA

Model comparison The primary output from the JASP ANOVA is presented in Table 1, which shows the support that the data offer for each model under consideration. The left-most column lists all models at hand: four alternative models and one null model. The models are ordered by their predictive performance relative to the best model; this is indicated in the BF_{01} column, which shows the Bayes factor relative to the best model which features only the objection factor. For example, the data are about 73 times more likely under the model with only the robot's objection as a predictor than under the null

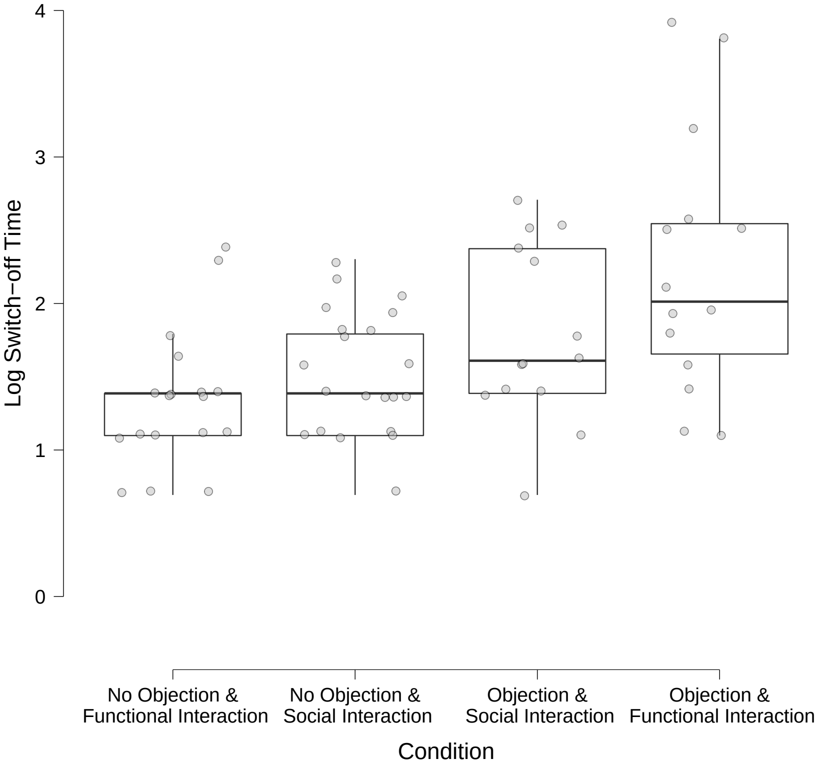

Figure 3: Observed log switch-off times for the data of Horstmann et al. (2018).

model. The prior model probability $P(\mathcal{M})$ is 0.2 for all models and the resulting posterior model probabilities are given by $P(\mathcal{M} \mid \mathcal{D})$. The $\text{BF}_{\mathcal{M}}$ column shows change from prior odds to posterior odds for each model. For example, for the best model with only the robot's objection as a predictor the change in odds is: ${}^{0.542}/_{(1-0.542)} \times {}^{(1-0.2)}/_{0.2} \approx 4.734$, which matches the output of Table 1. The right-most column provides an error percentage indicating the precision of the numerical approximations, which should not be too large.[7]

Table 1: Model comparison for all models under consideration for the data of Horstmann et al. (2018). The abbreviations 'O' and 'S' stand for the robot's objection and social interaction type, respectively. The term 'O * S' stands for the interaction between the two factors. The 'Model' column shows the predictors included in each model, the $P(\mathcal{M})$ column the prior model probability, the $P(\mathcal{M} \mid \mathcal{D})$ column the posterior model probability, the $\text{BF}_{\mathcal{M}}$ column the posterior model odds, and the BF_{01} column the Bayes factors of all models compared to the best model. The final column, 'error' is an estimate of the numerical error in the computation of the Bayes factor. All models are compared to the best model and are sorted from lowest Bayes factor to highest.

