

Running head: BAYESIAN ANALYSIS

Bayesian Analysis of Factorial Designs

Jeffrey N. Rouder

University of Missouri

Richard D. Morey

Cardiff University

Josine Verhagen

University of Amsterdam

April R. Swagman

University of Missouri

Eric-Jan Wagenmakers

University of Amsterdam

Abstract

This paper provides a Bayes factor approach to multiway ANOVA that allows researchers to state graded evidence for effects or invariances as determined by the data. ANOVA is conceptualized as a hierarchical model where levels are clustered within factors. The development is comprehensive in that it includes Bayes factors for fixed and random effects and for within-subjects, between-subjects, and mixed designs. Different model construction and comparison strategies are discussed, and an example is provided. We show how Bayes factors may be computed with `BayesFactor` package in R and with the JASP statistical package.

Bayesian Analysis of Factorial Designs

The importance of factorial designs and the associated ANOVA tests to experimental psychologists is difficult to overstate. Indeed, much experimental work relies on factorial designs and ANOVA analyses. Conventional ANOVA allows researchers to make decisions about the presence of main effects and interactions among factors. In this paper, we present a Bayesian approach to ANOVA. We advocate the use of Bayes factors which provide a principled approach for stating evidence from data for competing models (Edwards, Lindman, & Savage, 1963; Jeffreys, 1961; Berger & Sellke, 1987; Morey, Romeijn, & Rouder, 2013; Raftery, 1995).

There are two pragmatic advantages of Bayes factors for researchers: First, Bayes factors may be used to state evidence for the absence of an effect, which is better termed an *invariance*. The ability to state evidence for an invariance contrasts favorably to significance testing where researchers state a lack of evidence for an effect. Second, with Bayes factors, researchers need not make dichotomous reject and fail-to-reject decisions. Instead, they may state the Bayes factor as a graded measure of evidence without recourse to a dichotomous decision. Discussion and development of Bayesian model comparison techniques is a vast field. The comparison among different Bayesian and frequentist approaches has been made repeatedly. We do not wish to revisit this well trodden ground. Interested readers are referred to Berger & Berry (1988), Edwards et al. (1963), Efron (2005), Gelman, Shor, Bafumi, & Park (2007), Jaynes (2003), Kruschke (2011) and Wagenmakers (2007) among others.

The goal here is to provide a Bayes-factor approach to ANOVA, and this paper serves as a companion paper to our previous development in Rouder, Morey, Speckman, & Province (2012). Our previous development was technical in nature and is inaccessible to many psychologists, particularly to those who may not be familiar with Bayesian

methods. Moreover, the previous development leaves many practical issues unresolved, such as how to perform computations in practice and which models to compare. In this paper we provide a more relaxed and accessible treatment, providing an enhanced focus on the practical aspects of modeling and model comparison. We show how analysis is implemented in Morey & Rouder's (2014) `BayesFactor` package for R, and in Love et al.'s (n.d.) new open-source graphical statistical package, `JASP`. We discuss the types of comparisons to be made, and how to use multiple models within a modeling hierarchy to quantify evidence for targeted contrasts.

Bayesian Probability and Bayes Factors

This section provides a brief overview of Bayesian updating. The heart of statistical analysis is the concept of probability. In the frequentist conceptualization, probability is defined as a limit. It is the proportion of events in infinitely many repeated samples. Bayesians, however, define probability as a measure of subjective belief. For example, in the run up to the 2012 U.S. Presidential Election, statistician Nate Silver stated one-week out that Obama has a .75 chance of winning. This statement is understood as a Bayesian expression of belief—Silver believes that Obama will probably win. A probability statement expresses the subjective degree of uncertainty about given propositions.

Bayesian Updating

The focus of Bayesian analysis is not what to believe but rather how to change beliefs in light of data. Consider for example an unfair coin for which we would like to estimate the probability of landing heads. The coin is to be flipped 10 times. Before these flips, the analyst's beliefs are expressed as a probability distribution. For instance, the analyst may believe that all probabilities are equally likely, and this belief is expressed by the uniform distribution on probability shown in Figure 1A. This distribution is called the *prior distribution* on the parameter because it is specified before the data are considered.

Let θ denote the probability of a heads, and let $\pi(\theta)$ denote the prior distribution. For the prior in Figure 1A, $\pi(\theta) = 1$ for valid values of θ ($0 \leq \theta \leq 1$).

Now the coin is flipped, and say 7 heads are observed in 10 flips. The analyst uses a particular rule—Bayes’ rule—to rationally update her beliefs. The resulting distribution is shown, and it now reflects the new information. This updated distribution is called the *posterior distribution*, and it expresses beliefs conditional on observing specific data, seven heads in this case. The posterior is denoted $\pi(\theta|Y)$ where Y is the number of heads in 10 flips. Figure 1B shows a different set up. Here the analyst *a priori* believed that the coin was biased toward tails and observes 70 heads of 100 flips. Even though the prior was biased toward tails, the preponderance of heads from data has led to a large-scale revision of beliefs of θ near the value of .7. As more data are observed, the prior beliefs have less influence on the beliefs about θ .

Bayes’ rule, which provides for an optimal updating of beliefs (see Jeffreys, 1961), is

$$\pi(\boldsymbol{\theta} | \mathbf{Y}) = U(\boldsymbol{\theta}, \mathbf{Y}) \times \pi(\boldsymbol{\theta}).$$

where $\boldsymbol{\theta}$ are parameters and \mathbf{Y} are data. The term $U(\boldsymbol{\theta}, \mathbf{Y})$ is the updating factor that tells the analyst how to update the prior, $\pi(\boldsymbol{\theta})$, to reach the posterior, $\pi(\boldsymbol{\theta}|\mathbf{Y})$. The updating factor is:

$$U(\boldsymbol{\theta}, \mathbf{Y}) = \frac{f(\mathbf{Y} | \boldsymbol{\theta})}{f(\mathbf{Y})}.$$

The numerator of the updating factor, $f(\mathbf{Y}|\boldsymbol{\theta})$, is the likelihood function, that is, it is the probability of the observed data expressed as a function of the parameters. The denominator, $f(\mathbf{Y})$, is the probability of the data, which does not have a frequentist analog. Because parameters are viewed as distributions rather than fixed constants, they

may be conditioned upon. With this conditioning, the denominator may be re-expressed as

$$f(\mathbf{Y}) = \int f(\mathbf{Y} | \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

This integral is an expectation value, and the denominator is the weighted average of the likelihood function taken across all parameter values. Hence, the updating factor is the ratio of the likelihood at the evaluated parameter value to the average likelihood across all parameter values. The updating factor emphasizes parameter values that are more likely than the average and attenuates parameter values that are less likely.

Figure 1 shows how updating works. Panel C shows the likelihood and averaged likelihood (the flat line) for 7 heads in 10 flips. Let's consider a single point: $\theta = .73$. The likelihood value at that point is .261, which is 2.87 times greater than the average value of .0909. The updating factor at this point is therefore 2.87, and the posterior is 2.87 greater than the prior for $\theta = .73$. Panel D shows the same for 70 heads in 100 flips. The likelihood of any parameter value is reduced, but so is the average. The likelihood at $\theta = .73$ is .069 which is about 25 times greater than the average value of .0028. Hence, the posterior is 25 times greater than the prior at this point. The same updating works to lower the plausibility of estimates. For example, consider $\theta = .4$. For 7 heads in 10 flips, the likelihood is about half the average, and the posterior is lowered by this factor. For 70 heads in 100 flips, the attenuation is nearly perfect as .4 is incompatible with the observed data.

Bayes' rule may be written as a single equation. The most useful form here is

$$\pi(\boldsymbol{\theta} | \mathbf{Y}) = \frac{f(\mathbf{Y} | \boldsymbol{\theta})}{f(\mathbf{Y})} \times \pi(\boldsymbol{\theta}). \quad (1)$$

Bayesian Model Comparison

The preceding usage of Bayes' rule showed how beliefs about parameters in models should be updated. Yet, in ANOVA and in much of analysis in the psychological sciences, we are interested in comparing models themselves. Because Bayesians view probability as a measure of subjective belief, it is permissible to hold beliefs as probabilities on models themselves. Let \mathcal{M}_A be our previous model on θ , the probability of heads on a flip of a certain coin. The number of heads may be modeled as:

$$Y | \theta \sim \text{Binomial}(\theta, N)$$

$$\theta \sim \text{Uniform}(0, 1)$$

In the preceding usage, we discussed role of $f(Y)$, the probability of data, and it served as the denominator of the updating factor. The probability of data can also be viewed as the *prediction* of the model about where the data should be observed. We had calculated it as $f(Y) = \int f(Y|\theta)\pi(\theta)d\theta$, and noted that $f(Y) = .0909$ when $Y = 7$ for 10 flips (see Figure 1C). In fact, $f(Y) = .0909 = 1/11$ for all values of Y for this uniform prior. Figure 2A shows the probability of data for all the outcomes of the ten-flip experiment. These are literally the predictions of the model, that is, the probability of where data should fall before data have been observed. Notice that the distribution is proper and the probabilities sum to 1: the data have to occur somewhere.

Suppose we wish to compare this unfair-coin model to a fair-coin model, denoted \mathcal{M}_B . The fair-coin model is denoted

$$Y \sim \text{Binomial}(.5, N)$$

The predictions of the model are given by $\binom{N}{y}(.5)^{10}$ and shown in Figure 2B. With these

predictions we are ready to make model comparison. Had we observed a low number of heads, say $Y = 2$, we could note that the observation is better predicted under Model \mathcal{M}_A (with probability .0909) than under \mathcal{M}_B (with probability .044). The ratio is 2-to-1 in favor of the general model, and this value serves as an index of the predictive accuracy of one model relative to the other. Conversely, had we observed a moderate number, say $Y = 5$, then the observation is better predicted by the fair-coin model, Model \mathcal{M}_B than by the more general Model \mathcal{M}_A . The ratio here is 2.7-to-1 in favor of the fair-coin hypothesis. These ratios, called *Bayes factors*, are shown for all outcomes in Figure 2C.

An application of Bayes' rule shows the important role of the Bayes factor. In Bayesian analysis, beliefs may be placed on models themselves. Prior beliefs, those held before considering data, for Models \mathcal{M}_A and \mathcal{M}_B are denoted $\Pr(\mathcal{M}_A)$ and $\Pr(\mathcal{M}_B)$, respectively. Bayes' rule is used to update these beliefs:

$$\Pr(\mathcal{M}_A|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathcal{M}_A)\Pr(\mathcal{M}_A)}{f(\mathbf{Y})}, \quad \Pr(\mathcal{M}_B|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathcal{M}_B)\Pr(\mathcal{M}_B)}{f(\mathbf{Y})},$$

where $\Pr(\mathcal{M}_A|\mathbf{Y})$ and $\Pr(\mathcal{M}_B|\mathbf{Y})$ are posterior beliefs. It is convenient to express these posterior beliefs on models as odds:

$$\frac{\Pr(\mathcal{M}_A|\mathbf{Y})}{\Pr(\mathcal{M}_B|\mathbf{Y})} = \frac{f(\mathbf{Y} | \mathcal{M}_A)}{f(\mathbf{Y} | \mathcal{M}_B)} \times \frac{\Pr(\mathcal{M}_A)}{\Pr(\mathcal{M}_B)}. \quad (2)$$

The term $f(\mathbf{Y} | \mathcal{M}_A)/f(\mathbf{Y} | \mathcal{M}_B)$ is the *Bayes factor*, and it describes the updating factor or the extent to which the data cause revision in belief (Kass & Raftery, 1995). We denote the Bayes factor by B_{AB} , where the subscripts indicate which two models are being compared. A Bayes factor of $B_{AB} = 10$ means that prior odds should be updated by a factor of 10 in favor of model \mathcal{M}_A ; likewise, a Bayes factor of $B_{AB} = .1$ means that prior odds should be updated by a factor of 10 in favor of model \mathcal{M}_B . Bayes factors of $B_{AB} = \infty$ and $B_{AB} = 0$ correspond to infinite support from the data for one model over

the other with the former indicating infinite support for model \mathcal{M}_A and the later indicating infinite support for model \mathcal{M}_B . The deep meaning in Bayes' rule is that the updating factor on beliefs, the evidence in data, is the relative predictive accuracy of the models. Because Bayes factors index the evidence for models from data, their use has been recommended repeatedly for psychological research (e.g., Edwards et al., 1963; Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007).

