A Model-Based Test for Treatment Effects with Probabilistic Classifications

In press at *Psychological Methods*

Daniel R. Cavagnaro

Mihaylo College of Business and Economics, California State University, Fullerton

Clintin P. Davis-Stober

Department of Psychological Sciences, University of Missouri

Author Note

Correspondence: dcavagnaro@fullerton.edu. Clintin P. Davis-Stober received support from the National Science Foundation (SES-14-59866) and the National Institutes of Health (K25AA024182).

Abstract

Within modern psychology, computational and statistical models play an important role in describing a wide variety of human behavior. Model selection analyses are typically used to classify individuals according to the model(s) that best describe their behavior. These classifications are inherently probabilistic, which presents challenges for performing group-level analyses, such as quantifying the effect of an experimental manipulation. We answer this challenge by presenting a method for quantifying treatment effects in terms of distributional changes in model-based (i.e., probabilistic) classifications across treatment conditions. The method uses hierarchical Bayesian mixture modeling to incorporate classification uncertainty at the individual level into the test for a treatment effect at the group level. We illustrate the method with several worked examples, including a reanalysis of the data from Kellen et al. (2017), and analyze its performance more generally through simulation studies. Our simulations show that the method is both more powerful and less prone to type-1 errors than Fisher's exact test when classifications are uncertain. In the special case where classifications are deterministic, we find a near-perfect power-law relationship between the Bayes factor, derived from our method, and the p-value obtained from Fisher's exact test. We provide code in an online supplement that allows researchers to apply the method to their own data.

Keywords: Bayesian, Model Selection, Classification, Treatment Effects

A Model-Based Test for Treatment Effects with Probabilistic Classifications

In press at *Psychological Methods*

Within psychology, treatment effects are typically formalized in terms of changes in a dependent variable across a set of experimental conditions. When the dependent variable is measured on an interval scale, we can test for changes in its population mean under the general linear model (e.g., Maxwell & Delaney, 2004) or related approaches. On the other hand, when the dependent variable is categorical (i.e., discrete and non-ordinal), the relevant change is in the *distribution* of the dependent variable, which is typically assessed using a chi-squared test of a contingency table or related methods for categorical data analysis (e.g., Klugkist et al., 2010).

As a simple example, suppose we were interested in whether older adults have different risk preferences than younger adults when choosing among investment products. One way to proceed would be to recruit a sample of older adults and a sample of younger adults. Suppose the researcher had all participants complete an experiment where they are repeatedly offered different sets of investment products and are asked to indicate their preferred product. This researcher could then use these data to classify each participant according to one of three distinct risk attitudes: 1) risk seeking (prefers risky investments), 2) risk averse (prefers safe investments) or 3) risk neutral (chooses by the expected return of the investment). Clearly we can expect a certain degree of heterogeneity in the classifications within each age group. In other words, it is highly unlikely that all participants within an age group will be classified according to the same risk attitude. The hypothesis to test is whether the distribution of risk attitude classifications is the same for both age groups. The frequency of each risk attitude classification in each age group could be summarized in a contingency table, and a chi-squared test could be used to assess whether there was sufficient statistical evidence to conclude that the distribution of risk attitudes was different across the age groups[1].

[1] Many authors have noted that measuring risk (or similar) on a unidimensional scale and averaging across

In the above illustration, we treated the classifications as deterministic (i.e., measured without error), which made a chi-squared test appropriate. However, many classifications in psychological research are actually probabilistic (i.e., noisy or uncertain). For example, we later re-analyze the data of Kellen et al. (2017), which examined differences in risk attitude among young and older adults using a probabilistic classification method. In clinical psychiatric research, classification of mental illness is increasingly informed by parametric fits to computational models of generative psychological processes (Wiecki et al., 2015; Aranovich et al., 2017). In vision research, classification of visual impairments such as amblyopia are based on statistical estimation of a contrast sensitivity function (Hou et al., 2010; Lesmes et al., 2010). In judgment and decision making research, statistical models are used to classify decision makers according to decision-making strategies such as exemplar or prototype categorization (Scheibehenne & Pachur, 2015), transitive or intransitive preference (Cavagnaro & Davis-Stober, 2014), optimal exploitation/exploration or heuristic choice in bandit problems (Steyvers et al., 2009), and reinforcement learning or heuristic search in the Iowa Gambling Task (Worthy et al., 2013; Bishara et al., 2009). In each of these examples, categorical classifications are based on either the model with the highest value of a selection statistic, such as the Akaike Information Criterion (AIC, Akaike, 1976), or on the parameter of a model with the highest maximized likelihood (e.g., Bröder & Schiffer, 2003; Glöckner, 2009). These statistics are inherently probabilistic, since they are based on random samples, hence the classifications based on them are probabilistic as well.

Probabilistic classifications present formidable challenges for statistical tests of distributional changes. Standard methods for categorical data analysis, such as the chi-squared test or Fisher's exact test, are based on tabulating the frequency of each category in each treatment group. Such frequencies are not well defined when the

participants can yield highly misleading results, especially when group members are heterogenous in their underlying risk attitudes (e.g., Estes, 1956; Luce, 2000). Therefore, this analysis cannot be reduced to a between-groups t-test.

classifications are probabilistic. Although one could simply dichotomize the probabilities to facilitate such an analysis, numerous studies have documented significant problems with this practice (e.g., Cohen, 1983; Maxwell & Delaney, 1993; MacCallum et al., 2002; Streiner, 2002; Young, 2016). Problems include the loss of effect size and power, spurious statistical significance, and generally misleading results.[2]For example, suppose that in one experimental condition of a binary choice experiment, every participant was assessed to be 10% likely to have intransitive (i.e., irrational) preferences, while in the other condition, every participant was assessed to be 40% likely to have intransitive preferences. If these classifications were dichotomized then every participant in both conditions would seem to have transitive preferences, hence we would conclude that there was no treatment effect, despite the fact that participants in the second condition were four times as likely to have intransitive preferences as participants in the first condition. Similarly, a researcher could encounter the case where two (or more) classifications yield nearly identical AIC (or BIC) values, with one just barely larger than the others. In this case, simply selecting the best performing classification and discarding the others for purposes of statistical analyses is, quite simply, ignoring potentially important information. Further, if two or more classifications are close in AIC values on one sample, it is possible that a different classification would be selected as best performing on a second sample of data - leading to a problem of irreplicable results.

To address these challenges, we present a Bayesian approach to evaluating distributional changes in categorical classifications that are probabilistic, i.e., noisy or uncertain. Our approach is based on Bayesian model selection. In the simplest case, we formalize two competing hypotheses as Bayesian hierarchical models: (1) that there is a

[2] In some cases, one could use multivariate multiple regression (i.e., the general linear model) as an alternative to dichotomizing. The analysis would treat the classification probabilities as dependent variables and use dichotomous predictors to represent the experimental conditions. However, this approach has many undesirable properties due to probabilities being bounded between zero and one: the model can predict probabilities greater than zero or larger than one, the effects of the explanatory variables will tend to be nonlinear, and the variance will tend to decrease as the mean approaches one of the boundaries, resulting in errors that are heterosketastic and non-normal. Therefore, this approach can only be justified in a few, limited cases.

treatment effect, and (2) that there is not a treatment effect. We then use Bayesian model selection to quantify the evidence for selecting between them. The hierarchical models differ in their assumptions about the distribution of classifications across participants. In the null hypothesis, it is assumed that the distribution is the same in both treatment groups (i.e., no treatment effect), while in the alternative it is assumed that the distribution may vary between groups (i.e., a treatment effect). These assumptions are implemented with equality constraints (or a lack thereof) on the parameters governing the relative frequency of each classification in each condition. We also show how this approach extends naturally to testing more nuanced hypotheses about treatment effects in experiments with multiple treatment conditions and crossed factors.

Our approach to testing for a treatment effect can be conceptualized as an application of Occam's razor to psychological research – we should choose the simplest model that explains the data well (Myung & Pitt, 1997). In this case, the simpler model is the null hypothesis because it has fewer free parameters. Therefore, we should not select the more complex alternative (i.e., conclude that there is a treatment effect) unless the data justify that added complexity. Bayesian model selection provides a useful and appropriate tool for evaluating models according to this principle (Myung, 2000).

The Bayesian framework of our approach allows the researcher to infer statistical support for either the null or the alternative hypothesis, and to interpret the results intuitively in terms of predicting future data (Pitt & Myung, 2002). Evidence in favor of the alternative hypothesis suggests that better predictions of future data are likely to be obtained by assuming the existence of a treatment effect. Evidence in favor of the null hypothesis suggests the opposite. Rouder et al. (2009) and others have argued that being able to state evidence for the null hypothesis is critical for identifying invariance properties, i.e., those elements that stay constant when others change, which are critical to the advancement of psychological science. Said differently, our method allows researchers to accept the null hypothesis when warranted by the data, unlike classical approaches.