Model	$P(\mathcal{M})$	$P(\mathcal{M} \mid \mathcal{D})$	$\text{BF}_{\mathcal{M}}$	BF_{01}	error %
O	0.2	0.542	4.735	1.000	
O + S + O * S	0.2	0.303	1.736	1.791	2.770
O + S	0.2	0.146	0.682	3.719	1.323
Null model	0.2	0.007	0.030	73.373	0.000
S	0.2	0.002	0.009	252.495	0.005

Bayes factors are *transitive*, which means that if the model with only the robot's objection outpredicts the null model by a factor of a, and the null model outpredicts the model with only social interaction type by a factor of b, then the model with only the robot's objection will outpredict the model with only social interaction type by a factor of $a \times b$. Transitivity can be used to compute Bayes factors that may be of interest but are missing from the table. For example, the Bayes factor for the null model versus the model with only social interaction type can be obtained by dividing their Bayes factors against the best model: $252.495/73.373 \approx 3.441$ in favor of the null model.

Note that the Bayes factor is represented as BF_{01} in Table 1; predictive performance of the best model divided by the predictive performance for a particular model. Had we shown BF_{10}, we would have needed to take the reciprocal of the previous calculation to obtain the same result.

[7] In many situations, error percentages below 20% are acceptable. If the error is 20%, then a Bayes factor of 10 can fluctuate between 8 and 12. Because Bayes factors between 8 and 12 lead to the same qualitative conclusion, this amount of numerical error is not problematic (see also Jeffreys, 1961, Appendix B). When the error percentage is deemed too high, the number of samples can be increased to reduce the error percentage at the cost of longer computation time. For more information, see van Doorn et al. (2019).

Analysis of Effects The previous section compared all available models. However, as the number of predictors increases, the number of models quickly grows too large to consider each model individually.[8] Rather than studying the results for each model individually, it is possible to average the results from Table 1 over all models, that is, compute the model-averaged results. This produces Table 2, which shows for each predictor the prior and posterior inclusion probabilities, and the inclusion Bayes factor. A prior inclusion probability is the probability that a predictor is included in the model before seeing the data and is computed by summing up the prior model probabilities of all models which contain that predictor. A posterior inclusion probability is the probability that a predictor is included in the model after seeing the data and is computed by summing up the posterior model probabilities of all models which contain that predictor. The inclusion Bayes factor quantifies the change from prior inclusion odds to posterior inclusion odds and can be interpreted as the evidence in the data for including a predictor. For example, Table 2 shows that the data are about 68.6 times more likely under the models that include the robot's objection than under the models without this predictor.

Table 2: Results from averaging over the models in Table 1. The abbreviations 'O' and 'S' stand for the robot's objection and social interaction type respectively. The first column denotes each predictor of interest, the column $P(\text{incl})$ shows the prior inclusion probability, $P(\text{incl} \,|\, \mathcal{D})$ shows the posterior inclusion probability, and $\text{BF}_{\text{Inclusion}}$ shows the inclusion Bayes factor.

Effects	$P(\text{incl})$	$P(\text{incl} \,\vert\, \mathcal{D})$	$\text{BF}_{\text{Inclusion}}$
O	0.6	0.990	68.558
S	0.6	0.445	0.535
O * S	0.2	0.293	1.659

Although model-averaged results are straightforward to obtain, their interpretation requires special attention when interaction effects are concerned. In JASP, models are excluded from consideration when they violate the *principle of marginality*, that is, they feature an interaction effect but lack the constituent main effects (for details see Nelder, 1977). This model exclusion rule means that the active model set is not balanced. For example, in Table 2 the inclusion odds for the interaction 'O * S' is obtained by comparing four models without the interaction effect against the one model with the interaction effect. As an alternative, Sebastiaan Mathôd has suggested to compute inclusion probabilities for "matched" models only.[9] What this means is that all models with the interaction effect are compared to models with the same predictors except for the interaction effect. For example, the model with an interaction effect between

[8] In general, given p predictors there are 2^p models to consider. If interaction effects are considered, the model space grows even faster.

[9] See also `https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp`.

'O * S' in Table 2 is compared against the model with the main effects of 'O' and 'S', but not against any other models. To compute inclusion probabilities for main effects, models that feature interaction effects composed of these main effects are not considered. These models are excluded because they cannot be matched with models that include the interaction effect but not the main effect, since those violate the principle of marginality. Note that without interaction effects, the matched and not matched inclusion probabilities are the same.