The distinction between Bayes factors—the effect of the data to change beliefs—and the posterior odds—the resulting beliefs themselves—is in our view intellectually appealing. Two researchers may start with different prior odds. For example, Daryl Bem, a proponent of ESP (Bem, 2011), may hold the ESP hypothesis to be as likely as a no-ESP hypothesis. We, however, disagree, and have stated lower prior beliefs in the ESP hypothesis than in the no-ESP hypothesis because there is no known mechanism for ESP and the existence of ESP would require radical revisions to the laws of physics and biology (Rouder & Morey, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Even though we may not come to an agreement with Bem over prior odds and consequently will not agree on posterior odds, we may nonetheless agree on the evidence for these models from the data, and, consequently, agree how to update our respective beliefs. Rouder & Morey (2012) recommend researchers report Bayes factors as evidence. They may additionally provide prior and posterior odds as value-added guidance, perhaps from consideration of context.

Bayes factor is one of several Bayesian model comparison approaches. Alternatives include inference by observation of critical posterior quantities (Aitkin, 1991; Gelman & Shalizi, 2013; Kruschke, 2012) and selection with the deviance information criteria (DIC, Spiegelhalter, Best, Carlin, & van der Linde, 2002). As previously mentioned, the contrasts and comparisons among these methods is well represented in the literature and interested readers may go to these papers and subsequent commentaries.

ANOVA as a Hierarchical Model

We demonstrate here our Bayesian approach to ANOVA. Our modeling framework is influenced by Gelman (2005), who conceptualizes ANOVA as a hierarchical model in which effects are clustered within factors. We outline this approach, beginning with a brief overview of Bayesian hierarchical modeling with a focus on the role of the prior.

Consider an investigator who is exploring whether numbers are represented as analog magnitudes or as distinct symbolic categories (Dehaene, 2003; Gallistel & Gelman, 1992). A classic finding in this domain, first reported by Moyer & Landauer (1967) and replicated repeatedly is the *distance-from-five effect*. The participant is presented a single digit, say 4, and must judge whether this digit is less-than or greater-than five. Accuracy is near ceiling in such tasks for most adults, and the main variable of study is response time. Moyer & Landauer (1967) noted that in psychophysical experiments it takes longer to compare two closely-related analog magnitudes, such as two similar weights (e.g., Falmagne, 1990). If numbers are represented as analog magnitudes, then one might predict it would take longer to compare 4 to 5 than to compare 1 to 5. Hence, response time should speed as the presented digit is further from 5. Moyer and Landauer report such a distance-from-five effect where digits further from five are judged more quickly than digits closer to five. This effect is considered a landmark finding supporting the proposition that humans represent numerical magnitudes in an analog fashion.

We model the response times to digits as a function of their distance from five. Digits 4 and 6 are nearest to five (distance 1); digits 3 and 7 are next nearest (distance 2), and so on. For single digits between 1 and 9, there are four distance levels. Consider 40 participants performed judgments, each at a single distance level. For example 10 participants judged 4 and 6, another 10 judged 3 and 7, and so on. We will expand the design to the mixed version where all participants judge all distances subsequently, but for now, it is helpful to start with this simpler though less reasonable between-subjects design.

We start with the following top-level model: Let Y_{ij} be the mean response time for the j th participant in the cell with the i th distance condition. The model of mean response time

$$Y_{ij} = \mu + c_i + \epsilon_{ij},$$

where μ is a grand mean, c_i is the effect of the i th distance, ϵ_{ij} is a zero-centered noise term with variance σ^2 , and all the noise terms are independent. One immediate problem with this model is that there are more parameters contributing to the cell means than there are cell means, and, consequently, the parameters cannot be identified uniquely. A standard approach is to add a sums-to-zero constraint ($\sum_i c_i = 0$). With this constraint, sample means may serve as estimators:

$$\hat{\mu} = \bar{Y}, \tag{3}$$

$$\hat{c}_i = \bar{Y}_{i\cdot} - \bar{Y}, \tag{4}$$

where \bar{Y} is the grand sample mean and $\bar{Y}_{i\cdot}$ is the sample mean for the i th cell. We refer to the constraint as the *sums-to-zero* constraint and the above estimates as the *sample-mean estimators*.

Bayesian analysis offers a different approach that outperforms the sums-to-zero constraint. Consider the following prior specification for grand mean and effects:

$$\mu \sim \text{Normal}(0, \sigma_\mu^2), \tag{5}$$

$$c_i \sim \text{Normal}(0, \sigma_c^2). \tag{6}$$

The key constraint is that on c_i . For reasonable settings of σ_c^2 , say those on the order of milliseconds-squared in this example, the c_i parameters cannot be arbitrarily different. For example, it is highly unlikely that the effect of distance is on the order of milliseconds for

three of the levels and on seconds for the fourth. So long as σ_c^2 is finite, then the model is identifiable and estimable, and the smaller σ_c^2 , the more constraint on the variation of c_i .

One of the consequences of the shared prior on each c_i is that estimates of c_i are smaller in magnitude than the sample-mean estimates from (4). These estimators are known as *shrinkage estimators* because the estimate is “shrunk” toward a zero effect. Shrinkage estimators are well known in both Bayesian and frequentist statistical literatures (e.g., Efron & Morris, 1977; James & Stein, 1961), and in most situations are more accurate than sample means (Stein, 1959). Figure 3 shows an example. In this case, there are four levels, one for each distance, and the true means are shown on the top row. These are equally spaced, that is, the true effect of distance is additive on response time. Shown below are the data, and there are ten observations per level. The sample means are shown below, along with Bayesian estimates with prior settings of $\sigma_\mu^2 = \infty$ and $\sigma_c^2 = 1$. In this last row, we see shrinkage to the grand mean, that is, the Bayesian estimates are closer to each other than the sample means are to each other. This shrinking results in a lower root mean squared error (RMSE) in this case. Though a lower RMSE is not guaranteed for all sample data, it does hold on average (James & Stein, 1961).

A key question is how σ_c^2 should be set. If it is too large, then there is little constraint and shrinkage. If it is set too small, then the means will be too constrained to be similar and important variation across levels might be missed. In the next two sections, we address the issue.

Effect Size, The Natural Place To Add Constraint

One helpful approach in setting the group-level variance is to reparameterize the model in terms of effect size. In the previous mean-RT model, there is a separate parameter for each level of the factor. This parameter has a unit of measure, the unit of the dependent variable. For instance, if Y is response time, and it is measured in seconds,

then c_i is in seconds and σ_c^2 , the parameter of the prior, is in squared seconds. Priors must therefore be adjusted if one researcher uses seconds and another uses milliseconds.

Moreover, the priors must be specified for specific measures, and will change for other dependent measures, say for scalp voltage or for IQ points. It is more desirable to specify priors in a more standard fashion that is not so dependent on the unit of measure.

Fortunately, psychologists are well-versed in the solution to this problem—parameterization through effect sizes. Effect size is a unitless number, and its interpretation is the same regardless of the unit of the dependent variable. Consider the following parameterization of the top-level of the one-way ANOVA model:

$$Y_{ij} = \mu + \sigma d_i + \epsilon_{ij}. \quad (7)$$

Here, d_i is the effect size, $d_i = c_i/\sigma$. Note that this model is equivalent to the previous top-level model; the difference is in what the parameters mean, or the parameterization.

In the effect-size parameterization, noninformative priors may be placed on μ and σ^2 , the parameters with units, but proper informative priors must be placed on effect size, the parameter without units (Jeffreys, 1961). We take these prior specifications in turn. The noninformative prior for μ is a flat prior with equal weight on all values; the noninformative prior for variance is a flat prior on $\log \sigma^2$, that is, all values of the log of variance are equally weighted (Jeffreys, 1961; Box & Tiao, 1973). Following Jeffreys, this prior is expressed as

$$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Mirroring the approach we took above, the prior on the effect size parameter d_i is

$$d_i \sim \text{Normal}(0, g_d). \quad (8)$$

This prior results in shrinkage for d_i . The question remains on how to set g_d , the variance of effect sizes. Some researchers simply stipulate it, say $g_d = 1$. This choice is reasonable because we have an intuitive understanding, and the notion here is that while most effect sizes are *a priori* small, some will be large. The average spread, the standard deviation, is 1.0. This choice, while informative, is reasonable for many experimental contexts. In fact, the $g_d = 1$ prior setting is made in the derivation of the BIC goodness-of-fit statistic (Raftery, 1995). In our view, settings of $\sqrt{g_d}$ between say .2 and 1.0 seem defensible depending on the context. There is, however, a better approach.

A Hierarchical Model for Adaptive Shrinkage

A different approach is to treat g_d as a parameter to be estimated from the data. If there are large group effects in the data, then the d_i s will be well separated and g_d will tend to be large, and, importantly, there will be less shrinkage. If there are small group effects, then g_d will tend to be small and there will be more shrinkage. This adaptive nature in shrinkage is one of the appeals of hierarchical modeling in frequentist and Bayesian contexts (e.g., Efron & Morris, 1977). Not surprisingly, this hierarchical specification is popular among Bayesian analysts; examples with linear models include Bayarri & Garcia-Donato (2007), Gelman (2005), Liang, Paulo, Molina, Clyde, & Berger (2008), and Zellner & Siow (1980).

When g_d is treated as a parameter, a prior is needed for it. One convenient and flexible choice is a scaled inverse χ^2 distribution with a single degree of freedom (Zellner & Siow, 1980). The density of this prior is

$$f(g; h) = \frac{h^2}{2g^{3/2}\Gamma(1/2)} \exp(h^2/2g),$$

where Γ is the gamma function (Abramowitz & Stegun, 1965). The constant h^2 serves as a tuning parameter that sets the *a priori* expected range of effect sizes. The key specification

is the choice of h^2 , which must be set *a priori*. We can rewrite the model on d as:

$$\begin{aligned}d|g_d &\sim \text{Normal}(0, g_d), \\g_d &\sim \text{Inverse } \chi^2(1, h^2)\end{aligned}$$

Perhaps the easiest way of seeing the effect of this setting is to look at the influence of h^2 on d itself. To do so, we integrate out variation in g_d , that is, we examine the marginal distribution of d rather than the conditional one. This marginal distribution is a scaled Cauchy distribution, which is also a scaled t distribution with one degree of freedom. It too has a characteristic scale h . Figure 4A shows the prior distribution of effect sizes d for a few choices of h . If this scale is set too small or too large, say $h = 10$ or $h = .02$, respectively, then the marginal distribution has too little mass on reasonable effect-size values. Good choices are $.2 < h < 1$. We use $h = .5$ as a default setting both here and in our software. We recommend researchers tune the prior according to context. In the discussion section, we consider how inference varies as a function of h .