To briefly outline the how the method works, consider an experiment with a finite set of experimental conditions. These may be different levels of a single factor, or there may be crossed factors. Suppose that each participant is assigned to a single condition, with possibly different numbers of participants in each condition. Next, suppose that there is a finite set of latent classifications (i.e., individual-level models) under consideration for each participant, which can be evaluated using standard model selection statistics such as the AIC or Bayesian information criterion (BIC). For example, the classifications could be two decision making strategies: one a utility model and the other a lexicographic heuristic. Our method begins where typical model selection analyses leave off. That is, we assume that a model selection statistic such as the AIC or BIC has been computed for each classification, for each participant. With those in hand, the method comprises three steps, which are illustrated in the flowchart in Figure 1. The first step is to estimate the probability of each classification for each participant, which can be done with a straightforward transformation of standard model selection statistics. The second step is to estimate the distribution of classifications within each condition under the two different hierarchical Bayesian mixture models described above. The final step is to compute the Bayes factor between the two models, which yield the posterior odds of a treatment effect. The three steps of the method can be viewed as applying at increasing levels of generality: Step 1 applies to each participant, Step 2 applies to each condition, and Step 3 applies to the entire experiment.

The approach we propose is mathematically sophisticated and would be computationally intensive to evaluate directly. However, we bring together a sequence of simplifications and approximations from the literature, which render it quite simple to implement. Essentially, the analysis boils down to summing the classification probabilities (relative likelihoods) across participants, and then performing low-dimensional Monte Carlo integration of a closed-form density function, which is straightforward to compute. In the online supplement to this article, we provide general Matlab code that allows the user to apply our method to an arbitrary number of models, participants, and experimental conditions. Further, this code allows users to specify arbitrary sets of equality constraints among the different experimental conditions. This allows users to apply our technique to higher-order designs with multiple factors, e.g., two-way designs with multiple levels. As a pedagogical tool, we provide an interactive spreadsheet for computing the group-level components of the analysis (i.e., estimating and selecting among the hierarchical models) for a simple design with two conditions, and for a fully-crossed design with two factors and two levels of each factor. Users can input the probability of each classification for each participant and the spreadsheet will instantly show the Bayes factor.

The paper is organized as follows. First, we present a mixture modeling framework for probabilistic classification. Then, we describe our method for evaluating treatment effects within this framework. We then illustrate how the method works with two examples. The first is a toy example that demonstrates how the method can pick up treatment effects that would be missed if data were dichotomized.[3] The second worked example reanalyzes the data from Kellen et al. (2017), which investigated differences in risk attitude among older and younger adults within the context of a decision making under risk task. Next we analyze the performance of the method more generally by carrying out Monte Carlo simulations. The first set of simulations compares the type-1 and type-2 error rates of standard approaches (e.g., Fisher's exact test on dichotomized data) with Bayes factor results from our proposed method. We find that our method is virtually immune to type-1 error when the sample size is sufficiently large (e.g., greater than 30 in each condition), while achieving lower type-2 error rates than Fisher's exact test with the same sample size. The next set considers the special case of two classifications that are revealed with certainty (i.e., all probabilities equal to zero or one), in which we find a near-perfect power-law relationship between the Bayes factor, derived from our method, and the p-value, obtained from Fisher's exact test ($R^2 > 0.99$). While our approach is much more general, we conclude that it works comparably to existing methods in cases where both

[3] All worked examples can be reproduced using the pre-programmed Excel spreadsheets provided in the online supplement.

apply. Finally, we generalize the method to multiple conditions and higher order designs. We conclude with a summary and discussion.

Mixture models of probabilistic classifications

We begin by defining a hierarchical model for N participants in an experiment, each of whom generate a vector of data, denoted x_n for $n = 1, \ldots, N$. The model consists of two components: one that operates at the level of the participant, and another that operates at the level of the group. The participant-level component is a parametric likelihood function for a single participant's data, which we refer to as a *classification*. The group-level component is a Dirichlet-multinomial distribution over classifications. In that sense, our basic setup is comparable to that of a Latent Dirichlet Allocation (LDA, Blei et al., 2003). The basic idea of the model is that each participant is represented as a random mixture over latent classifications, where each classification is itself a latent parameter model of the data.

For example, in the opening, we described an experiment aimed at testing whether older adults have different risk preferences than younger adults. The treatment groups in that experiment were older adults and younger adults. The data were choices between financial products, and the classifications of interest were "risk seeking," "risk averse," and "risk neutral." Under our modeling framework, each participant would be assumed to have a latent classification, either risk seeking, risk averse, or risk neutral, which could be estimated probabilistically based on their observed choices between financial products. In turn, each treatment group would be characterized by a mixture of classifications representing the probability of each classification among members of the group.

Formally, suppose that there is a finite set of M-many possible classifications, denoted c_m for $m = 1, \ldots, M$. Suppose that each classification corresponds to a particular Bayesian latent parameter model defined by a parametric likelihood function for the data and a prior distribution over the latent parameter. We write $\mathcal{L}_m(\theta|x_n) = p(x_n|\theta, c_m)$ for the parametric likelihood function, where θ is the latent parameter of the classification, and

$p(\theta|c_m)$ for the prior. In principle, the likelihood function does not need to have a closed form as long as it can be simulated, such as in a diffusion or neural network model. What we refer to as a classification here is typically called a model, but to avoid ambiguity we will reserve that term for the group-level hierarchical model, of which the latent classifications are a component.

For the group-level component, let $\boldsymbol{\phi} \sim Dirichlet(\boldsymbol{\alpha})$ be a distribution over classifications, which we call the *classification mixture*, and let $\mathcal{C}_n \sim Multinomial(\boldsymbol{\phi})$ be the latent classification of participant n. The M-dimensional Dirichlet distribution has parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ and is conjugate to the multinomial distribution, which will be critical in our development of an efficient algorithm for Bayesian statistical inference within this framework.

Combining the participant-level and group-level components yields the following, generative, probabilistic model:

1. Generate a classification mixture $\boldsymbol{\phi} \sim Dirichlet(\boldsymbol{\alpha})$
2. For each participant, $n = 1, \ldots, N$:
   (a) Generate a classification $\mathcal{C}_n \sim Multinomial(\boldsymbol{\phi})$
   (b) Generate a parameter θ_n from $p(\theta_n|\mathcal{C}_n)$
   (c) Generate data x_n from $p(x_n|\theta_n, \mathcal{C}_n)$.

Given the Dirichlet parameter $\boldsymbol{\alpha}$, the joint distribution of a classification mixture $\boldsymbol{\phi}$, a set of N latent classifications $\boldsymbol{\mathcal{C}}$, a set of N latent classification parameters $\boldsymbol{\theta}$, and a set of N data sets $\mathbf{x}$, is given by:

$$p(\boldsymbol{\phi}, \boldsymbol{\mathcal{C}}, \boldsymbol{\theta}, \mathbf{x}) = p(\boldsymbol{\phi}) \prod_{n=1}^{N} p(\mathcal{C}_n|\boldsymbol{\phi}) p(\theta_n|\mathcal{C}_n) p(x_n|\theta_n). \tag{1}$$

Now, consider a generic experiment aimed at testing for a treatment effect. Suppose that there are K conditions (e.g., K levels of a single factor). Let N_k denote the number of participants in condition k, for $k = 1, \ldots, K$, and let $x_{n,k}$ denote the observed data from participant n in condition k. With only a minor change in notation, adding subscripts for the conditions, the model defined above can be applied to each condition separately. Let $\boldsymbol{\phi}_k \sim \mathit{Dirichlet}(\boldsymbol{\alpha}_k)$ be the classification mixture in condition k, and let $\mathcal{C}_{n,k}$ and $\theta_{n,k}$ be the latent classification and parameter for participant n in condition k, respectively. Then, assuming that the conditions are independent, the joint distribution of the data from all conditions is a product of distributions of the form defined in Equation 1:

$$p\left[(\boldsymbol{\phi}_1, \boldsymbol{\mathcal{C}}_1, \boldsymbol{\theta}_1, \mathbf{x}_1), \ldots, (\boldsymbol{\phi}_K, \boldsymbol{\mathcal{C}}_K, \boldsymbol{\theta}_K, \mathbf{x}_K)\right] = \prod_{k=1}^{K} p(\boldsymbol{\phi}_k, \boldsymbol{\mathcal{C}}_k, \boldsymbol{\theta}_k, \mathbf{x}_k). \tag{2}$$

We refer to the model defined by Equation 2 as the *encompassing model*, and denote it by H_e. It allows the classification mixtures to differ across conditions (i.e., it instantiates the hypothesis that there is a treatment effect). A graphical representation of H_e is shown on the left-hand side of Figure 2.