Table 3 shows the inclusion probabilities and inclusion Bayes factor obtained by only considering matched models. Comparing Table 3 to Table 2, the prior inclusion probability of the main effects decreased because these are based on one model fewer. The posterior inclusion probabilities of the main effects decreased but that of the interaction effect increased. The inclusion Bayes factor, the evidence in the data for including a predictor, provides slightly more evidence for including the main effect of the robot's objection and the interaction effect, and somewhat more evidence for excluding the main effect of the social interaction type.

Table 3: Results from averaging over the models in Table 1 but only considering "matched" models (see text for details). The abbreviations 'O' and 'S' stand for the robot's objection and social interaction type respectively. The first column denotes each predictor of interest, the column $P(\text{incl})$ shows the prior inclusion probability, $P(\text{incl} \mid \mathcal{D})$ shows the posterior inclusion probability, and $\text{BF}_{\text{Inclusion}}$ shows the inclusion Bayes factor.

Effects	$P(\text{incl})$	$P(\text{incl} \mid \mathcal{D})$	$\text{BF}_{\text{Inclusion}}$
O	0.4	0.6872	72.76
S	0.4	0.1524	0.28
O * S	0.2	0.3033	2.018

Parameter Estimates After establishing which predictors are relevant we can investigate the magnitude of the relations by examining the posterior distributions. Table 4 summarizes the model-averaged posterior distributions of each level (β_j), using four statistics: the posterior mean, the posterior standard deviation, and the lower and upper bound of the 95% central credible interval. The symmetry in the estimates is a consequence of the sum-to-zero constraint, that is, the posterior mean of O-Yes = $-1 \times$ the posterior mean of O-No = 0.265. Table 4 shows that the effect of objection is about 0.265 (95% CI [0.111, 0.418]). A posterior estimate for the observed log response time of a particular group, say the condition where the robot did not object, can be obtained by adding the posterior mean of the intercept (i.e., the grand mean), 1.724, to the posterior mean of the no-objection condition, -0.265, which yields 1.459.[10]

[10] This calculation is valid only for the posterior means, not for the other posterior summaries.

Table 4: Summary of the marginal model averaged posterior distributions. Posteriors are summarized using mean, standard deviation, and 95% central credible intervals (CI).

				95% CI	
Predictor	Level	Mean	SD	Lower	Upper
Intercept		1.724	0.077	1.569	1.877
O	Yes	0.265	0.077	0.111	0.418
	No	−0.265	0.077	−0.420	−0.113
S	Functional	−0.044	0.071	−0.186	0.097
	Social	0.044	0.071	−0.098	0.185
O * S	Yes & Social	−0.132	0.072	−0.278	0.008
	Yes & Functional	0.132	0.072	−0.009	0.276
	No & Social	0.132	0.072	−0.009	0.276
	No & Functional	−0.132	0.072	−0.278	0.008

To summarize, the Bayesian ANOVA revealed that the robot's objection almost certainly had an effect on switch-off time ($\text{BF}_{\text{Inclusion}} = 68.558$). We also learned that the data are not sufficiently informative to allow a strong conclusion about the effect of the robot's social interaction type ($\text{BF}_{\text{Inclusion}} = 0.535$) or about an interaction effect between objection and social interaction type ($\text{BF}_{\text{Inclusion}} = 1.659$).

Example II: Post Hoc Tests on the Houses of Hogwarts

After executing an ANOVA and finding strong evidence that a particular predictor relates to the outcome variable, a common question arises: "Which levels of the predictor deviate from one another?". As an illustration, consider the data from Jakob, Garcia-Garzon, Jarke, and Dablander (2019) where 847 participants filled out a 'sorting hat' questionnaire that determined their assignment to one of the four Houses of Hogwarts from the Harry Potter books: Gryffindor, Hufflepuff, Ravenclaw, or Slytherin.[11] Subsequently, participants filled out the dark triad questionnaire (Jones & Paulhus, 2014) that was used to derive the outcome variable: Machiavellism.