Previously, we discussed how Bayes factors were ratios of predicted probabilities under competing models. The one-way effect size model may be compared to a null model given by $Y_{ij} = \mu + \epsilon_{ij}$. Figure 5 shows the construction of Bayes factors for the one-way case with two levels for 30 observations. Figure 5A provides a representation of a null model on d as well as the effects model—the marginal model on d for $h = .5$. Figure 5B provides the predicted densities for different observed effect size. If the observed effect size is large in magnitude, say at $.5$, then the predicted density is higher for the effects model than the null model. Likewise, if the observed effect size is small in magnitude, say at $.2$, then the predicted density is higher for the null model than the effects model. Figure 5C shows the Bayes factor as a function of observed effect size. In the following development we expand the ANOVA models to account for multiple factors, mixed designs, and fixed

effects. In all cases, however, the mechanics of model comparison work the same. Predictions for competing models are obtained by marginalizing over the parameters, and the evidence for one model relative to another is the ratio of the densities of these predictions evaluated at the observed data.

Multiway ANOVA: A Hierarchical Clustering of Factors

In factorial designs, experimental factors are manipulated through several levels each. For example, consider a researcher who asks whether there is a numeric distance effect for word-form number terms (e.g., “seven”) as well as digit-form number terms (e.g., “7”). This second factor is called here presentation—either the numbers are presented as words or digits. If the factors are crossed, then there are four levels of distance and two levels of presentation (words vs. digits). The design spans a total of 8 cells. For concreteness, consider a design where 80 participants judge digits with 10 participants in each cell. The critical questions are about whether there is a main effect of distance, a main effect of presentation type, and an interaction between the two. We use the term *covariate* in this context to refer to individual additive effects, say the main effect of distance, the main effect of presentation, or the interaction.

We start with the following top-level model: Let Y_{ijk} be the response time for the k th participant in the cell with the i th distance and j th presentation, $i = 1, \dots, 4$, $j = 1, \dots, 2$, $k = 1, \dots, 10$. The top-level model of mean response times is

$$Y_{ijk} = \mu + \sigma[d_i + p_j + (dp)_{ij}] + \epsilon_{ijk},$$

where μ is a grand mean, d_i is the effect size of the i th distance, p_j is the effect size of the j th presentation, $(dp)_{ij}$ is the interaction term for the ij th combination of distance and presentation, and ϵ_{ijk} is a zero-centered noise term with variance σ^2 . As before, all error

terms are independent. Models are placed on effect sizes as follows:

$$\begin{aligned}d_i|g_d &\sim \text{Normal}(0, g_d), \\p_j|g_p &\sim \text{Normal}(0, g_p), \\(dp)_{ij}|g_{dp} &\sim \text{Normal}(0, g_{dp}).\end{aligned}$$

with

$$g_d \sim \text{Inv } \chi^2(1, h_d), \quad g_p \sim \text{Inv } \chi^2(1, h_p), \quad g_{dp} \sim \text{Inv } \chi^2(1, h_{dp}).$$

This prior specification captures the notion of clustering effects within factors. Because the model is hierarchical with a parent distribution across level effects within a covariate, these level effects are correlated and shrinkage occurs. Yet, because there are separate parent distributions for each covariate, the effects for one factor are independent of those for the other. It is for this reason that Gelman (2005) describes ANOVA as being a hierarchical clustering of effects within factors.

The flexibility in hierarchical clustering is well-suited for experimental psychology. In many experiments, researchers combine disparate factors where effects may vary considerably. The most common example is that people often vary greatly in their performance while manipulations have a much smaller effect. For the numeric distance effects, we can document three different effects all on different orders of magnitude. The effect of people is the greatest, with different individuals having means that vary by 100s of milliseconds. Distance effects typically span say 50 ms or so (Rouder, Lu, Speckman, Sun, & Jiang, 2005). Other effects, say the effect of whether the previous trial was a word or digit, are very small (< 10 ms) should they exist at all. In the hierarchical approach, level effects within a factor, say that of distance, are expected to have about the same scale or magnitude, and a distance effect of 200 ms would be unexpected and shrunk in considerably. Yet, simultaneously, participant effects of 200 ms would not be shrunk

nearly as much as this is an expected effect given the range of other item effects.

Shrinkage is adaptive, but with different adaptations for different factors. Hence, with the hierarchical structure, researchers may run designs with factors that vary considerably in potential effect sizes.

Fixed vs. Random Effects

The distinction between fixed and random effects has a long and somewhat controversial history in statistics (for an enlightened discussion, see Gelman, 2005). It may seem that effects in Bayesian models are random effects by default as each is a random variable that has a prior distribution. In the current case, the effects within a factor are treated as exchangeable, that is, they are treated explicitly as samples from a common parent distribution over effects. This treatment reinforces the random-effects interpretation, and with this interpretation, analysts can draw inferences not only about the effect of each level, but about the population of effects as well.

There is, however, a critical distinction between fixed and random effects in Bayesian analysis on the dimensionality of the model. In frequentist treatments, the sums-to-zero constraints reduce the number of parameters by one per constraint. In a one-way design, for example, there is a single sums-to-zero constraint that reduces the dimensionality by one parameter. The simplest way to impose this constraint in a Bayesian model is to consider only the first $I - 1$ effects and let the last one be the negative of the sum of the rest. For example, if there are four distance effects, the effect sizes for the first three levels may be parameters as d_1 , d_2 , and d_3 , and the remaining effect size is simply $d_4 = -(d_1 + d_2 + d_3)$ such that the sum of all four effect sizes is zero. Priors are placed on parameters d_1 , d_2 , and d_3 , but not on d_4 as this parameter is a function of the other three. In this case, the dimensionality of the problem is reduced by one parameter. This reduction is a form of constraint much like the hierarchical priors are

a form of constraint. In this case, there is constraint provided by both the dimension reduction and the hierarchical priors; for instance, the estimates of the $I - 1$ parameters would be more diffuse if a flat prior rather than a hierarchical prior were used.

In practice, we do not eliminate one effect size. Instead we project the set of I main effects into $I - 1$ parameters such that the sums-to-zero constraint holds and the prior model on each effect is the same. These projections are more of a technical nicety, closely related to orthogonal contrasts; the details are provided in Rouder et al. (2012). Readers unfamiliar with projections need only note that they act like sums-to-zero constraints in that they reduce model dimensionality.

The notion of modeling fixed effects as a reduction in model dimensionality extends gracefully to interactions. In the above example with four levels of distance and two levels of presentation, the unconstrained interaction covariate was comprised of eight levels. In the usual fixed-effect treatment, separate sums-to-zero constraints are applied across rows and columns of the interaction matrix resulting in five sums-to-zero constraints. These constraints reduced the number of interaction terms to three from eight. Sums-to-zero constraints can be applied to both factors if both are treated as fixed, or to one factor but not the other if one but not the other is treated as fixed, or to no factors. Perhaps the most interesting case is the one where one factor is fixed and the other is random. Here, the treatment is equivalent to imposing sums-to-zero constraints on either rows or columns of the interaction matrix, an approach advocated by Cornfield & Tukey (1956). This leaves four different interaction models. If both factors are random, then there are no sums-to-zero constraints resulting in eight interaction parameters. If distance is random and presentation is fixed, then there are four separate sums-to-zero constraints, one for each distance level, and a total of four interaction parameters. If distance is fixed and presentation is random, there are two separate sums-to-zero constraints, one for each presentation level, and a total of six interaction parameters. Finally, when both are fixed,

there are separate sums-to-zero constraints for each row and column, but one of these is redundant. Hence, there are five constraints and three interaction parameters.

Dimension reduction through sums-to-zero constraints is a different constraint than the hierarchical clustering. The two constraints are not mutually exclusive and may be applied simultaneously. One may impose the sums-to-zero constraint and then cluster the resulting parameters hierarchically. For example, let \mathbf{d} , \mathbf{p} , and (\mathbf{dp}) be vectors of I , J , and IJ main effects and interactions, respectively, and let \mathbf{d}^* , \mathbf{p}^* , and $(\mathbf{d}^*\mathbf{p}^*)$, be vectors of $I - 1$, $J - 1$ and $(I - 1)(J - 1)$ main effects and interactions after dimension reductions. Then the following priors may be put on these effects:

$$\begin{aligned} d_i^* &\sim \text{Normal}(0, g_d), \\ p_j^* &\sim \text{Normal}(0, g_p), \\ (d^*p^*)_{ij} &\sim \text{Normal}(0, g_{dp}). \end{aligned}$$

As before, the levels within a factor are related by the parent distribution, while there is no such relation across levels of different factors. That is, the effects are clustered within factors.

Should researchers use the reduced-dimension fixed-effects models or the more heavily dimensioned random-effects models? The question to us is one of model specification. Recall that models tend to be specified to help understand the structure in data, and in particular, whether theoretically-motivated constraints tend to hold. In this context, the sums-to-zero constraint provides a focus on differences between levels rather than the levels themselves. In line with the conventional wisdom, this focus seems best when the researcher wishes to learn about the particular levels in an experiment rather than new levels not present in the experiment. Conversely, the more complicated random-effects models may be used for generalizing to new levels.

Between-Subjects, Within-Subjects, and Mixed Manipulations,
Designs, and Models

Experimental psychologists and statisticians have developed somewhat different nomenclature for describing how experimental factors are crossed with participants. The following is fairly conventional for experimental psychologists: If a factor is manipulated in a between-subjects manner, then each participant observes one level of the factor. Conversely, if a factor is manipulated in a within-subject manner, then each participant observes all levels of the factor. In a between-subjects design, all factors are manipulated in a between-subjects manner; in a within-subjects design, all factors are manipulated in a within-subject manner, and in a mixed design, some factors are manipulated within subjects while others are manipulated between subjects. Experimental psychologists will refer to ANOVA as “between-subject,” “within-subject” and “mixed” if it applies to between-subject, within-subject, and mixed designs, respectively. Moreover, experimental psychologists do not count subjects as a factor when discussing designs. For instance the design where distance and presentation are manipulated in a within-subject manner would be referred to as a two-way design even though in the corresponding ANOVA, the variation across participants and the variation within a participant are jointly modeled with separate variance components. Statisticians in contrast would consider the design a three-way design (with factors of people, distance, and presentation), and the corresponding ANOVA model to be mixed. We adopt here the experimental psychologists’ nomenclature for within-subject and between-subject and mixed designs, but use more standard statistical nomenclature to characterize the number of factors and type of model.

The hierarchical ANOVA model is perfectly set up to account for between-subject, within-subjects, and mixed designs. If all factors are manipulated between subject, then there is no need for a subject factor as the effect of subjects serves as an unidentifiable part of the error variance. If any factor is within-subject, however, then it is possible to

separate participant variability from residual error. In these cases, we add a factor for people, and each person serves as a level. For example, if distance and presentation are manipulated as within-participant factors, then the following top-level model may be used:

$$Y_{ijk} = \mu + \sigma[s_k + d_i^* + p_j^* + (d^*p^*)_{ij}] + \epsilon_{ijk},$$

where s_k is the random effect of each subject. Note that the distance and presentation are treated as fixed effects and have a reduced dimensionality. Also note the lack of interactions between subjects and the other factors. This choice is discussed at length in the next section. The specification of the model is completed with the following priors:

$$\begin{aligned} d_i^* &\sim \text{Normal}(0, g_d), \\ p_j^* &\sim \text{Normal}(0, g_p), \\ (d^*p^*)_{ij} &\sim \text{Normal}(0, g_{dp}), \\ s_k &\sim \text{Normal}(0, g_s), \end{aligned}$$

which encodes independent pooling across participants.