If there are no treatment effects in an experiment, then the classification mixture should be the same in every condition. Therefore, we instantiate the hypothesis that there is no treatment effect with an equality constraint on H_e. Specifically, we define a model H_0, equivalent to H_e, except with the constraint that $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 = \ldots = \boldsymbol{\phi}_K$. Under this constraint, there is just one Dirichlet parameter ($\boldsymbol{\alpha}$) and one free mixture parameter ($\boldsymbol{\phi}$), which can be conditioned on the data from every participant, regardless of their group membership. Thus, under H_0, as under H_e, the joint distribution for the full experiment is a product of distributions of the form in Equation 1, but it can be expressed as

$$p\left[\boldsymbol{\phi}, (\boldsymbol{\mathcal{C}}_1, \boldsymbol{\theta}_1, \mathbf{x}_1), \ldots, (\boldsymbol{\mathcal{C}}_K, \boldsymbol{\theta}_K, \mathbf{x}_K)\right] = p(\boldsymbol{\phi}) \prod_{k=1}^{K} \prod_{n=1}^{N_k} p(\mathcal{C}_{n,k} | \boldsymbol{\phi}) p(\theta_{n,k} | \mathcal{C}_{n,k}) p(x_{n,k} | \theta_{n,k}).$$

A graphical representations of H_0 is shown on the right-hand of Figure 2.

We propose to evaluate the evidence for (and against) a treatment effect using model selection between the hierarchical models H_0 and H_e. Specifically, we propose to compute the odds ratio for H_0 relative to H_e, given the data from the experiment, which is known as the Bayes factor. Formally, under the assumption that both models have the same prior likelihood, the Bayes factor for H_0 relative to H_e is defined as

$$BF_{0e} = \frac{p(D|H_0)}{p(D|H_e)}, \tag{3}$$

where D denotes the combined data from every participant in the experiment. The numerator and denominator of the Bayes factor in Equation 3 are the marginal posterior probabilities of the data given each model. Each is obtained by integrating the corresponding joint density function over Dirichlet parameters, classification mixtures, latent classifications, and latent parameters.

The Bayes factor between H_0 and H_e is a quotient of high-dimensional integrals, which does not have a closed form solution in general. Evaluating it by direct simulation of the generative model would be extremely computationally intensive. Therefore, we present an algorithm based on a sequence of computational techniques and approximations, which allows the Bayes factor for these models to be obtained easily from widely used model selection statistics, with no more than low-dimensional Monte Carlo integration. The algorithm proceeds according to the three steps described earlier and shown in Figure 1. The next section presents each step in detail.

Method for evaluating treatment effects

Step 1: Estimate classification probabilities for each participant

The first step in our method is to estimate the normalized posterior likelihood of each classification, for each participant, which we call the *classification probabilities*. These can be derived from widely used model selection statistics such as the AIC, BIC, or Bayes factor. We discuss each in turn. In the online supplement, we provide code for carrying out these transformations.

Using the AIC. The AIC of c_m for participant x_n is

$$AIC_n(c_m) = -2 \ln p(x_n|\hat{\theta}, c_m) + 2z, \tag{4}$$

where $\hat{\theta} = \arg\max_\theta \mathcal{L}(\theta|x_n, c_m)$ is the maximum likelihood estimate of θ for data x_n, and z is the number of parameters in the classification c_m. The AIC provides an asymptotically unbiased estimator of the expected Kullback-Leibler discrepancy between the generating model and the fitted approximating model. For participant n, the difference in AIC value between classification c_m and the worst performing model is

$$\Delta_m(AIC_n) = AIC_n(c_m) - \max_m AIC_n(c_m). \tag{5}$$

The relative likelihood of classification m for participant n can be derived, up to a normalizing constant, from the delta AIC as

$$\mathcal{L}(c_m|x_n) \propto \exp\{-\frac{1}{2}\Delta_m(AIC_n)\}, \tag{6}$$

where $\propto$ stands for "is proportional to." These relative likelihoods can then be normalized by dividing by the sum of the relative likelihoods for all classifications under consideration, yielding the normalized posterior probabilities (Wagenmakers & Farrell, 2004). Written out explicitly, the formula is

$$p(c_m|x_{n,k}) = \frac{\exp\{-\frac{1}{2}\Delta_m(AIC_n)\}}{\sum_{m=1}^{M} \exp\{-\frac{1}{2}\Delta_m(AIC_n)\}}. \tag{7}$$

Using the Bayes factor. Our ultimate goal is to obtain the Bayes factor between the mixture models of probabilistic classification defined earlier, for the entire group. However, the Bayes factor of interest in this step is that between different latent classifications for a given participant. It is simply the ratio of the posterior marginal likelihoods of the data given two different classifications, and it can be obtained by a number of different methods, including the product space method (Lodewyckx et al., 2011), the encompassing prior method (Klugkist & Hoijtink, 2007), or the Savage Dickey density ratio (Verdinelli & Wasserman, 1995; Wagenmakers et al., 2010). Many of these methods can be implemented in off-the-shelf software such WinBugs (Lunn et al., 2000) or JASP (Love et al., 2015).

For a given participant, assume that the Bayes factor for each classification has been calculated relative to a common, reference classification. Write $BF_{m0}(x_{n,k})$ for the Bayes factor for classification c_m relative to the reference classification, conditioned on the data $x_{n,k}$. Then, the normalized posterior likelihood of c_m for participant n in condition k can be derived as

$$p(c_m|x_{n,k}) = \frac{BF_{m0}(x_{n,k})}{\sum_{j=1}^{M} BF_{j0}(x_{n,k})}. \tag{8}$$

For classifications based on more complex models, such as those with many parameters, or process models, the Bayes factor may be intractable. In these cases, the classification probabilities can be approximated from simpler model selection statistics such as the AIC or BIC.

Using the BIC. The BIC of c_m for participant n is

$$BIC_n(c_m) = -2\ln p(x_n|\hat{\theta}, c_m) + z\ln|x_n|, \tag{9}$$

where $|x_n|$ denotes the sample size for participant n. The difference between the BICs of two models approximates minus twice the log of the Bayes factor between those models. That is, $BIC(c_1) - BIC(c_2) \approx -2\ln BF_{12}$. From this approximation, the relative posterior likelihoods of the candidate models can be derived as above using the Bayes factor.

Step 2: Estimate the classification mixture in each condition

After obtaining the posterior probability of each classification for each participant, the next step is to combine these results across participants within each condition. We do this using a random effects framework that furnishes a probability density over the classifications themselves. In this framework, the classification for each participant is treated as a draw from a population-level multinomial distribution. Essentially, we use the classification probabilities derived in step 1 to estimate the classification mixtures within models H_0 and H_e defined in the previous section.

Formally, let $\boldsymbol{\phi}_k = \phi_{1,k}, \phi_{2,k}, \ldots, \phi_{M,k}$ be the *classification mixture* in condition k, where $\phi_{m,k}$ is the population probability of classification m in condition k. Then, the classification of participant n in condition k is given by the multinomial random variable

$$\mathcal{C}_{n,k} \sim Multinomial(\boldsymbol{\phi}_k).$$

To enable Bayesian estimation of each $\boldsymbol{\phi}_k$, we define a prior probability distribution over it. Specifically, we define a Dirichlet prior with parameters $\boldsymbol{\alpha_k} = (\alpha_{1,k}, \ldots, \alpha_{M,k})$, as

$$p_0(\boldsymbol{\phi}_k) = Dir(\boldsymbol{\phi}_k; \boldsymbol{\alpha_k}) = \frac{1}{B(\boldsymbol{\alpha_k})} \prod_{m=1}^{M} \phi_{m,k}^{\alpha_{m,k}-1} \tag{10}$$

where $B(\boldsymbol{\alpha_k})$ is the multivariate Beta function, which can be expressed in terms of the gamma function as

$$B(\boldsymbol{\alpha_k}) = \frac{\prod_{m=1}^{M} \Gamma(\alpha_{m,k})}{\Gamma(\sum_{m=1}^{M} \alpha_{m,k})}. \tag{11}$$

In the absence of any prior knowledge of the treatment effect, we recommend assuming $\alpha_{1,k} = \alpha_{2,k} = \ldots = \alpha_{M,k} = 1$ in the prior distribution, in which case all possible vectors of multinomial parameters are equally likely. This assumption is essential to our derivation of a simple formula for the Bayes factor in the next section.

Using the variational Bayesian method proposed by Stephan et al. (2009), the posterior distribution of $\boldsymbol{\phi}_k$ can be approximated as

$$p(\boldsymbol{\phi}_k | x_{1,k}, x_{2,k}, \ldots, x_{N,k}) \approx Dir(\boldsymbol{\phi}_k; \alpha'_{1,k}, \alpha'_{2,k} \ldots, \alpha'_{M,k}), \tag{12}$$

where

$$\alpha'_{m,k} = \alpha_{m,k} + \sum_{n=1}^{N_k} p(\mathcal{C}_{n,k} = c_m | x_{n,k}), \tag{13}$$

and $p(c_m|x_{n,k})_{m=1,\ldots,M}$ are the probabilities of each classification for participant n in condition k, which were obtained in the previous step[4]. The quantity $\sum_{n=1}^{N_k} p(\mathcal{C}_{n,k} = c_m|x_{n,k})$ in Equation 13 can be interpreted as the expected number of participants in condition k who are classified as c_m, although this value will typically not be an integer. Therefore, Equation 13 essentially adds these 'data counts' of the classifications to the 'prior counts,' given by $\boldsymbol{\alpha_k}$.