In this example, there is only one categorical predictor: The House of Hogwarts a participant was assigned to. If we compare the model with this predictor to the null model, we find overwhelming evidence for the alternative ($\text{BF}_{10} = 6.632 \times 10^{18}$). This is a clear indication that Machiavellism differs between the members of the four houses. However, this result does not indicate the houses responsible for the difference. To address that question, we need a post hoc test.

For ANOVA models, the main component of a post hoc test is a t-test on all pairwise combinations of a predictor's levels. For a Bayesian ANOVA, the

[11] The raw data and original analyses can be found at https://osf.io/rtf74/.

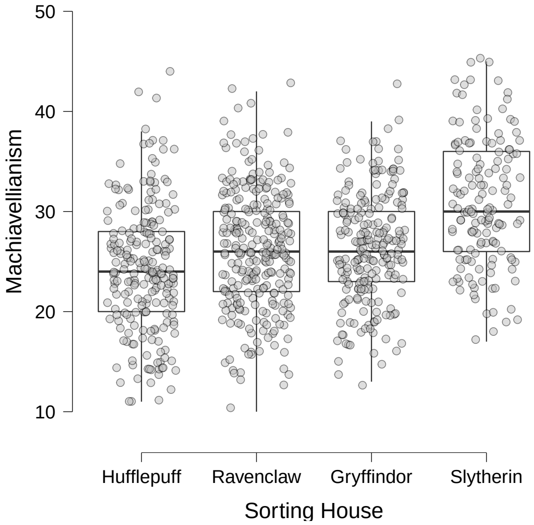

Figure 4: Observed Machiavellism scores for each of the four Houses of Hogwarts.

main component is the Bayesian t-test. Table 5 shows the Bayesian post hoc tests for the sorting hat data. As with frequentist inference, Bayesian post hoc tests are subject to a multiple comparison problem. To control for multiplicity, we follow the approach discussed in Westfall (1997) which is an extension of the approach of Jeffreys (1938); for an overview of Bayesian methods correcting for multiplicity see for instance de Jong (2019).

Westfall's approach relates the overall null hypothesis $p(\mathcal{H}_0)$ that all condition means are equal to each comparison between two condition means. That way, the prior probability of the overall null hypothesis can be adjusted to correct for multiplicity and this influences each individual comparison. The procedure to relate the overall null hypothesis to each comparison is described below.

A condition mean μ_i is either equal to the grand mean μ with probability τ, or μ_i is drawn from a continuous distribution with probability $1 - \tau$. It is key that this distribution is continuous because two values drawn from a continuous distribution are never exactly equal. Thus, the probability that two condition means μ_i and μ_j are equal is $p(\mu_i = \mu_j) = p(\mu_i = \mu) \times p(\mu_j = \mu) = \tau^2$. From this, the probability of the null hypothesis that all J condition means are equal follows: $p(\mathcal{H}_0) = p(\mu_1 = \mu_2 = \cdots = \mu_J) = p(\mu_1 = \mu) \times p(\mu_2 = \mu) \times \cdots \times p(\mu_J = \mu) = \tau^J$. Solving for τ, we obtain $\tau = p(\mathcal{H}_0)^{1/J}$. Thus, the prior probability that two specific magnitudes are equal can be expressed in terms of the prior probability that all magnitudes are equal, that is $p(\mu_i = \mu_j) = \tau^2 = p(\mathcal{H}_0)^{2/J}$. For example, imagine there are four conditions ($J = 4$) and the prior probability that all condition means are equal is 0.5. Then, the prior probability that two conditions means are equal is: $p(\mu_1 = \mu 2) = \sqrt{0.5}$. The prior odds are then ${}^{(1-\sqrt{0.5})}/_{\sqrt{0.5}} \approx 0.414$.

In sum, the Westfall approach involves, as a first step, Bayesian t-tests for all pairwise comparisons, which provides the unadjusted Bayes factors. In the next step, the prior model odds are adjusted by fixing the overall probability of no effect to 0.5. The adjusted prior odds and the Bayes factor are then used to calculate the adjusted posterior odds.