Model Comparison Strategies

Conventional ANOVA is comprised of a series of top-down model selection steps. Consider the two-way example above where the distance and presentation are varied. The full model, that is the model with the maximal number of covariates is

$$\mathcal{M}_f \quad Y_{ijk} = \mu + \sigma[d_i + p_j + (dp)_{ij}] + \epsilon_{ijk}. \quad (9)$$

In conventional ANOVA as practiced by psychologists, the effect of one covariate, say the main effect of distance, is tested while the other covariates are estimated freely. This

approach may be considered a top-down model selection approach. The main effect of distance is a comparison of the full model to a model with distance effects omitted, which is given by

$$\mathcal{M}_1 \quad Y_{ijk} = \mu + \sigma[p_j + (dp)_{ij}] + \epsilon_{ijk}. \quad (10)$$

If the top-level model compares favorably, then the main effect of distance is evidenced; if the model without the distance main effect compares favorably, then the distance main effect is unnecessary. Likewise, the main effect of presentation is obtained by comparing the full model to

$$\mathcal{M}_2 \quad Y_{ijk} = \mu + \sigma[d_i + (dp)_{ij}] + \epsilon_{ijk}, \quad (11)$$

and the interaction is obtained by comparing the full model to

$$\mathcal{M}_3 \quad Y_{ijk} = \mu + \sigma(d_i + p_j) + \epsilon_{ijk}. \quad (12)$$

Of course there are more models that may be considered, and the remainder is given by

$$\mathcal{M}_4 \quad Y_{ijk} = \mu + \sigma d_i + \epsilon_{ijk}, \quad (13)$$

$$\mathcal{M}_5 \quad Y_{ijk} = \mu + \sigma p_j + \epsilon_{ijk}, \quad (14)$$

$$\mathcal{M}_6 \quad Y_{ijk} = \mu + \sigma[(dp)_{ij}] + \epsilon_{ijk}, \quad (15)$$

and a null model with no effects,

$$\mathcal{M}_0 \quad Y_{ijk} = \mu + \epsilon_{ijk}. \quad (16)$$

Each of these models may be compared by Bayes factor. If a model without an effect is most supported, say $Y_{ijk} = \mu + \sigma d_i + \epsilon_{ijk}$, then there is evidence for a distance effect and evidence against a presentation effect.

One problem with considering all the models in a factorial design is the number of such models grows alarmingly fast with increasing numbers of factors. A two-way design provides for three possible covariates and for 8 possible models. A three-factor model provides three main effects, three two-way interactions, and a three-way interaction. The presence or absence of each of these 7 covariates yields 128 possible models. Four- and five-factor designs yield 32,768 and about 2.1 billion models, respectively. Note that with within-subject and mixed designs, the effect of subjects must be modeled as a factor, so that three, four, and even five factor designs are not uncommon. For example, consider a design in cognitive aging where a researcher wishes to study two manipulated factors across younger and older adults. There would be four factors in the design with factors for age and subjects in addition to the two manipulated factors. Consideration of thousands, millions, or billions of models is computationally infeasible as well as intellectually unappealing.

Fortunately, there are reasonable approaches for cutting down the number of models under consideration. The first is to not test effects in nuisance variables. Consider the variation across participants and items. This variation certainly needs to be modeled, but it need not be tested. Hence, all models may include variation from these factors, and models without this variation need not be considered. Such a reduction is common in practice as researchers often do not report tests of these nuisance factors in experimental settings.

A second complementary approach, discussed by Nelder (1998) and Venables (2000), is to note that models with interactions but without corresponding main effects are often not plausible. Model \mathcal{M}_6 is an example. In this model there is an interaction between

distance and presentation without a main effect of either. To understand the plausibility of such a model, consider a hypothetical context where a cross-over is believable. Suppose we perform an experiment to explore how much people would be willing to pay for a pint of vanilla ice-cream as a function of its sugar and fat content. Figure 6A shows a hypothetical example of a balancing phenomenon. Here, the sugar and fat need to be matched to maintain balance, otherwise the ice-cream tastes too sweet or too bland. As such, perhaps the lack of main effects may seem plausible. To show that it is not, consider Figure 6B which shows a more complete view of the relationship between sugar content, fat content, and value of the pint of ice-cream. The inner most circle shows the highest value, in this case above 6\$ a pint, and the successive rings show lessening value. The balance phenomena comes from the positive association: for each level of fat there is a balancing level of sugar. The points, at values of 2.5\$ and 5\$, correspond to the points in Figure 6A. For the shown levels there are no main effects. But, if the levels are chosen differently, then main effects surely appear. That is, the lack of main effects here results only for very carefully and precisely chosen levels, and main effects appear otherwise. In application, we do not have the information in Figure 6B — otherwise, we would not need to perform the experiment — and it seems implausible that one could fortuitously chose levels to precisely cancel out the main effects. For the two-factor models above, we exclude Model \mathcal{M}_1 , \mathcal{M}_2 , and \mathcal{M}_6 from consideration because they specify interactions without a corresponding main effect for the interacting factors. In our software, discussed next, the default is to exclude these models though this default may be overrode by researchers who disagree with the exclusion.

A third approach for reducing the set of models is to exclude from consideration models with subject-by-treatment interactions, or what is sometimes known as random slopes. Consider a researcher who is assessing a distance effect and decides to include a subject-by-distance interaction. For concreteness, we assume there are only two distances

in the design: say digits 4 and 6 (one away from five) and digits 3 and 7 (two away from five). The subject-by-distance interaction implies that there may be some participants who have a true distance effect in the usual direction effect but others who have a true effect in the reverse distance direction (faster responses for the smaller distance). Such interactions make it difficult to interpret the main effects of distance. For instance, suppose 60% of the population has a 20 ms reverse distance effect while 40% has a 60 ms ordinary distance effect. There is an average 12 ms ordinary distance effect, yet this 12 ms effect is not interpretable given that it mischaracterizes a sizable proportion of the people. There are two approaches to dealing with this problem. The first is not to include these interactions when assessing main effects of variables of theoretical interest. Eliminating models with subject-by-treatment interactions is analogous to conventional analyses where such interactions are part of the statistical error in the model. The second is to place constraints on subject-by-treatment interactions much as is done in item-response-theory models. To date, we do not know of the development of Bayes factors for constrained interactions in factorial designs, and as such and for the time being, we recommend the first approach of not considering subject-by-treatment interactions. As with any analysis, exploratory data analyses and theoretical considerations can help analysts decide which is the most appropriate approach for the problem at hand.

The three elimination strategies — not testing nuisance effects, eliminating models with interactions and no corresponding main effects, and eliminating subject-by-treatment interactions — reduces the number of models considerably. For example, in the four-factor case above with factors of distance, presentation, age, and subjects, the strategy reduces the number of models from 32,769 to 19.

Bayes Factor Computation

When researchers run factorial designs, they often wish to perform inference on main effects and interactions. Although Bayes factors provide an ideal means of doing, they may be difficult to compute in practice. The difficulty comes in the computation of the predictions, that is, of the marginal probability of data conditional on a model, $Pr(\mathbf{Y} | \mathcal{M})$. This marginal probability may be expressed as

$$Pr(\mathbf{Y} | \mathcal{M}) = \int_{\boldsymbol{\theta} \in \Theta} Pr(\mathbf{Y} | \boldsymbol{\theta}, \mathcal{M}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where $\pi(\boldsymbol{\theta})$ is the prior. The term $Pr(\mathbf{Y} | \boldsymbol{\theta}, \mathcal{M})$ is the probability of the data at a given parameter vector, the likelihood, and is often convenient to evaluate. Even so, the integration, however, is often computationally difficult. If model dimensionality is small, then the integral may be evaluated with standard numerical methods. The difficulty occurs for larger dimension models, which, unfortunately, are common in even relatively simple applications in psychological science. For instance, consider the previous example with 4 distance levels and 2 presentation levels where the levels of each factor are crossed with participants in a mixed design. Let's say there are 20 participants. Even if we consider all effects as fixed and exclude all participant-by-treatment interaction terms, there are 31 parameters to integrate across. Numerical methods become increasingly less reliable and more time intensive in these cases, so much so that it becomes inconvenient in usual applications.

Fortunately, integration across the parameters for the presented ANOVA models is tractable. The key insight provided by Zellner & Siow (1980) is that the integration of all parameters save the g parameters may be expressed in an analytic form. The remaining g parameters may be integrated out via a Monte Carlo simulation, and the computation of the Bayes factors is fairly quick and precise at least for models with typically small

number of factors. Rouder et al. (2012) provide the derivation of the computations, and these are implemented in Morey & Rouder's `BayesFactor` package for R. In the following example, we show how the package may be used to compute Bayes factors in analysis.

Consider a hypothetical example where younger and older adults perform the less-than-five task for the four distances and across the two presentation types. Distance and presentation are manipulated within participants; age is manipulated between participants. Each participant observes 25 trials in each of the 8 cells, and the eight cell mean response times serve as the raw data for analysis. We describe analysis with R with the `BayesFactor` package in the next section. Following, we discuss a new and evolving statistics package, JASP, which is an open-source, user-friendly Bayesian-statistics alternative to SPSS. JASP provides a graphical interface to the `BayesFactor` package, and it may be used to compute Bayes factors without any knowledge of R.

Analysis in R with the BayesFactor package

Appendix A provides details on loading the hypothetical data set into R and installing the `BayesFactor` package, which is used in the following analysis. Figure 7 shows the cell means for the hypothetical example. Visually, the trends suggest a main effect of age, distance, and presentation, and, perhaps, an age-by-distance interaction. Table 1 provides the conventional ANOVA analysis. There is a significant main effect of distance ($F(3, 144) = 30.9, p < .05$) and a significant age-by-distance interaction ($F(3, 144) = 6.4, p < .05$). No other effects are significant, including the main effect of age.

The Bayes factor ANOVA can be performed using the `anovaBF` function:

```
bf = anovaBF(rt~a*d*p+s, data = dat, whichModels="withmain",
             whichRandom="s", iterations = 100000)
```

where the variable `bf` stores the result. The result of a Bayes factor analysis is not

simply a table; the object `bf` contains all the information necessary to compare any two models, to estimate parameters from any specific model, and to reproduce the analysis.

It is helpful to sort the models in `bf` so that the models will be presented in order of the magnitude of their Bayes factor:

```
bf=sort(bf, decreasing = TRUE)
bf
```

The result of the Bayes factor ANOVA is shown in Table 2. There are 18 lines, one for each of the 18 non-null models. The number of models reflects not only the model formula — which in this case includes four factors — but the options `whichModels` and `whichRandom`. Option `whichRandom` indicates which factors are random, and these factors are not tested and enter in only as main effects for the reasons discussed previously. The option `whichModels` indicates how the hierarchy of models is to be constructed. For ANOVA, there are three options of interest: the default `withmain`, `all`, and `top`. Option `all` constructs a hierarchy with all possible models except for interactions with random factors. In this case, it would include three additional models where an interaction between two factors is present but at least one of the main effects is missing. Option `withmain` is similar to `all`, but it excludes all those models where interactions are present without corresponding main effects. Option `top` includes the most general model and one where a single covariate is removed. This option is most similar to classical ANOVA where contrasts are made top-down with the inclusion of all other covariates. We recommend that the default option, `withmain`, be used so that difficult-to-interpret models are excluded; however, analysts may wish to use one of the other settings in specific circumstances. Finally, the argument `iterations` controls the number of iterations performed by the estimation algorithm.