The estimation process described above for the hierarchical model H_e yields separate estimates of the classification mixture in each condition. Specifically, the fit of this hierarchical model to the experimental data provides a point estimate of the relative frequency of each classification in each condition, or equivalently, an estimate of the probability of each classification for a randomly selected participant in each condition. The probability of a given classification in a given condition can be estimated by either the mean of the posterior distribution, which is given by $\frac{\alpha'_j}{\alpha'_1+\ldots+\alpha'_M}, j = 1, \ldots, M$, or the mode of the posterior distribution, which is given by $\frac{\alpha'_j-1}{\alpha'_1+\ldots+\alpha'_M-M}, j = 1, \ldots, M$.

When the estimated classification mixtures vary across conditions, it suggests that there is a treatment effect. However, a competing explanation for the data is that the true mixtures are actually the same in every condition, and that any apparent differences are due to sampling error, approximation error, or stochastic variation. We instantiate this competing hypothesis with the hierarchical model H_0, which is nested within H_e by an equality constraint on the classification mixtures. Specifically, we set $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 = \ldots = \boldsymbol{\phi}_N$. Hence, H_0 can be parameterized with a single multinomial distribution with prior distribution $p_0(\boldsymbol{\phi}|H_0) = Dir(\boldsymbol{\phi}; \boldsymbol{\alpha}_0)$. This common multinomial distribution can be estimated just as above by treating all of the data as coming from a single condition.

[4] It is worth noting that this is only an approximate posterior and the true posterior does not necessarily have the exact form of a Dirichlet distribution. Nevertheless, the approximation proves to be quite precise in simulation. See Stephan et al. (2009) for details.

The third step in the method, described next, entails Bayesian model selection between H_0 and H_e. From a model complexity standpoint, the null model is more parsimonious (i.e., less complex) than the encompassing model because it has fewer free parameters. Conceptually, it is more parsimonious to model the data without the assumption of a treatment effect than with it. Therefore, H_0 will be favored in Bayesian model selection unless the improved fit of H_e, relative to H_0, is enough to overcome the complexity penalty.

Step 3: Test the equality constraint on classification mixtures

In this step, we compute the Bayes factor for H_0 relative to the H_e. To do so, we take advantage of the fact that H_0 is nested in H_e, which allows the Bayes factor to be computed using an encompassing prior generalization of the Savage-Dickey density ratio (Wetzels et al., 2010). Under this approach, the Bayes factor simplifies to the ratio of the posterior and prior (parameter) densities under the encompassing model, at the equality constraint.

More generally, the Savage-Dickey density ratio can be explained as follows. For a null hypothesis H_0 that posits a parameter value $\boldsymbol{\xi}$ to be constrained to some special value, i.e., $\boldsymbol{\xi} = \boldsymbol{\xi}_0$, and an alternative hypothesis that assumes $\boldsymbol{\xi}$ is free to vary, and writing D to denote the observed data, the Bayes factor for H_0 relative to H_e can be expressed as

$$BF_{0e} = \frac{p(D|H_0)}{p(D|H_e)} = \frac{p(\boldsymbol{\xi} = \boldsymbol{\xi}_0|D, H_e)}{p_0(\boldsymbol{\xi} = \boldsymbol{\xi}_0|H_e)}, \tag{14}$$

which equals the ratio of the heights for the posterior and the prior distribution for $\boldsymbol{\xi}$ at $\boldsymbol{\xi}_0$ in the encompassing model (Wagenmakers et al., 2010).

In the present case, the equality constraint in H_0 is between classification mixtures: $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 = \ldots = \boldsymbol{\phi}_N$. Although these are actually distributions, we treat them as parameters in the hierarchical models. The encompassing prior generalization of the Savage-Dickey density ratio (Wetzels et al., 2010) extends Equation 14 to this type of equality constraint, so that the Bayes factor can be computed analogously, as follows. Let Ω denote all possible classification mixtures (i.e., all possible values of $\boldsymbol{\phi}$). Then, under the encompassing model, the (marginal) *prior* density at the equality constraint is

$$p(\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 = \ldots = \boldsymbol{\phi}_K | H_e) = \int_\Omega \prod_{k=1}^{K} p_0(\boldsymbol{\phi}_k = \boldsymbol{\phi}) \, d\boldsymbol{\phi}, \tag{15}$$

and, writing $\mathbf{x}_k = \{x_{n,k}\}_{n=1,\ldots,N_k}$ to denote data for every participant in condition k, and letting $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the (marginal) *posterior* density at the equality constraint is

$$p(\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 = \ldots = \boldsymbol{\phi}_K | D, H_e) = \int_\Omega \prod_{k=1}^{K} p(\boldsymbol{\phi}_k = \boldsymbol{\phi} | \mathbf{x}_k) \, d\boldsymbol{\phi}. \tag{16}$$

The Bayes factor for H_0 relative to the H_e can be computed as the ratio of these densities:

$$BF_{0e} = \frac{\int_\Omega \prod_{k=1}^{K} p(\boldsymbol{\phi}_k = \boldsymbol{\phi} | \mathbf{x}_k) \, d\boldsymbol{\phi}}{\int_\Omega \prod_{k=1}^{K} p_0(\boldsymbol{\phi}_k = \boldsymbol{\phi}) \, d\boldsymbol{\phi}}.$$

This formula can be simplified further by noting that when the prior over classification mixtures is uniform in every condition, the density function in the denominator reduces to the constant value $\Gamma(M) = (M-1)!$, which yields

$$BF_{0e} = \Gamma(M)^{-K} \int_\Omega \prod_{k=1}^{K} p(\boldsymbol{\phi}_k = \boldsymbol{\phi} | \mathbf{x}_k) \, d\boldsymbol{\phi}. \tag{17}$$

The value of BF_{0e} is the odds ratio for H_0 relative to H_e. A Bayes factor of 1.0 indicates that both H_0 and H_e are equally likely. Larger values of BF_{0e} indicate stronger evidence *against* a treatment effect. In testing for treatment effects, it can be more intuitive to emphasize the evidence *for* a treatment effect instead of against it. In that case, we will use the reciprocal of BF_{0e}, denoted BF_{e0}, for which larger values indicate stronger evidence for a treatment effect.

The integral in Equation 17 is straightforward to compute via Monte Carlo integration because the posterior densities are Dirichlet with parameters given by Equations 12 and 13. It suffices to simulate a large, uniform sample of multinomial parameter vectors and compute the density at each one via the Dirichlet probability density function. The dimensionality of the integral is K, so if there are only two experimental conditions then this calculation can be carried out in spreadsheet software, such as the one provided in the online supplement. Our more general Matlab code in the online supplement allows this integral to be computed for an arbitrary number of experimental conditions and classifications.

Illustration of the method

Toy Example

Consider a hypothetical experiment aimed at testing whether an experimental manipulation affects the distribution of three classifications in a population. Suppose that each of 60 participants were randomly assigned to one of two experimental conditions: Condition 1 or Condition 2. Participants in both conditions answered a short questionnaire, and each participant's data were fit to all three classifications: A, B, and C. A model selection analysis was then performed, yielding the AIC of each classification for each participant.

A set of hypothetical results from this analysis are given in Table 1. The table shows the AIC of each classification for each participant in each condition. For emphasis, it also shows the preferred classification for each participant (i.e., the one with the lowest AIC).

To illustrate the importance of incorporating probabilistic information about classifications, we will analyze these data in two different ways[5]. First, without considering classifications as probabilistic, we simply count the number of participants in each condition for whom each classification had the lowest AIC, and summarize the results in a contingency table. The result is shown in Table 2. There, we see that in Condition 1, five participants each are classified as A and B, while 20 are classified as C, whereas in

Condition 2, 12 participants are classified as A, three as B, and 15 as C. The contingency table is too sparse to use a Chi-squared test, so we use the Freeman-Hilton extension of Fisher's exact test to assess whether there is a change in the classification mixtures across conditions (Soper, 2017). We find that the difference between conditions is not statistically significant ($p = 0.138$).

Next, we test for a treatment effect using our Bayesian method, taking into account the continuous information provided by the AIC (i.e., without dichotomizing). First, we transform the raw AIC values into classification probabilities using Equation 7. Table 3 shows shows the derived probability of each classification for each participant in each condition. Next, we compute the alpha parameters of the Dirichlet posteriors according to Equation (2). Essentially, there is one alpha parameter for each column of probabilities in Table 3, and each parameter is estimated as one plus the sum of the column. They are shown in the last row of Table 3. These parameters define the posterior densities over classification mixtures in each condition:

$$\begin{aligned} p(\boldsymbol{\phi}_1|\mathbf{x}_1) &= Dir(\boldsymbol{\phi}_1; 5.024, 6.721, 21.545) \\ p(\boldsymbol{\phi}_2|\mathbf{x}_2) &= Dir(\boldsymbol{\phi}_2; 13.966, 3.750, 15.283). \end{aligned}$$

Finally, we evaluate Equation 17 with respect to these posterior densities. The major operation in Equation 17 is the marginal of the product of the two densities, where the marginal is taken over all possible common mixtures for the two conditions. For this example, we used the spreadsheet in the online supplement to do Monte Carlo integration with 20,000 trials, but other software could be used, including the Matlab code that is also provided in the supplement. The constant outside the integral in Equation 17 is

[5] These data could also be analyzed with a two-sample t-test on the AIC values, across conditions. However this approach has several limitations. For one, its statistical properties rely on normality assumptions that are likely to be violated with classification probabilities in heterogeneous populations. More importantly, one would need to test for differences on each model separately (i.e., multiple comparisons), leading to either a loss of power or an inflated family-wise type-1 error rate. Finally, the nonlinear transformation of AIC values into classification probabilities may yield contradictory results.