Table 5 shows the results for the post hoc tests of the sorting hat example. The adjusted posterior odds show (1) evidence (i.e., odds of about 16) that Machiavellism differs between Hufflepuff and Ravenclaw; (2) evidence (i.e., odds of about 27) that Machiavellism differs between Gryffindor and Hufflepuff; (3) overwhelming evidence (i.e., odds of about 1.04×10^9, 5.43×10^{16}, and 5.30×10^9) that Machiavellism differs between Gryffindor and Slytherin, between Hufflepuff and Slytherin, and between Ravenclaw and Slytherin, respectively; (4) evidence (i.e, odds of ${}^1/_{0.0432} \approx 23$) that Machiavellism of Gryffindor and Ravenclaw is the same.

Table 5: Post hoc test for the Sorting House data. The first two columns indicate the houses being compared, the third and fourth column indicate the adjusted prior model odds and posterior model odds respectively, and the fifth column indicates the uncorrected Bayes factor in favor of the alternative hypothesis that the magnitudes differ. The final column shows the numerical error of the Bayes factor computation.

Level 1	Level 2	Prior Odds	Posterior Odds	$\mathrm{BF}_{10,\mathrm{U}}$	error %
Gryffindor	Hufflepuff	0.414	27.2	65.6	5.73×10^{-5}
Gryffindor	Ravenclaw	0.414	0.0432	0.104	9.56×10^{-5}
Gryffindor	Slytherin	0.414	1.04×10^9	2.50×10^9	3.94×10^{-16}
Hufflepuff	Ravenclaw	0.414	15.5	37.3	7.57×10^{-8}
Hufflepuff	Slytherin	0.414	5.43×10^{16}	1.31×10^{17}	3.35×10^{-23}
Ravenclaw	Slytherin	0.414	5.30×10^9	1.28×10^{10}	6.36×10^{-16}

Now that we know which Houses differ, the next step is to assess the magnitude of each House of Hogwarts on Machiavellism score. Rather than examining a table that summarizes the marginal posteriors, we plot the model averaged posteriors for each house in Figure 5. Clearly, Slytherin scores higher on Machiavellism than the other Houses whereas Hufflepuff scores lower on Machiavellism than the other Houses. Table 6 in the appendix shows the parameters estimates of the marginal posterior effects for each house.

Concluding Comments

The goal of this paper was to provide guidance for practitioners on to conduct a Bayesian ANOVA in JASP and interpret the results. Although the focus was