The first column in the table provides the models for comparison, and second column provides the Bayes factor relative to a common comparison model, defined on the

last line indicated by the “Against Denominator” label. In this case, the denominator is $rt \sim \mu + \text{subjects} + \epsilon$. More generally, the null model includes the grand mean, the noise terms, and any random effects indicated by the “whichRandom” option to the function. These random effects are common in all models, and the focus is squarely on assessing the evidence for the fixed effects. Consider, for example, the 15th model, which is $rt \sim \mu + \text{age} + \text{subjects} + \epsilon$. The Bayes factor is the comparison of this model against the null model with no age effect, and this age-effect model is less preferable to the null model by 0.69-to-1. The last column shows the estimated error from the numerical integration that is used to estimate the Bayes factor. For some of these models, the errors seem large; however, in comparison to Bayes factor values they are not particularly worrisome. Greater precision may be had by initially increasing the number of iterations (with the `iterations` argument to the function) or sampling additional iterations with the `recompute` function. Details are available in the `BayesFactor` manual.

To draw inferences from the output, it is helpful to focus on the most preferred model, the one with main effects of distance and age and an interaction between the two. One way we can assess the evidence for each covariate by comparing this most preferred model with one missing the covariate in question. For example, the evidence for the age-by-distance interaction is comprised of the Bayes factor between the 1st (winning) model, and the 4th model, the best-fitting model without the interaction. The Bayes factor values relative to the null are 1.7×10^{16} and 3×10^{14} , respectively. The ratio of these values is about 56-to-1, which is the Bayes factor of the winning model relative to the one without the interaction. This value serves as the evidence in favor of the age-by-distance interaction; the interpretation is that there is substantial evidence that older adults have a larger distance effect than younger adults. The same process yields evidence for the inclusion of age (a comparison of the winning model to the 4th model yields a Bayes factor of 56-to-1), and for the inclusion of distance (a comparison of the

winning model to the 15th model yields a Bayes factor of 2×10^{16} -to-1).

Bayes factors may also be used to state evidence against effects as well: that is, evidence for invariances. The winning model did not include the factor of presentation type, indicating evidence that the results were invariant to whether the stimuli were digits or words. The strength of the evidence for this invariance is given by the comparison of the winning model to the best model with presentation included, the 2nd model. The Bayes factor is 5-to-1, which is positive evidence for the invariance or lack of an effect.

The Bayes Factor package includes a number of useful utilities. For example, the command `head(bf)` will produce output for the six most preferred models. Bayes factors may be plotted as well, and Figure 8 may be obtained in the Bayes factor package with the command

```
plot(head(bf))
```

The above comparison approach is based on selecting the best model and comparing it to competitors to state the evidence for or against main effects and interactions. An alternative approach, based on model averaging, is discussed in Appendix B.

Analysis in JASP

JASP is designed to be a user-friendly, open-source replacement for SPSS that performs Bayesian analysis (Love et al., n.d.). The main advantage of JASP for Bayesian inference is that it is easy-to-use and requires no special knowledge. It makes use of simple menus and boxes, and may be mastered quickly by researchers with familiarity with SPSS or Excel. Currently, JASP has functionality for performing t-tests, ANOVA, regression, and cross-tabulations. JASP uses the `BayesFactor` package as a back-end computational engine. Consequently, it serves as a graphical interface for Bayes-factor computations. JASP is freely available at <http://jasp-stats.org>.

Because JASP is a menu-driven, graphical interface, the explanation of how to use it is best provided as a screencast (<https://www.youtube.com/watch?v=wVJemH3G11I>). The screencast uses the above lexical-distance-effect example. Data can be loaded into JASP in a comma-separated format (.csv); see the Appendix for details on the location of this file. Figure 9 shows a screen shot from the analysis. The top row shows the menus, and ANOVA has been selected. Much like SPSS, variables are identified by dragging from one box to another. The right side shows the output. The top-right box, which is highlighted, shows Bayes factors for all the models. It provides the same values as provided in Table 2. The bottom-right box shows the model-averaged probabilities on factors and interactions and is discussed further in Appendix B.

The Role of The Priors Revisited

The main goal in the above analyses is to compute Bayes factors for different ANOVA models that encode effects and invariances of factors. In Bayes factor computation, the prior plays a critical role, and in fact, we use different priors on effect sizes to specify different models. The prior settings of note in all models are those on the g parameters. These g parameters determine the *a priori* dispersion of effects around zero. At first glance it may seem wise to specify priors that favor large values of g , or a large *a priori* variance on effect sizes. Such a view may be seen as “noninformative” as a researcher does not commit to any particular range of effect sizes. Yet, this prior setting is unwise. As DeGroot (1982) notes, wide priors do contain a commitment or information, but the commitment is to very large effects. If prior mass is spread across unreasonably large effect sizes, there must necessarily be little prior mass on reasonable ranges, say for effect sizes less than 3.0 in magnitude. If moderate effect sizes are observed, say those around .5, then they are more compatible with a null-effect model than with one that *a priori* overweights large effect sizes. Therefore, it is critical that the prior on g be

judiciously chosen rather than arbitrarily wide. Figure 4 showed how h , the scale on the variability of effects, tunes the marginal prior on effect size. The useful range is between $h = .2$ and $h = 1$.

Choosing an appropriate prior is a thoughtful process about which there are different prevailing views (Berger, 2006; Goldstein, 2006). We tend to think that the prior should reflect a reasonable set of beliefs informed by context and the literature (Dienes, 2014). In the above example, we used a default scale, which is a scale of $h = .5$ in version 0.98 of the `BayesFactor` package. This prior places mass in reasonable ranges without being overcommitted to any one point. Further justification for this default choice comes from consideration of previous research in distance effects, where effects span about 50 ms from a distance of 1 to a distance of 3 (Rouder et al., 2005). This 50 ms effect is recast as a ± 25 ms from grand mean. Participants have about 200 ms of variability, and given that there are 25 trials per cell, cell means have about 40 ms in variation. Hence, we expect effects on the order of $25/40$ or about .62. So, overall, the default value of .5 is quite reasonable for these types of data. We note that the value of the resulting Bayes factors are quite stable across a range of reasonable settings, and variation in Bayes factors across reasonable differences in prior settings should be accepted as part of the value-added nature of modeling in analysis.

The fact that Bayes factors are dependent on prior settings has fueled the critique that they are too subjective (e.g., Liu & Aitkin, 2008; Gelman & Shalizi, 2013; Kruschke, 2013). We think this critique is misguided for two reasons. First, the prior setting is seen as part of the alternative model, and the need to specify the alternative carefully and judiciously should be seen as part of good modeling. Second, specification at this level is necessary for models to yield predictions on data. The fact that well-specified data actually predict constraint on data is appealing. Evidence for the null should be calibrated against a belief in how big effect sizes would be if the null was false, and the dependence

of the Bayes factor on prior settings captures this intuition. The parameterization of ANOVA in effect sizes makes choosing a judicious alternative relatively straightforward. Figure 4B shows how setting h affects the Bayes factor value for a t -test. As can be seen, the Bayes factors do change with the setting of h , but not drastically so. In fact, the Bayes factor is far more sensitive to the t -value than the prior setting. This dependence of inference on reasonable variations in tuning is desirable. Researchers should be able to use their expectations as part of model specification to add value by calibrating inference.

All commands that produce Bayes factors in the `BayesFactor` package have an option for setting this scale h on effect sizes.

Conclusions

ANOVA remains the workhorse of experimental psychology. The Bayes factor approach highlighted in this report provides researchers with a principled approach to state relative evidence from data for or against effects and invariances. Models that encode effects and invariances are proposed, and the evidence is the probability of the data for a given model relative to that for an alternative model. This odds ratio is directly interpretable much as the odds ratio of a horse race is directly interpretable. Because the odds are directly interpretable, they may be reported without any further qualification and without reference to decision-making criteria. That is, if a researcher finds a Bayes factor of 2-to-1, such a value may be reported as is, and then the researcher can engage the reader in an argument over interpretation. For example, it may be judicious to consider a value of 2-to-1 unpersuasive if the effect is an extrasensory perception effect because such an effect if true would be in contravention of theories of physics and biology. Yet, it may also be judicious to be quite concerned about a value of 2-to-1 in a different context, say if the goal is to assess the effectiveness of controlling a confound through matching. Bayes factor evidence should not be evaluated by strict criterion, but in

reference to the context and goals of the researcher.

The ANOVA approach provided here is similar to that of Gelman (2005) who uses separate variance components for each covariate. The approach differs from Gelman in that the effect of the covariates is assessed through Bayes factors. To make the Bayes factor computationally convenient, we use the g -prior approach of Zellner & Siow (1980), and for ANOVA, use a separate g parameter for each factor.

Psychologists have not stressed the importance of stating evidence for invariances. Some such as Cohen (1994) and Meehl (1978) argue that true invariances never exist, others such as Cumming (2014) argue that cataloging effect sizes and confidence intervals is all that is needed. We briefly address these points: First, there is a sense in which advocacy of Bayes factors is not threatened with an *a priori* belief that invariances do not hold. Those who wish to replace the invariances with models in which effect sizes vary in a constrained equivalence region or with order restrictions are free to do so, and Morey & Rouder (2011) and Morey & Wagenmakers (2014) develop Bayes factors for these respective cases. Second, the claim that invariances never hold should be testable in principle. In fact, Bayes factors provide an ideal test of the proposition. One can compute the Bayes factor evidence for the invariance against other alternatives, say one in which the effect size is presumed small but nonzero. In this manner researchers who believe that invariances do not hold can attempt to build data-driven evidence for the position rather than assert it as a matter of faith. Third, even if invariances never hold they are still useful. Consider by analogy the invariances in physical laws, say the law that planets circle the sun on an ellipse. It is true that at some level this law is violated due to the tug of other planets and the presence of friction from interstellar materials. Even so, the platonic form of the law is exceedingly useful in uniting mechanics on earth with those in space. Stating invariances, even when they hold only approximately, will assuredly lead to better theory development. The alternative approach that has garnered recent attention,

the presentation of effect sizes and associated CIs, is desirable because there is never any abstraction of these effect sizes in theoretical terms (Morey, Rouder, Verhagen, & Wagenmakers, in press). We believe that it is this abstraction into effects and invariances that will drive better theoretical development, and this belief serves as additional motivation for adopting Bayesian model comparison.

References

- Abramowitz, M., & Stegun, I. A. (1965). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. New York: Dover.
- Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(1), 111-142. Retrieved from <http://www.jstor.org/stable/2345730>
- Bayarri, M. J., & Garcia-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, *94*, 135-152.
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407-425. Retrieved from <http://dx.doi.org/10.1037/a0021524>
- Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*, 385-402.
- Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*, 159-165.
- Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*(397), 112-122. Retrieved from <http://www.jstor.org/stable/2289131>
- Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. New York: Addison-Wesley.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.
- Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics*, *4*, 907-949.

- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.
- DeGroot, M. H. (1982). Lindley's paradox: Comment. *Journal of the American Statistical Association*, *77*(378), 336-339. Retrieved from <http://www.jstor.org/stable/2287246>
- Dehaene, S. (2003). The neural basis of the Weber-Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, *7*, 145-147.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Quantitative Psychology and Assessment*. Retrieved from 10.3389/fpsyg.2014.00781
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193-242.
- Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, *100*(469), 1–5.
- Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.
- Falmagne, J. C. (1990). *Elements of psychophysical theory*. New York: Oxford.
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439-453. Retrieved from <http://psycnet.apa.org/doi/10.1037/a0015251>
- Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*, 43-74.
- Gelman, A. (2005). Analysis of variance – why it is more important than ever. *Annals of Statistics*, *33*, 1–53.
- Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 57-64.