$\Gamma(3)^{-2} = 0.25$, because there are 3 classifications and 2 conditions. Multiplying the constant and the integral yields a Bayes factor of approximately $BF_{0e} = 0.32$. Taking the reciprocal yields $BF_{e0} = 3.125$, which indicates an odds ratio of more than 3:1 in favor of a treatment effect.

Empirical demonstration: Re-Analysis of Kellen et al. (2017)

As an empirical illustration of our framework, we re-analyze data from Kellen et al. (2017) (hencefore KMD17). KMD17 investigated differences in risk attitude among older and younger adults within the context of a decision-making-under-risk task. The task involved binary choices on pairs of risky gambles. Based on the observed choices, KMD17 sought to identify each participant's risk-attitude profile: either risk seeking (preference for the riskier gamble in the pair), risk averse (preference for the less risky gamble), or risk neutral (indifferent between the two gambles).

Their experimental design consisted of two conditions, which were characterized by whether the participants were younger adults (n=30; median age: 22, range: 18 to 34) or older adults (n=30; median age: 65.5, range: 61 to 78). Every participant completed 300 trials of a choice task, which involved selecting between two gambles that were shown on a computer screen. Each gamble in every trial had two possible outcomes, each of which could be either a gain or a loss. A gamble with only gains is called a 'G' gamble, a gamble with only losses is called an 'L' gamble, and a gamble with a gain and a loss is called an 'M', or 'Mixed' gamble. KMD17 considered five types of trials, called 'contexts': 'GG', two G gambles; 'LL', two L gambles; 'MM', both gambles have a gain and a loss outcome; 'MG', a mixed gamble paired with a G gamble, and; 'ML', a mixed gamble paired with an L gamble.

We focus on two hypotheses from KMD17 regarding the MG and ML contexts. First, they hypothesized that older adults were more likely than younger adults to be risk seeking in both contexts. Second, they hypothesized that the *reflection effect* Kahneman &

Tversky (1979), i.e., risk-seeking in the ML context and risk averse in the MG context, would be more pronounced for the younger adults than the older adults.

To test these hypotheses, KMD17 analyzed the choice data of all 60 participants at the individual level. They used a multinomial distribution to model choice responses generated from the three possible risk attitudes, and estimated this model separately for each context. They then used model selection, via normalized maximum likelihood, to classify each individual according to their risk attitude profile for each context. Once the best performing model was identified for each trial type, the resulting participant classifications were treated as deterministic in subsequent statistical tests of their hypotheses. The results of their significant G^2 tests supported their hypotheses, but they were based on the incorrect assumption that every individual was classified correctly and with certainty. Therefore, we will reanalyze their data using our Bayes factor method, taking full account of classification uncertainty.

The treatment effects that we seek to test can be formalized as differences between the distributions of risk-attitude classifications among younger adults and older adults. The first hypothesis is that the proportion of older adults whose profiles are risk-seeking in both contexts is higher than that of younger adults. The second hypothesis is that the proportion of younger adults whose profiles are risk-seeking in the ML context and risk-averse in the MG context is higher than that of older adults. Therefore, we define six classifications corresponding to different combinations of risk-profiles in the two conditions:

- Classification A: risk seeking for MG and ML conditions
- Classification B: risk averse for MG and ML conditions
- Classification C: risk neutral for MG and ML conditions
- Classification D: risk averse for MG, risk seeking for ML (reflection effect)
- Classification E: risk neutral for MG, risk seeking for ML
- Classification F: risk averse for MG, risk neutral for ML.

To obtain the Bayesian posterior probability of each classification for each participant, we use the methodology of Davis-Stober & Brown (2013). For every participant, we calculate the Bayes factor for each of these six classifications against a baseline classification, denoted G, which places no restrictions on choice behavior. We then derive the Bayesian posterior probability of each of these seven classifications from the Bayes factors as previously described. Posterior estimates of the group-level models, i.e., the alpha parameters for each model A-G, follow directly from these Bayesian posterior probabilities. These are reported in Panel A of Table 4. Using these estimates, we find the Bayes factor *against* a treatment effect to be BF=0.157, which is an evidence ratio of approximately 6:1 in favor of a treatment effect. This means that the encompassing model, which allows younger and older adults to have differing distributions over the classifications A-G, is six times more likely to have generated the data than the null model, which requires younger and older adults to have the same classification mixture.

The previous analysis was a more 'global' test of the classification rates between younger and older adults than what was hypothesized. The research hypotheses say nothing about the proportions of Classifications B, C, E, F and G among these two groups. We can test more specifically for the two effects reported by KMD17 by considering submodels in which we aggregate Classifications B, C, E, F and G; leaving Classifications A and D. In this case, the null model assumes that older and younger adults have identical incidence rates of both risk seeking (Classification A) and the reflection effect (Classification D). It makes no additional assumptions about the incidence rates of the remaining models. On the other hand, the encompassing model assumes different incidence rates of both risk seeking and the reflection effect, while making no additional assumptions about the rates of the remaining classifications.

Panel B of Table 4 reports the resulting α parameters and Bayes factor against a treatment effect. The Bayes factor turns out to be 0.026, which is an evidence ratio of approximately 40:1 in favor of a treatment effect. Therefore, consistent with the conclusions of KMD17, we find it to be highly unlikely that the older and younger adults have identical incidence rates of both risk seeking (Model A) and the reflection effect (model D). It is 40 times more likely that the incidence rates of both risk seeking and the reflection effect are different between the groups.

It is notable that the strength of evidence for a treatment effect is much higher in the submodel than in the global test. This suggests that the global encompassing model may be unnecessarily complex. It has a free parameter for each of the seven classifications, essentially assuming that there are group differences on the incidence rates of each and every combination of risk profiles. Such a highly parameterized model is unnecessary, and may result in overfitting if some of these assumed differences do not actually exist. The submodel that has only three parameters, one for classification A, one for D, and another for the remaining classifications, seems to strike a better balance between complexity and parsimony in describing these data. It allows for the incidence rates of Classifications A and D to be different between the groups, without making strong assumptions about the rates of the remaining classifications. Thus, our analysis thus confirms the findings of KMD17 while also providing a more precise and nuanced evaluation of the strength of evidence in favor of these two effects, without possible artifacts resulting from dichotomizing probabilities.

Analysis of the method

Type-1 and type-2 error rates

Our analysis of the Toy Example suggests that our Bayesian test may be able to pick up treatment effects that would be missed under standard approaches that use dichotomized probabilities. To study this property more systematically, we next conduct Monte Carlo simulations to analyze the type-1 and type-2 error rates of our approach relative to Fisher's exact test with dichotomized classifications, under one possible set of experimental conditions. Although type-1 and type-2 error rates are frequentist concepts, we can explore them for our Bayesian method by simulating experimental data (i.e., classification probabilities) from known, underlying distributions, and then assessing whether the correct inference would be made using our method.

First, we instantiate the hypothesis that there is *no* treatment effect by generating classification probabilities from the same distribution in both conditions. For simplicity, we will assume that there are two models under consideration. Specifically, we use a beta distribution with parameters $\alpha = 4$ and $\beta = 6$, which has a mean of 0.4 and a standard deviation of 0.15.

We ran separate simulations of 10,000 trials each with n=10, n=30, and n=100 participants in each condition. On each simulated data set, we compute the Bayes factor against a treatment effect via our method, as well as the p-value of Fisher's exact test, which requires the probabilities to be dichotomized. We assess the type-1 error rate of Fisher's exact test using significance levels of 0.10, 0.05, and 0.01, respectively. The results are shown in Panel A of Table 5. For the Bayesian test, we consider the result to be a type-1 error if the evidence for a treatment effect is sufficiently strong. We use cutoffs of 1.0, 3.0, and 10.0 to represent slight, substantial, and strong evidence of a treatment effect, respectively (Jeffreys, 1961). The Bayesian type-1 error rates using these cutoffs are shown in Panel B of Table 5.

Fisher's exact test is known to be conservative, and the results of our simulation show that it remains conservative even with the dichotomized dependent variable. The actual type-1 error rate increases with sample size, but remains below the nominal rate at each of the sample sizes we tested. On the other hand, the type-1 error rate for the Bayesian test *decreases* with sample size. Moreover, for sample sizes of 30 or more, the Bayesian test did not make a single type-1 error in any of the 10,000 trials, even at the extremely aggressive cutoff of BF<1.0. Thus, although Fisher's exact test is conservative, the Bayesian test is virtually immune to type-1 errors when the sample size is sufficiently large.