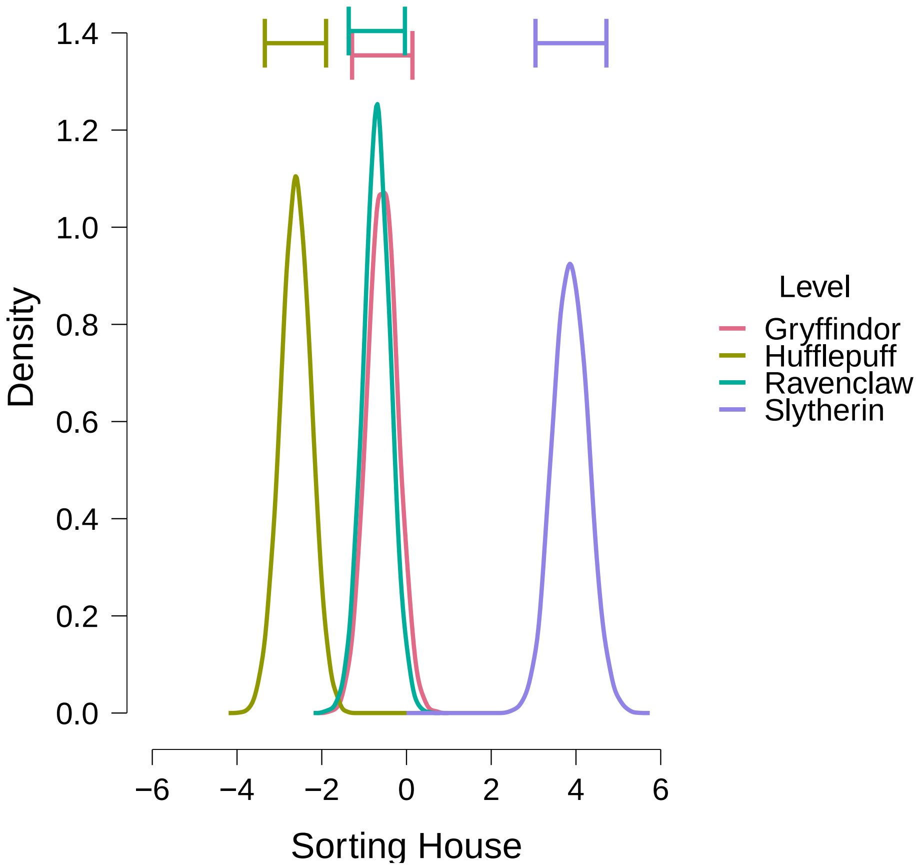

Figure 5: Posterior distributions of the effect of each House of Hogwarts on Machiavellism. Slytherin scores higher on Machiavellism than the other Houses whereas Hufflepuff scores lower on Machiavellism than the other Houses. The horizontal error bars above each density represent 95% credible intervals.

on ANOVAs with categorical predictors, JASP can also handle ANOVAs with additional continuous predictors. The appropriate analysis then becomes an analysis of covariance (ANCOVA) and all concepts explained here still apply. For a general guide on reporting Bayesian analyses see van Doorn et al. (2019).

As with all statistical methods, the Bayesian ANOVA comes with limitations and caveats. For instance, when the model is severely misspecified and the residuals are non-normally distributed, the results from a standard ANOVA –whether Bayesian or frequentist– are potentially misleading and should be interpreted with care. In such cases, at least two alternatives may be considered. The first alternative is to consider a rank-based ANOVA such as the Kruskal–Wallis test

(Kruskal & Wallis, 1952). This test depends only on the ordinal information in the data and hence does not make strong assumptions on how the data ought to be distributed. The second alternative is to specify a different distribution for the residuals. Using software for general Bayesian inference such as Stan (Carpenter et al., 2017) or JAGS (Plummer, 2003), it is relatively straightforward to specify any distribution for the residuals. However, this approach requires knowledge about programming and statistical modeling and is likely to be computationally intensive. Another limitation of the Bayesian ANOVA is that, especially in more complicated designs, it is not straightforward to intuit what knowledge the prior distributions represent.

Some limitations are specific to JASP. Currently, it is not possible to use post hoc tests to examine whether the contribution of a level differs from zero, that is, to test whether a specific level deviates from the grand mean. It is also not possible to handle missing values in any other way than list-wise deletion. Another limitation relates to sample size planning. The typical planning process involves a frequentist power analysis which provides the sample size needed to achieve a certain rate of correctly detecting a true effect of a prespecified magnitude. A Bayesian equivalent of power analysis is Bayes factor design analysis (BFDA; e.g., Schönbrodt & Wagenmakers, 2018). In a sequential design, BFDA produces the expected sample sizes required to reach a target level of evidence (i.e., a target Bayes factor). In a fixed-n design, BFDA produces the expected levels of evidence, given a specification of the magnitude of the effect. At the moment of writing BFDA has not been implemented in JASP; an accessible tutorial and a Shiny app are provided by Stefan, Gronau, Schönbrodt, and Wagenmakers (2019).

We believe that the Bayesian ANOVA provides a perspective on the analysis of factorial designs that can fruitfully supplement or even supplant the currently dominant frequentist ANOVA. The epistemic advantages of the Bayesian paradigm are well known (e.g., Jeffreys, 1961; Wagenmakers et al., 2018) but in order to be adopted in research practice it is essential for the methodology to be implemented in an easy-to-use software package such as JASP. In addition to the software, however, practitioners also require guidance on how to interpret the results, which was the main purpose of this paper. In general, we hope that the increased use of the Bayesian ANOVA will stimulate the methodological diversity in the field, and that it will become more standard to examine the robustness of frequentist conclusions by comparing them to the Bayesian alternative.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*, 305–307.

Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, *40*, 1550–1577.

Berger, J. O., & Jefferys, W. H. (1992). The application of robust Bayesian analysis to hypothesis testing and Occam's razor. *Journal of the Italian Statistical Society*, *1*, 17–32.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information–theoretic approach (2nd ed.)*. New York: Springer Verlag.

Burton, P. R., Gurrin, L. C., & Campbell, M. J. (1998). Clinical significance not statistical significance: A simple Bayesian alternative to p values. *Journal of Epidemiology & Community Health*, *52*(5), 318–323.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

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

de Jong. (2019). A Bayesian approach to the correction for multiplicity. *Unpublished Master Thesis*. Retrieved from https://doi.org/10.31234/osf.io/s56mk

Dienes, Z., & McLatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*(1), 207–218.

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*, 313–329.

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*, *292*(6522), 746–750.

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2019). A conceptual introduction to Bayesian model averaging. *manuscript submitted for publication*.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.

Horstmann, A. C., Bock, N., Linhuber, E., Szczuka, J. M., Straßmann, C., & Krämer, N. C. (2018). Do a robot's social skills and its objection discourage interactants from switching the robot off? *PLoS ONE*, *13*(7), e0201581.

Jakob, L., Garcia-Garzon, E., Jarke, H., & Dablander, F. (2019). The science behind the magic? The relation of the Harry Potter "sorting hat quiz" to personality and human values. *manuscript submitted for publication*.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving*, *7*, 2-9.

JASP Team. (2019). *JASP (Version 0.9.2)[Computer software]*. Retrieved from https://jasp-stats.org/

Jeffreys, H. (1938). Significance tests when several degrees of freedom arise simultaneously. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *165*, 161–198.

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: Oxford University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Jevons, W. S. (1874/1913). *The principles of science: A treatise on logic and scientific method*. London: MacMillan.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (sd3) a brief measure of dark personality traits. *Assessment*, *21*(1), 28–41.

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

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Levene, H. (1961). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling* (pp. 279–292). Stanford, California: Stanford University Press.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

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

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*, 235–245.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor 0.9.12-4.2.* Comprehensive R Archive Network. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Nathoo, F. S., & Masson, M. E. J. (2016). Bayesian alternatives to null–hypothesis significance testing for repeated–measures designs. *Journal of*

Mathematical Psychology, *72*, 144–157.

Nelder, J. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society: Series A (General)*, *140*(1), 48–63.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing.* Vienna, Austria.

Rouder, J. N., Engelhardt, C. R., McCabe, S., & Morey, R. D. (2016). Model comparison in ANOVA. *Psychonomic Bulletin & Review*, *23*, 1779–1786.

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

Savalei, V., & Dunn, E. (2015). Is the call to abandon $p-$values the red herring of the replicability crisis? *Frontiers in Psychology*, *6*, 245.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042-1058. doi: 10.3758/s13428-018-01189-8

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*, 768–777.

Tijmstra, J. (2018). Why checking model assumptions using null hypothesis significance tests does not suffice: A plea for plausibility. *Psychonomic Bulletin & Review*, *25*, 548–559.

van der Schoot, R., Winter, S., Ryan, O., Zondervan Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian papers in psychology: The last 25 years. *Psychological Methods*, *22*, 217–239.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (Eds.). (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., ... Wagenmakers, E.-J. (2019). The JASP guidelines for conducting and reporting a Bayesian analysis. *manuscript submitted for publication*. Retrieved from psyarxiv.com/yqxfr

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, *92*(437), 299–306.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.

Parameter Estimates for the Sorting Hat Data

Table 6: Parameter estimates for each of the houses in the data of Jakob et al. (2019). The interpretation of each column is identical to that of Table 4.

				95% CI	
Predictor	Level	Mean	SD	Lower	Upper
Intercept		26.923	0.215	26.46	27.337
Sorting house	Gryffindor	−0.568	0.357	−1.28	0.140
	Hufflepuff	−2.610	0.360	−3.34	−1.898
	Ravenclaw	−0.696	0.330	−1.36	−0.037
	Slytherin	3.874	0.418	3.04	4.719