- Gelman, A., Shor, B., Bafumi, J., & Park, D. (2007). Rich state, poor state, red state, blue state: What's the matter with Connecticut? *Quarterly Journal of Political Science*, *2*, 345-367.
- Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*, 403-420.
- James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth berkeley symposium on mathematical statistics and probability* (p. 361-379).
- Jaynes, E. (2003). *Probability theory: the logic of science*. Cambridge, UK: Cambridge University Press.
- Jeffreys, H. (1961). *Theory of probability (3rd edition)*. New York: Oxford University Press.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299-312.
- Kruschke, J. K. (2012). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*.
- Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*, 573-603.
- 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. Retrieved from <http://pubs.amstat.org/doi/pdf/10.1198/016214507000001337>

- Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *56*, 362-375. Retrieved from <http://dx.doi.org/10.1016/j.jmp.2008.03.002>
- Love, J., Selker, R., Verhagen, J., Smira, M., Wild, A., Marsman, M., . . . Wagenmakers, E.-J. (n.d.). *JASP (version 0.40)*. (<http://www.jasp-stats.org>)
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834. Retrieved from <http://www.psych.umn.edu/faculty/meehlp/113TheoreticalRisks.pdf>
- Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2013). The humble Bayesian: model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, *66*, 68-75. Retrieved from <http://dx.doi.org/10.1111/j.2044-8317.2012.02067.x>
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406-419. Retrieved from <http://dx.doi.org/10.1037/a0024377>
- Morey, R. D., & Rouder, J. N. (2014). *BayesFactor 0.9.6*. Comprehensive R Archive Network. Retrieved from <http://cran.r-project.org/web/packages/BayesFactor/index.html>
- Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (in press). Why hypothesis tests are essential for psychological science: A comment on Cumming. *Psychological Science*.
- Morey, R. D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics & Probability Letters*, *92*(0),

121 - 124. Retrieved from

<http://www.sciencedirect.com/science/article/pii/S0167715214001862>

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519-1520.

Nelder, J. A. (1998). The selection of terms in response-surface models-How strong is the weak-heredity principle? *American Statistician*, *52*, 315-318.

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

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin and Review*, *12*, 195-223.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682-689. Retrieved from <http://dx.doi.org/10.3758/s13423-011-0088-7>

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877-903.

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.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225-237. Retrieved from <http://dx.doi.org/10.3758/PBR.16.2.225>

- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *64*, 583-639.
- Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *The Annals of Mathematical Statistics*, *30*(4), 877-880. Retrieved from <http://www.jstor.org/stable/2237432>
- Venables, W. N. (2000). *Exegeses on linear models*. Paper presented to the S-PLUS User's Conference. Retrieved from <http://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf>
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, *14*, 779-804.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. A comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426-432. Retrieved from <http://dx.doi.org/10.1037/a0022790>
- Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: Proceedings of the First International Meeting held in Valencia (Spain)* (pp. 585-603). University of Valencia.

Appendix A

The hypothetical data may be found at

<http://pcl.missouri.edu/exp/aovExample.txt> and can be read into R using the `read.table` function:

```
dat=read.table(url('http://pcl.missouri.edu/exp/aovExample.txt'), head=TRUE)
```

After reading in the data and prior to the analysis, it is recommended to check that all columns representing categorical factors are understood by R to be factors:

```
dat$s=factor(dat$s)
dat$a=factor(dat$a)
dat$d=factor(dat$d)
dat$p=factor(dat$p)
```

The classical ANOVA in Table 1 can be obtained with the following R command:

```
aovResult=aov(rt~a*d*p+Error(s/(d*p)), data=dat)
summary(aovResult)
```

In order to install the Bayes factor package, the `install.packages` function is used:

```
install.packages('BayesFactor', dependencies = TRUE)
```

The `install.packages` function only needs to be run once; so long as the package is installed, one can load it with the `library` function:

```
library(BayesFactor)
```

This command that loads the library must be run at the beginning of every session in which `BayesFactor` will be used.

The data file to use for JASP is found at

<http://pcl.missouri.edu/exp/Lessthan5.csv>.

Appendix B

An alternative approach to making statements about factors is to use Bayesian model averaging. In this approach models may be given *a priori* probabilities, perhaps the 19 models under consideration would be given *a priori* probabilities of $1/19$. Then these prior probabilities are converted into odds, updated via multiplication by the Bayes factor, and reconverted to become posterior probabilities. Models may then be partitioned according to whether they do or do not contain a factor. For example, consider the invariance to presentation. The following models do not contain any effect of presentation: the null model, Model 1, Model 4, Model 5, and Model 15. The posterior probabilities derived from the Bayes factors for these models are about 0, 0.752, 0.014, 0.013, and 0. The total posterior probability across all of these models that do not contain the presentation effect is 0.779, and the complement of this value is the probability across all models that do contain a presentation effect. We may compute posterior odds for this total probability, which is 3.518. To compute a composite Bayes factor for the invariance, we may total up the prior probabilities on the models that do not contain the effect, $5/19$, and convert it to prior odds (0.357). The Bayes factor for the invariance is just the posterior odds (3.518) divided by the prior odds (0.357), and it evaluates to 10-to-1. JASP provides prior and posterior model-averaged probabilities and Bayes factors in the bottom-right output panel under the columns $P(\text{inc})$, $P(\text{inc}|Y)$, and $\text{BF}_{\text{Inclusion}}$, respectively. Keep in mind, however, that this estimate is sensitive to Monte Carlo error.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
a	1	0.7911	0.7911	2.7644	0.1029
Residuals	48	13.7369	0.2862		
d	3	0.1559	0.0520	30.8855	0.0000
a:d	3	0.0323	0.0108	6.3971	0.0004
Residuals	144	0.2422	0.0017		
p	1	0.0023	0.0023	1.3459	0.2517
a:p	1	0.0007	0.0007	0.4259	0.5171
Residuals	48	0.0806	0.0017		
d:p	3	0.0017	0.0006	0.3581	0.7833
a:d:p	3	0.0028	0.0009	0.5921	0.6212
Residuals	144	0.2299	0.0016		

Table 1

The classical ANOVA results for the hypothetical example.

```

## Bayes factor analysis
## -----
## [1] a + d + a:d + s           : 1.687919e+16 9.62%
## [2] a + d + a:d + p + s       : 3.508808e+15 8.37%
## [3] a + d + a:d + p + a:p + s : 1.166088e+15 34.25%
## [4] d + s                     : 3.035337e+14 0.26%
## [5] a + d + s                 : 2.971541e+14 30.92%
## [6] a + d + a:d + p + d:p + s : 1.170903e+14 3.7%
## [7] a + d + p + s             : 7.550958e+13 1.98%
## [8] d + p + s                 : 6.304038e+13 0.64%
## [9] a + d + a:d + p + a:p + d:p + s : 2.358671e+13 5.16%
## [10] a + d + p + a:p + s        : 7.469271e+12 10.09%
## [11] d + p + d:p + s          : 2.361781e+12 1.36%
## [12] a + d + a:d + p + a:p + d:p + a:d:p + s : 2.214369e+12 5.57%
## [13] a + d + p + d:p + s       : 1.562008e+12 10.65%
## [14] a + d + p + a:p + d:p + s : 540292602711 4.5%
## [15] a + s                     : 0.6918868    4.88%
## [16] p + s                     : 0.1814085    0.57%
## [17] a + p + s                 : 0.1155655    3.48%
## [18] a + p + a:p + s          : 0.03811532   1.79%
##
## Against denominator:
##   rt ~ s
## ---
## Bayes factor type: BFlinearModel, JZS

```

Table 2

The Bayes factor ANOVA results for the hypothetical example.

Figure Captions

Figure 1. Updating with Bayes' rule. **A, B:** Prior and posterior for 7 heads in 10 flips and for 70 heads in 100 flips, respectively. **C, D.** Updating factors for 7 heads in 10 flips and 70 in 100 flips, respectively.

Figure 2. **A, B:** The probability of outcomes under the general model, \mathcal{M}_A and under the fair-coin model, \mathcal{M}_B , respectively. **C:** The ratio of these probabilities is the Bayes factor between the models. This ratio is plotted on a log scale.

Figure 3. The effect of shared priors is to shrink estimates away from extreme values. The top row shows four true mean values, which are equally spaced on the axis of measurement (time). The data, comprised of 10 observations per level, are shown directly below the true means, and there is a separate line for each group. Conventional sample mean and shared-prior estimates are shown at the bottom, and the shrinkage of the shared estimates toward the grand mean is evident.

Figure 4. The key specification in the ANOVA models is the scale of variability in effect sizes, h^2 . **A** Distribution of effect size for four values of h . Values of $h = 10$ and $h = .02$, top left and bottom right panels, respectively, are too extreme to be useful. Values of $h = 1$ and $h = .2$, top right and bottom left panels respectively, serve as reasonable limits in most contexts. **B** Bayes factor as a function of prior specification (h) shows that while the specification matters as it must, the changes across reasonable specifications (shown as the shaded area) are not great, especially when compared to the sensitivity of the Bayes factor on the data (t -values). The sample size of the data is $N = 50$.

Figure 5. The construction of Bayes factors for the one-way model with two levels. **A** The null and effect models on effect size ($h = .5$). **B.** The predictive densities for each model. **C.** The Bayes factor as a function of observed effect size.

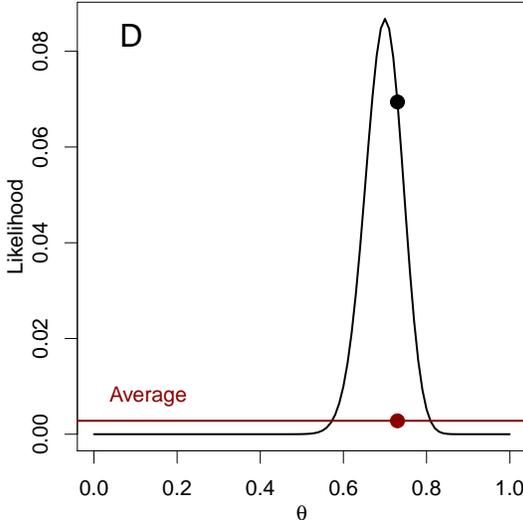
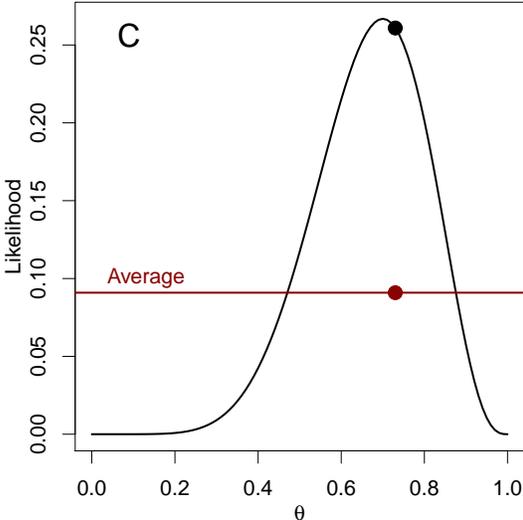
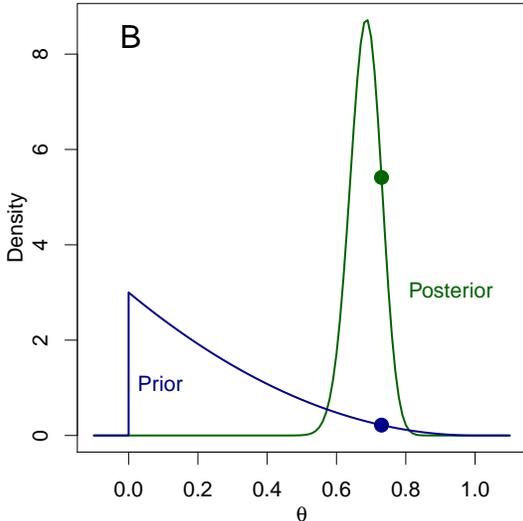
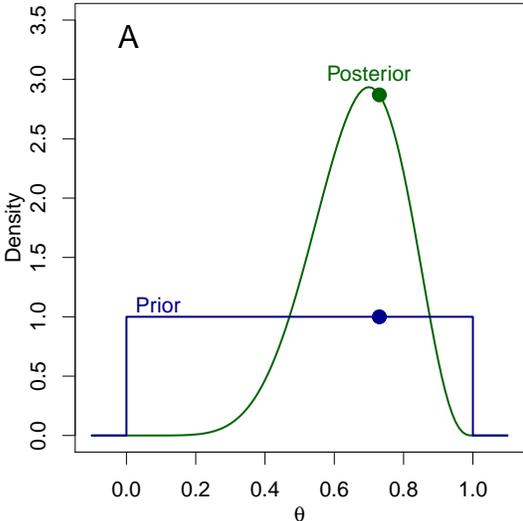
Figure 6. A. An interaction without a main effects for the price offered for a pint of ice-cream as a function of sugar and fat content. *B.* The same dependency across a continua of levels. From this panel it is clear that the lack of main effects reflects an implausibly fortuitous choice of levels.