The Bayesian test has an additional advantage over Fisher's exact test in being able to infer that there is no treatment effect. This does not mean simply failing to reject the null hypothesis, but rather accumulating sufficient evidence to conclude that there is no treatment effect. Retaining BF_{e0} as a measuring stick, we used cutoffs of 1.0, 0.3, and 0.1, which can be interpreted as slight, substantial, and strong evidence against a treatment effect, respectively. Panel C of Table 5 summarizes the rate at which the Bayesian test was able to correctly infer that there was no treatment effect. Although the test never reaches a Bayes factor less then 0.1 at these sample sizes, it reaches a Bayes factor less than 0.3 in every simulation with n=100, 84% of the time with n=30, and 14% of the time with n=10.

Next, we move to the type-2 error rate and simulate data with a treatment effect by generating classification probabilities from different beta distributions in the two conditions. We generate classification probabilities using Beta(2,18) for Condition 1 and Beta(8,12) for Condition 2. We conduct 10,000 simulation trials each with n=20, n=30, and n=40 participants in each condition.

The type-2 error rate of Fisher's exact test on these simulated trials, using cutoffs of 0.01, 0.05, and 0.10, respectively, are shown[6] in Panel A of Table 6. We find that Fisher's Exact test is alarmingly underpowered at all three sample sizes. At the .05 significance level, even with n=100 in both conditions, it resulted in type-2 error 26% of the time.

Panel B of Table 6 shows the type-2 error rate of the Bayesian test at each sample size, using cutoffs of 1.0, 3.0, and 10.0 on BF_{e0}, respectively. In contrast to the results from Fisher's Exact Test, we find that the Bayesian test is able to correctly infer a treatment effect in virtually every trial with n=100 in each condition. Even with only n=30 in each condition, and using a cutoff of BF<0.3 for inferring an effect, the type-2 error rate is just 2.0%. We attribute this lack of power to the fact that the classification probabilities had to

[6] Given the continuous nature of the dependent variable in this example (i.e., the classification probability) one might argue that a linear regression analysis would be more appropriate than Fisher's exact test. Indeed, in this particular case, for simplicity, we considered only two possible classifications, and we generated classification probabilities from homoskedastic beta distributions that are not far from normal. However, it would be straightforward to generate an example that could "fool" the regression analysis by generating data from bimodal or strongly skewed distributions, with different variances. Furthermore, the linear regression analysis does not generalize to cases with more than two possible classifications, as in Example 1.

be dichotomized before conducting the test.

The deterministic special case

An important special case of probabilistic classification is when the probabilities are degenerate (i.e., equal to either zero or one). In that case, classical tests for analyzing contingency tables, like Fisher's Exact, can be applied without dichotomizing. If our method works as advertised, then it should agree with the results of Fisher's exact test in such cases, as the latter is a well-established method for detecting a change in the distribution of (deterministic) classifications across conditions.

For example, consider a hypothetical experiment aimed at testing whether an experimental intervention affects participants' preference between two snacks options, one that is healthy and another that is decadent. Suppose that each of 60 participants was randomly assigned to either the treatment condition, in which they would receive the intervention, or the control condition, in which they would not receive the intervention. The dependent variable is the choice of snack at the end of the study period. Hypothetical results are summarized in the two-by-two contingency table in Panel A of Table 7. The table indicates that 19 of 30 participants in the control condition chose the decadent snack, compared to just 10 of 30 in the intervention condition. This suggests that the intervention may reduce preferences for the decadent snack. That finding is confirmed to be significant by Fisher's Exact test (p=0.0379), as well as two other classical test that are also applicable in this case: a z-test for the difference between population proportions (p=0.0201) and a chi-squared test of homogeneity (p=0.0148).

The same data could also be analyzed using our Bayes factor method. We simply treat the classification probabilities as degenerate (i.e., p=0.0 or p=1.0). Let c_1 represent choice of the healthy snack and c_2 represent choice of the decadent snack. Then, we can compute the posterior Dirichlet parameter for each classification (i.e., snack choice) in each condition by adding the classification probabilities in each condition as follows:

$$\alpha'_{1,1} = 1.0 + (11 \times 1.0) + (19 \times 0.0) = 12.0$$

$$\alpha'_{2,1} = 1.0 + (11 \times 0.0) + (19 \times 1.0) = 20.0$$

$$\alpha'_{1,2} = 1.0 + (20 \times 1.0) + (10 \times 0.0) = 21.0$$

$$\alpha'_{2,2} = 1.0 + (20 \times 0.0) + (10 \times 1.0) = 11.0$$

where $\alpha'_{m,k}$ is the parameter for c_m in condition k. These parameters specify the Dirichlet density function that can be plugged into Equation (17) to obtain the Bayes factor *against* a treatment effect. Using Monte Carlo integration to evaluate the integral, we find that the Bayes factor is $BF_{0e} = 0.23$ (see Excel spreadsheet "Example 2" in the online supplement for details). This number quantifies the evidence *against* a treatment effect. It means that, based on these data, it is $\frac{1}{0.23} = 4.35$ times more likely than not that the population rate of choosing the decadent snack is different in the intervention condition than in the control condition. Put differently, the evidence ratio is approximately 4:1 *in favor* of a treatment effect. Thus, the Bayes factor result confirms the result of the classical hypothesis tests while also offering an interpretation in terms of the odds ratio between the two hypotheses.

On the other hand, consider a different set of hypothetical data from the same experiment. Suppose that the proportion of participants choosing the healthy option is 10 out of 30 for *both* conditions, suggesting that there is no treatment effect. Indeed, all three classical hypothesis tests: z-test, Chi Squared, and Fisher's exact test would yield p-values of exactly 1.000, meaning that the null hypothesis cannot be rejected.[7]. The Bayes factor result, on the other hand, directly yields the odds ratio for the null hypothesis. The Bayes factor for this example is equal $BF_{0e} = 3.39$, meaning that the null hypothesis is 3.39 times more likely than the alternative. In other words, the odds ratio in favor of a treatment effect is approximately 1:3.

[7] Under specific assumptions about the true population proportions, one could do a power analysis and estimate the probability that the result is type-II error. In the Bayesian analysis, analogous assumptions

The example above represents the extreme case in which classification probabilities are degenerate (i.e., equal to either zero or one) and the classification mixtures are identical across conditions. This pattern of results yields the strongest possible evidence against a treatment effect. While frequentist tests will always yield a p-value of 1.0 in this case, the Bayes factor depends on the sample size and the specific (common) classification mixture. To illustrate, Table 8 shows the Bayes factor for such data under different combination sample size (in each condition) and classification mixture (e.g., the proportion of participants classified as “healthy”). In general, the Bayes factor increases with sample size, and is largest when the common mixture is closer to 1.0 or 0.0. We caution that for some combinations, e.g., $N_k < 20$ and a common mixture between 0.3 and 0.7, it is not possible to exceed the threshold for “substantial” evidence against a treatment effect ($BF_{0e} \geq 3.0$). These are highlighted in red in Table 8. In addition, for sample sizes of 100 or less in each condition, the Bayes factor only exceeds 10.0 when every subject has the same classification (i.e., when the common proportion is either 0.0 or 1.0). We conclude that the burden of proof is high for inferring that there is *not* a treatment effect. In contrast, the Bayes factor can show very strong evidence for a treatment effect, even when the sample size is relatively small. For instance, with just ten subjects in each condition, if classifications are unanimous and opposite across the two conditions, the odds ratio in favor of a treatment effect is more than 32,000:1, and the p-value of Fisher’s exact test is essentially zero.

To further explore the relationship between the Bayes factor and Fisher’s exact test in the deterministic special case, we conducted both tests on a large set of randomly generated, two-by-two contingency tables. To generate each table, we assumed a balanced design with either N=10, N=30, or N=100 participants in each condition, and generated the frequency of each choice (i.e., classification) in each condition uniformly at random. We

about the true population proportions could be encoded in the prior distribution in the alternative hypothesis. However, one would need integrate this prior over the equality constraint to enable calculation of the Bayes factor in Equation 17. A uniform prior assumes that all proportions are equally likely. The frequentist power analysis assumes a specific set of proportions, which could be encoded in the alternative hypothesis as a degenerate prior with all of the weight on one set of proportions.

did this 10,000 times with each sample size, and for each one we computed both the p-value, obtained from Fisher's exact test, and the Bayes factor.

The results of the simulations are plotted in the graphs in Figure 3. Panels A, B, and C of the figure depict the results for N=10, N=30, and N=100, respectively. In each panel, the x-axis is the p-value and the y-axis is the Bayes factor. Each dot represents one simulated data set and is color coded by the sample size. The smallest values of both the p-value and the Bayes factor were on the order of 10^{-54}, so we use a log scaling of both coordinate axes and truncate both axes in each graph at 10^{-7} to maintain perspective in the range of p-values and Bayes factors that are more typical (i.e., near $p = 0.01$ and $BF_{0e} = 1.0$).