Figure 7. Cell means for the hypothetical data set in which age (younger vs. older adult), distance (1 through 4), and presentation (digit vs. word) served as fixed effects and people served as random effects. Age entered as a between-subject variable while distance and presentation were manipulated in a within-subject fashion.

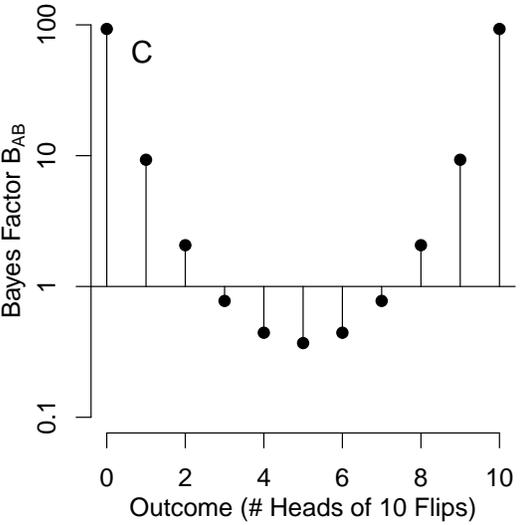
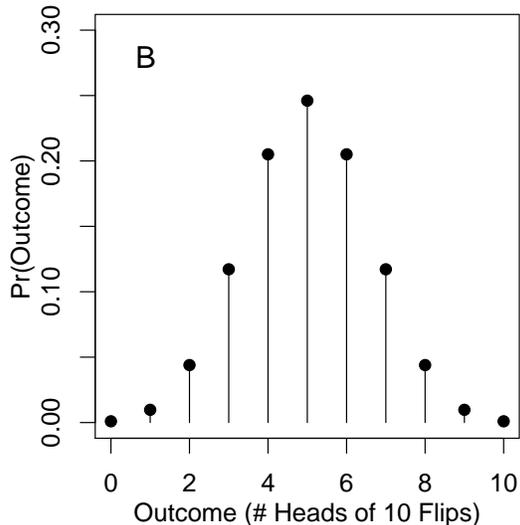
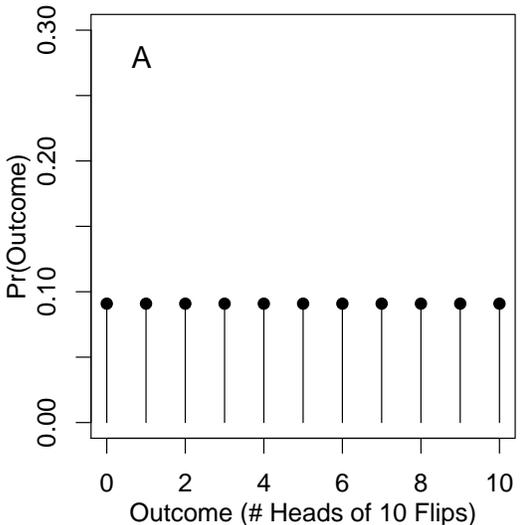
Figure 8. Bayes factors for the six most preferred models. The Bayes factors are relative to the null model with participant random effects. The (very small, barely visible) error bars show the estimation error for each Bayes factor.

Figure 9. A screen shot from JASP, an easy-to-use graphical statistical package that provides Bayesian inference. JASP uses the `BayesFactor` package and is suitable for researchers more comfortable in SPSS than R. Variable selection is straightforward, and multiple output panels allow researchers to make nuanced statements about their data.

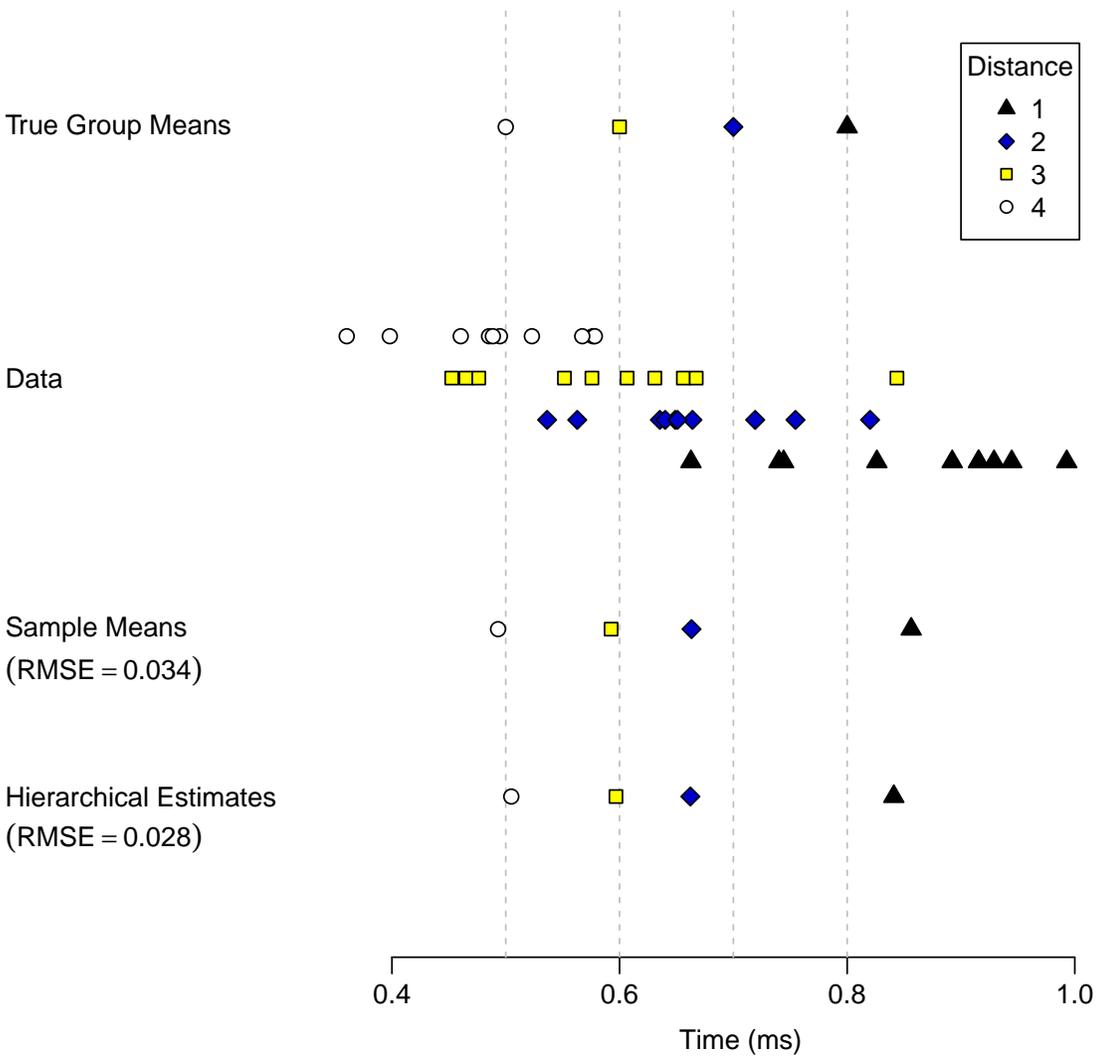
Bayesian Analysis, Figure 1



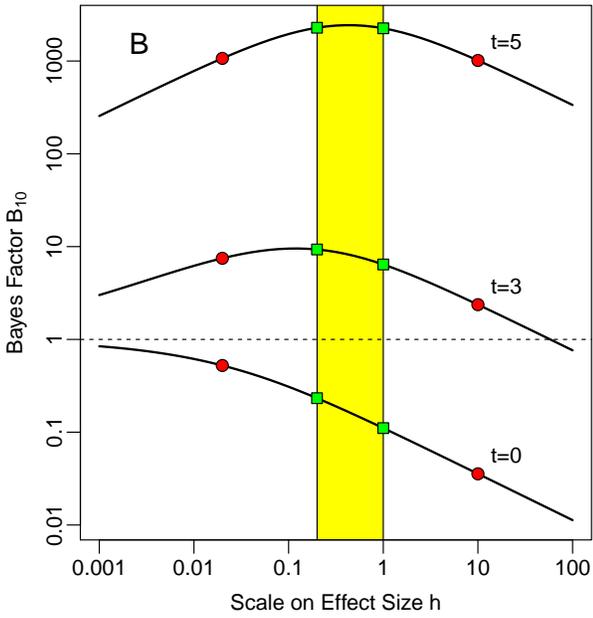
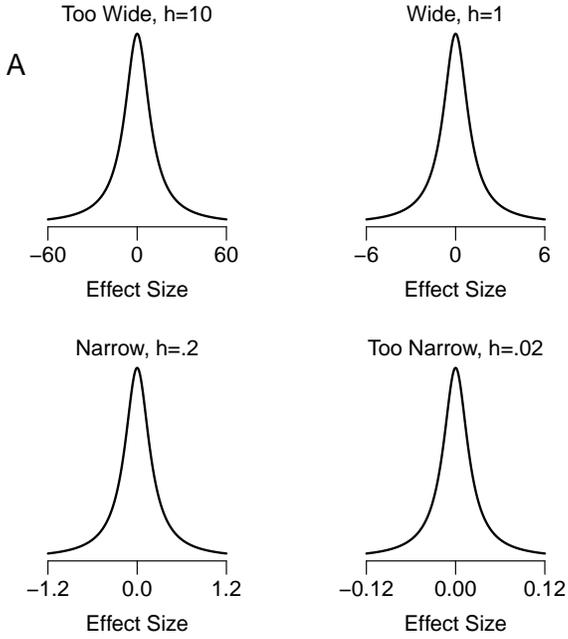
Bayesian Analysis, Figure 2



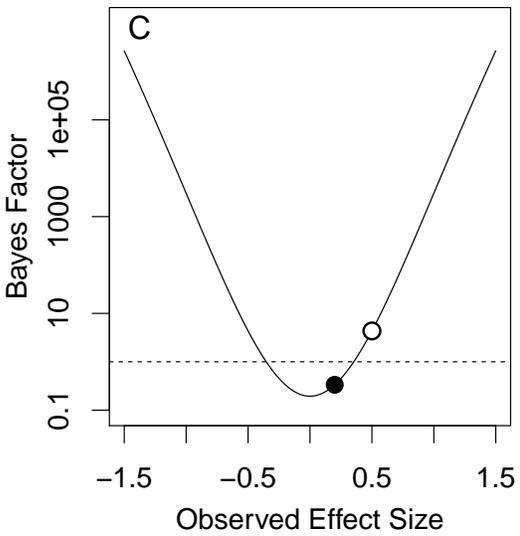
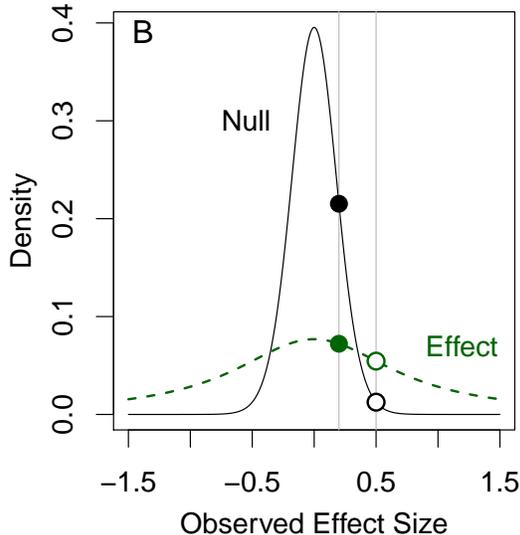
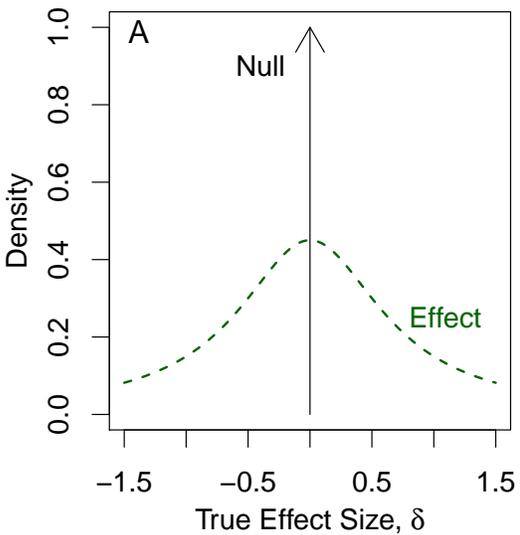
Bayesian Analysis, Figure 3



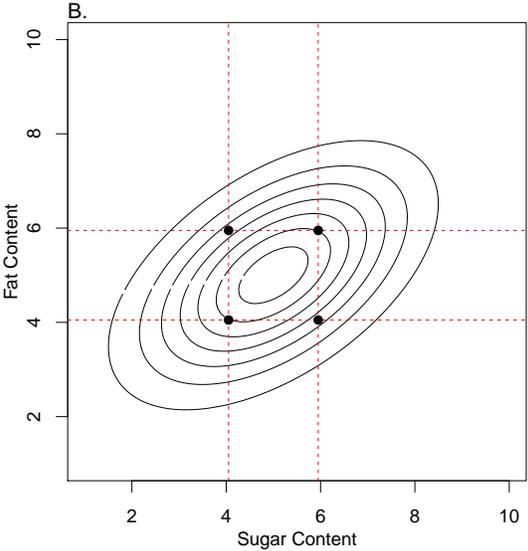
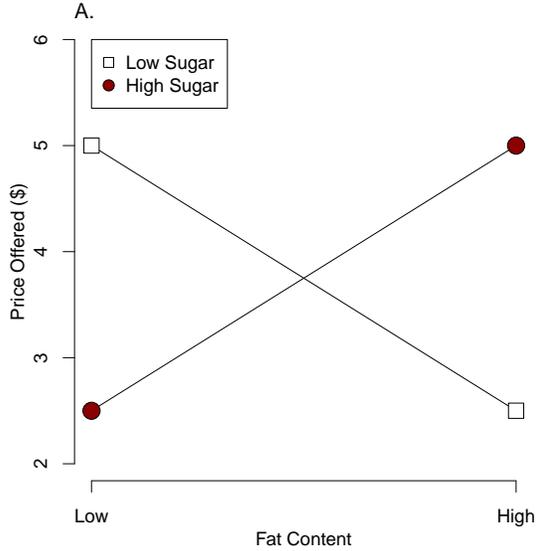
Bayesian Analysis, Figure 4

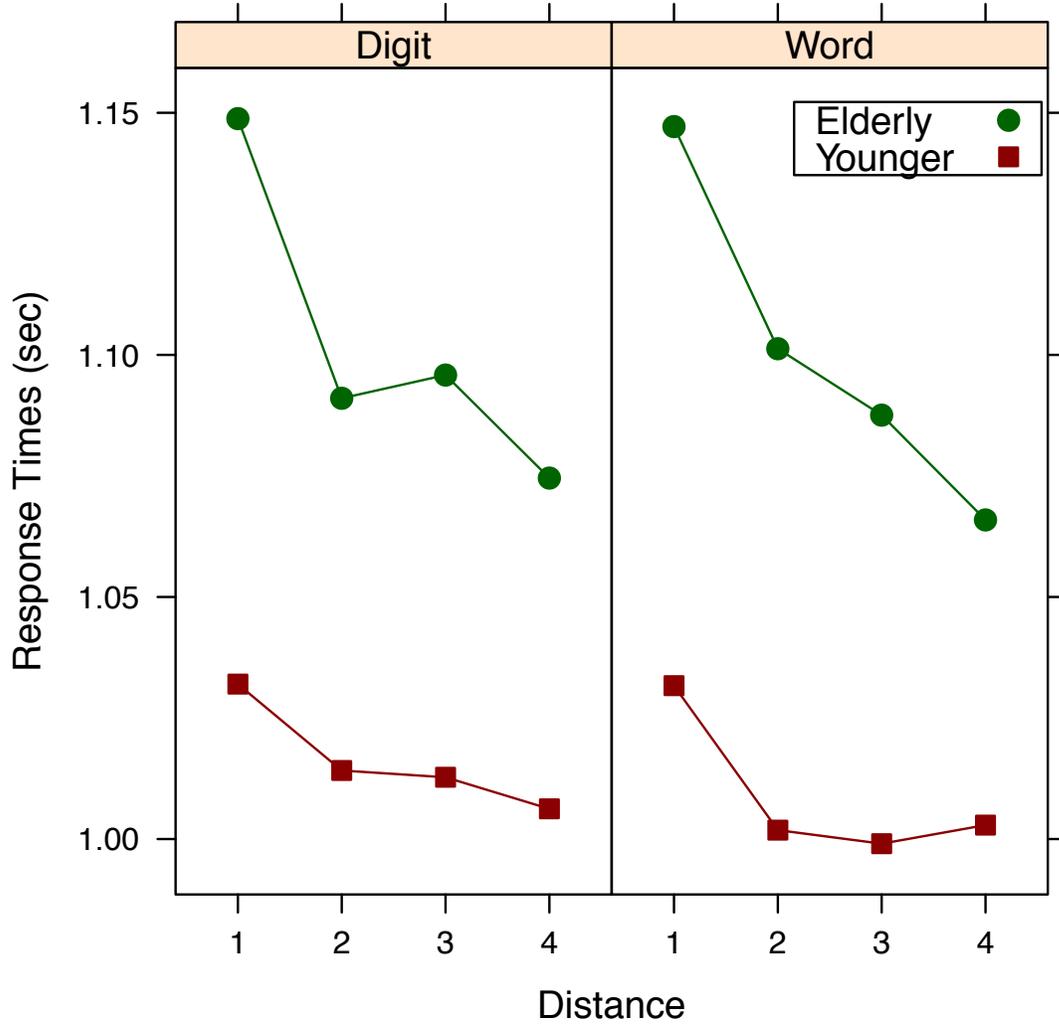


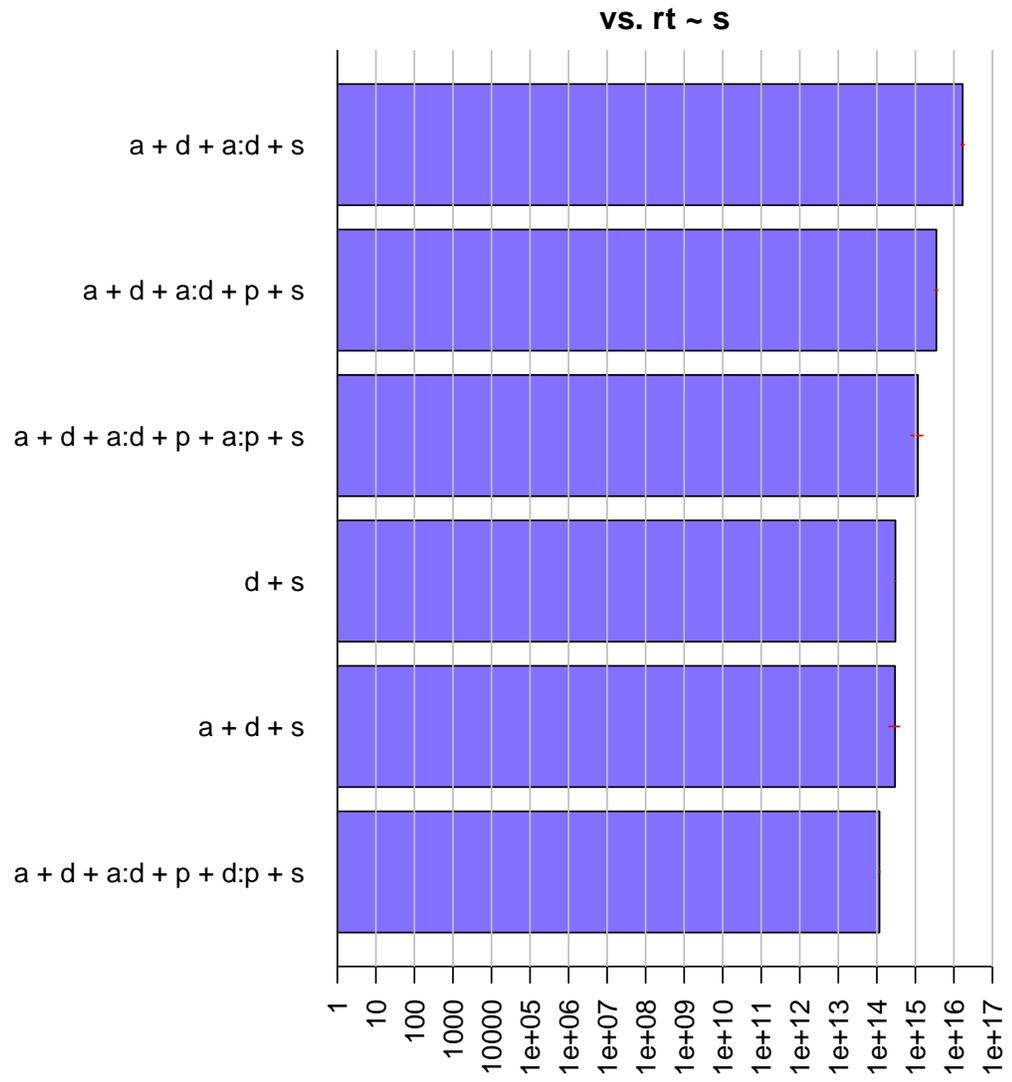
Bayesian Analysis, Figure 5



Bayesian Analysis, Figure 6







Bayesian Analysis, Figure 9