The pattern of results is the same for each sample size. As expected, smaller p-values (closer to 0.0) correspond to a smaller Bayes factors (closer to $-\infty$). Most strikingly, there is an exceptionally strong linear relationship between the p-value and the Bayes factor in the log-log coordinates of the axes. This indicates a power relationship, Indeed, we find that a power curve fits each graph with $R^2 > 0.99$.

Panel D overlays the results for all three sample sizes. There, we see clearly that the same p-value corresponds to a lower Bayes factor (stronger evidence of a treatment effect) when the sample size is smaller. This pattern is consistent with the fact that the same p-value for a smaller sample size indicates a larger effect size. For instance, we find that a Bayes factor of 1.00, which indicates equal evidence for and against the null hypothesis, corresponds to a p-value of approximately 0.3 when $n = 10$, 0.2 when $n = 30$, and 0.07 when $n = 100$. A p-value of 0.01 corresponds to a Bayes factor of about $BF_{0e} = 0.03$ when $n = 10$, which is an evidence ratio of approximately 33:1 in favor of a treatment effect. The corresponding evidence ratio at p=0.01 drops to about 100:6 ($BF_{0e} = 0.06$) and 100:12 ($BF_{0e} = 0.12$) when $n = 30$ and $n = 100$, respectively.

In conclusion, although our Bayesian method is not a direct theoretical extension of Fisher's exact test, they bear a strikingly close relationship where both methods apply (i.e., when classifications are deterministic). The fact that these tests agree where they both apply strengthens the notion that our method identifies treatment effects in a reasonable, meaningful, and interpretable way.

Extension to multiple conditions and higher-order designs

The Bayes factor between H_0 and H_e is akin to the omnibus test in the analysis of variance (ANOVA). It tests the extreme hypothesis that all conditions are identical against the alternative that there is at least one condition unlike the others. While this test may be sufficient for simple designs, such as one with two levels of a single factor, more complex designs with crossed factors and multiple factor levels may require more nuanced hypotheses analogous to testing simple main effects in classical ANOVA. Such hypotheses can be formalized and tested in the current framework by defining additional submodels of H_e with the appropriate equality constraints.

For example, consider an experiment with three treatment conditions. Without loss of generality, suppose that Condition 1 is a control condition while Conditions 2 and 3 are two different experimental treatments. Let H_e be the encompassing model in which each condition k has its own classification mixture $\boldsymbol{\phi}_k$. Let H_0 be the null model in which $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 = \boldsymbol{\phi}_3$ (i.e., no treatment effects). These two models, H_0 and H_e, instantiate the extreme cases in which either both treatments have an effect and the effects are different (H_e) or neither treatment has an effect (H_0). It may also be relevant to consider the intermediate cases in which only one treatment has an effect, or in which both treatments have the same effect. This can be done by defining additional hierarchical models with different equality constraints on the classification mixtures. To that end, we define another hierarchical model, H_1, to be identical to H_e but with the restriction that $\boldsymbol{\phi}_2 = \boldsymbol{\phi}_3$. Under H_1, both treatments have an effect, since the classification mixtures in Conditions 1 and 2 may differ from the control condition, but the effects are identical. Let H_2 and H_3 also be defined identically to H_e, except with the restriction that $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_3$, or $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$, respectively (i.e., one treatment has an effect but the other does not).

The null model, H_0, is nested in each of H_1, H_2, H_3, and H_e, so it serves as a natural baseline against which to compute the Bayes factor for each of the other hypotheses. The calculation is done as in Equation (17) except that the product over conditional Dirichlet densities inside the integral has only as many terms as there are unconstrained Dirichlet distributions (i.e., classification mixtures). If multiple treatment conditions are assumed to have the same classification mixture, then the corresponding multinomial distribution is conditioned on the combined data from all of those treatment conditions. For example, H_2 assumes that $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_3$, so the Bayes factor for H_0 over H_2 would be computed as

$$BF_{02} = \Gamma(M)^{-3} \int_{\Omega} p(\boldsymbol{\phi}|\mathbf{x}_1, \mathbf{x}_3) p(\boldsymbol{\phi}|\mathbf{x}_2)\, d\boldsymbol{\phi}.$$

The result can be interpreted as the odds ratio for the hypothesis of no treatment effect relative to the hypothesis that there is an effect in Condition 2 but not in Condition 3. Inverting this Bayes factor yields BF_{20}, which is the strength of evidence *for* an effect in condition 2 but not in condition 3. Put differently, BF_{20} measures the strength of evidence for an effect in Condition 2 under the assumption that there is not effect in Condition 3.

The method extends similarly to two-way designs. Consider a fully crossed design with two factors, F_1 and F_2, and two levels of each factor. Let $\mathbf{x}_{a,b}$ denote the data from all participants at the a^{th} level of F_1 and the b^{th} level of F_2, and let $\boldsymbol{\phi}_{a,b}$ denote the mixture of classifications at the a^{th} level of F_1 and b^{th} level of F_2. In the null model, the mixtures are assumed to be identical in every condition. Formally, $\boldsymbol{\phi}_{a,b} = \boldsymbol{\phi}_{c,d}$ for all a, b, c, d (i.e., there are no treatment effects). In the encompassing model, both factors may have treatment effects. We instantiate the hypothesis that there is a treatment effect of F_1 but not F_2 by adding appropriate restrictions to the encompassing model. Specifically, define H_1 to identical to H_e but with the restrictions that $\boldsymbol{\phi}_{1,1} = \boldsymbol{\phi}_{1,2}$ and $\boldsymbol{\phi}_{2,1} = \boldsymbol{\phi}_{2,2}$. Then H_1 has two free parameters, which we will denote by $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, defined as $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_{1,1} = \boldsymbol{\phi}_{1,2}$ and

$\boldsymbol{\phi}_2 = \boldsymbol{\phi}_{2,1} = \boldsymbol{\phi}_{2,2}$. Since H_0 is nested in H_1, we can compute the Bayes factor for H_0 relative to H_1, using Equation 17, as

$$BF_{01} = \Gamma(M)^{-2} \int_{\Omega} p(\boldsymbol{\phi}_1 = \boldsymbol{\phi} | \mathbf{x}_{1,1}, \mathbf{x}_{1,2}) p(\boldsymbol{\phi}_2 = \boldsymbol{\phi} | \mathbf{x}_{2,1}, \mathbf{x}_{2,2})\, d\boldsymbol{\phi}. \tag{18}$$

In contrast, the Bayes factor for H_0 relative to H_e would be computed as

$$BF_{0e} = \Gamma(M)^{-4} \int_{\Omega} \prod_{a,b} p(\boldsymbol{\phi}_{a,b} = \boldsymbol{\phi} | \mathbf{x}_{a,b})\, d\boldsymbol{\phi}. \tag{19}$$

The Bayes factor for H_1 relative to H_e can then be obtained as $BF_{1e} = \frac{BF_{0e}}{BF_{01}}$. A small value of BF_{1e} (smaller than 1.0) means it is more likely than not that there is a treatment effect of F_1, while a large value indicates evidence against a treatment effect.

Alternatively, we could instantiate the hypothesis that F_2 has a treatment effect but F_1 does not by defining a model H_2 to be identical to H_e except for the restrictions that $\boldsymbol{\phi}_{1,1} = \boldsymbol{\phi}_{2,1}$ and $\boldsymbol{\phi}_{1,2} = \boldsymbol{\phi}_{2,2}$. This model also has two free parameters and the Bayes factors for it relative to H_0 and H_e can be obtained as above.

In this way, our framework can be applied to test any hypothesis that can be formulated as (sets of) equality constraints among the multinomial parameters governing the distribution of classifications in the experimental conditions. The next example illustrates this for a two-way design. Included in the online supplement is a simple example evaluating a hypothesis in a 2 × 2 × 2 design. Also included in the online supplement is Matlab code that can be used to calculate Bayes factors for models defined by arbitrary equality constraints among experimental conditions.

Two-way example

Recall Example 1, in which a hypothetical experiment was aimed at testing whether an experimental manipulation affects the distribution of three classifications in a population. Suppose that this hypothetical experiment had a second factor, also with two levels, which was crossed with the first. For example, the first factor could be the primary factor of interest, such as the intervention and control conditions from the snack experiment, and the second could be a blocking factor such as gender. For convenience, we will refer to the levels of the first factor as Condition 1 and Condition 2, and to the levels of the second factor as Block 1 and Block 2. Suppose that there were ten participants in each combination of the two factors, and that a model selection analysis had yielded the probability of each classification for each participant. A hypothetical set of results from this analysis are given in Panel A of Table 9. The results were constructed to make the pattern of effects obvious, so that there would be clear expectations for the types of effects that should be picked up by the Bayesian test.

To analyze these data under our method, we define four multinomial distributions – one for each combination of factors. They are denoted $\boldsymbol{\phi}_{1,1}$, $\boldsymbol{\phi}_{1,2}$, $\boldsymbol{\phi}_{2,1}$, and $\boldsymbol{\phi}_{2,2}$. Then, we define four hierarchical models corresponding different hypotheses about the relationships between these classification mixtures:

$$H_0 : \boldsymbol{\phi}_{1,1} = \boldsymbol{\phi}_{1,2} = \boldsymbol{\phi}_{2,1} = \boldsymbol{\phi}_{2,2}$$

$$H_1 : \boldsymbol{\phi}_{1,1} = \boldsymbol{\phi}_{1,2} \neq \boldsymbol{\phi}_{2,1} = \boldsymbol{\phi}_{2,2}$$

$$H_3 : \boldsymbol{\phi}_{1,1} = \boldsymbol{\phi}_{2,1} \neq \boldsymbol{\phi}_{1,2} = \boldsymbol{\phi}_{2,2}$$

$$H_e : \boldsymbol{\phi}_{1,1} \neq \boldsymbol{\phi}_{2,1} \neq \boldsymbol{\phi}_{1,2} \neq \boldsymbol{\phi}_{2,2}$$

Essentially, H_0 is that neither the condition nor the block has an effect, H_1 is that the condition has an effect but not the block, H_2 is that the block has an effect but not the condition, and H_e is that both the condition and the block have effects.

The encompassing model, H_e, comprises four distinct multinomial distributions (i.e., classification mixtures) – one for each combination of the two factors, whereas H_1 and H_2 each comprise just two distinct distributions – one for each condition or one for each block. Each of these multinomials has a Dirichlet prior, which is updated based on only the data from the corresponding level, or combination of levels. For example, in H_e, the parameters of the (Dirichlet) posterior for $\boldsymbol{\phi}_{1,1}$ are computed as one-plus-the-sum of the probabilities among the participants in Condition 1 and Block 1. On the other hand, in H_1 there is a common Dirichlet posterior for $\boldsymbol{\phi}_{1,1}$ and $\boldsymbol{\phi}_{1,2}$, for which the parameters (α's) are computed as one-plus-the-sum of the probabilities among all participants in Condition 1, regardless of blocks. These calculations are illustrated in the spreadsheet for Example 4 in the online supplement.

We computed the Bayes factor for each of H_1, H_2, and H_e relative to H_0 using the spreadsheet for Example 4 in the online supplement, and also using the Matlab code provided in the online supplement. The latter can achieve greater precision in the same amount of computing time, while the former is more illustrative of the steps and underlying calculations in the procedure. In the end, we found that the ten thousand simulation trials implemented in the Excel spreadsheet were enough to provide reasonable convergence. The results are shown in Panel B of Table 9.

Overall, we find $BF_{e0} = 130$, indicating conclusive evidence in favor of the hypothesis that there are effects of both Condition and Block. On the other hand $BF_{10} = 0.21$, indicating substantial evidence against the hypothesis that there is an effect of Condition but not of Block. This should come as no surprise because, aggregating across blocks, the classification probabilities are identical between conditions. Similarly, we find $BF_{20} = 0.19$, indicating substantial evidence against the hypothesis that there is an effect of Block but not Condition. This also comes as no surprise because aggregating across conditions, the classification probabilities are identical between blocks.

Discussion

In summary, we presented a novel method for assessing treatment effects (as well as the lack thereof) within the context of probabilistic individual-level model classifications. Treatment effects are operationalized as changes in the distribution of classifications across populations. Our approach is Bayesian in nature, resulting in a Bayes factor that provides an evaluation of evidence either for, or against, the presence of a treatment effect. The test is quite general in that the individual-level models used for classification need not be nested. Nor are Bayes factors strictly required for computation purposes at the individual-level; approximations can be calculated via AIC or BIC values. Once the individual-level model classification values are in hand, the test itself is computationally simple and does not require high-dimensional integration.

Our methodology does not require the classifications themselves to be formulated as Bayesian models. It can be applied with any set of classifications for which AIC, BIC, or Bayes factor values can be calculated. In this way, the test for a treatment effect can be linked with a wide range of individual-level models across many disciplines, e.g., the general linear model and variations thereof, cognitive process models, etc. Our approach naturally treats each individual-level model as a potential data-generating process for each individual. From a practical standpoint, the individual-level models (i.e., classifications) should be defined to instantiate qualitative distinctions that align with the researcher's hypotheses and goals. For example, the models may instantiate qualitatively different decision strategies, mental states, or clinical diagnoses, or they may simply differ in terms of whether a particular behavioral property is satisfied.

Our test for a treatment effect compares a hierarchical model with equality constraints on classification mixtures against another hierarchical model without such equality constraints. More nuanced treatment effects could be formalized and tested by considering hierarchical models with inequality as well as equality constraints. For example, one may wish to consider the hypothesis that an experimental intervention increases or decreases the likelihood of a certain classification. Klugkist et al. (2010) present a Bayesian method for evaluating such hypotheses on contingency tables, under a similar conceptual framework. Our work extends this method by allowing the classifications to be probabilistic, while only considering the special case of equality constrained hypotheses. The advantage of this restriction is that the resulting test is mathematically and computationally simple relative to the general case.

This work follows in the vein of recent Bayesian extensions of frequentist analyses including the Bayesian t-test (Rouder et al., 2009), Bayesian ANOVA (Klugkist et al., 2005; Rouder et al., 2012, 2016), Bayesian evaluation of contingency tables (Klugkist et al., 2010), Bayesian structural equation modeling (Muthén & Asparouhov, 2012), Bayesian mediation analysis (Yuan & MacKinnon, 2009), Bayesian time-series models (de Vries & Morey, 2013), Bayes factor approaches for interval null hypotheses (Morey and Rouder, 2011), and Default Bayesian hypothesis test for correlations (Wetzels & Wagenmakers, 2012). Although our proposed method is not a direct theoretical extension of any particular frequentist test, we find that it has a remarkably close empirical relationship with Fisher's exact test, on data to which both methods apply (i.e., when classifications are deterministic). Specifically, on a sample of 10,000, randomly generated, two-by-two contingency tables, we find a near-perfect power-law relationship ($R^2 > 0.997$) between the Bayes factor, derived from our method, and the frequentist p-value, obtained from Fisher's exact test.

Our method could also be viewed as a variation of hierarchical Bayesian methods for group-level estimation of a single classification, such as a hierarchical drift diffusion model (Wiecki et al., 2013), or hierarchical cumulative prospect theory (Nilsson et al., 2011). In that approach, a single classification is assumed, and distributional assumption are made about its latent parameters (e.g., that they are normally distributed in the population). A set of hyper-parameters (e.g., the means and variances of the normal distributions) governing the group-level distribution of the latent parameters are then updated based the data from every individual in the group. One important distinction in our approach is that we do not make group-level distributional assumptions about the latent parameters within the classifications under consideration. Rather, we consider the prior distributions on latent parameters within a classification to be defining components, which can themselves carry psychological theory (Vanpaemel, 2010). Moreover, the within-group heterogeneity that we consider is categorical rather than incremental. The classifications under consideration may instantiate fundamentally different neural or psychological processes. Future work could aim to merge these two approaches.

Future work could also extended the method to within-participant designs, as well as mixed factor designs with pre-test and post-test components. Currently, we only consider between-participants designs and post-test comparisons. A key consideration would be to incorporate the dependencies across treatment conditions that naturally arise as part of this design. In principle, one could simply consider all possible combinations of classifications across conditions, but this can quickly lead to an intractable number of combinations when there are multiple conditions and several classifications.

References

Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion. *Mathematics in Science and Engineering*, *126*, 27–96.

Aranovich, G. J., Cavagnaro, D. R., Pitt, M. A., Myung, J. I., & Mathews, C. A. (2017). A model-based analysis of decision making under risk in obsessive-compulsive and hoarding disorders. *Journal of Psychiatric Research*, *90*, 126–132.

Bishara, A. J., Pleskac, T. J., Fridberg, D. J., Yechiam, E., Lucas, J., Busemeyer, J. R., … Stout, J. C. (2009). Similar processes despite divergent behavior in two commonly used measures of risky decision making. *Journal of Behavioral Decision Making*, *22*(4), 435–454.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Bröder, A., & Schiffer, S. (2003). Bayesian strategy assessment in multi-attribute decision making. *Journal of Behavioral Decision Making*, *16*, 193–213.

Cavagnaro, D. R., & Davis-Stober, C. P. (2014). Transitive in our preferences, but transitive in different ways: An analysis of choice variability. *Decision*, *1*(2), 102–122.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253.

Davis-Stober, C. P., & Brown, N. (2013). Evaluating decision maker “type” under p-additive utility representations. *Journal of Mathematical Psychology*, *57*(6), 320–328.

de Vries, R. M., & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological methods*, *18*(2), 165–185.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological bulletin*, *53*(2), 134–140.

Glöckner, A. (2009). Investigating intuitive and deliberate processes statistically: The multiple-measure maximum likelihood strategy classification method. *Judgment and Decision Making*, *4*(3), 186–199.

Hou, F., Huang, C.-B., Lesmes, L., Feng, L.-X., Tao, L., Zhou, Y.-F., & Lu, Z.-L. (2010). qcsf in clinical application: efficient characterization and classification of contrast sensitivity functions in amblyopia. *Investigative ophthalmology & visual science*, *51*(10), 5365.

Jeffreys, H. (1961). Theory of probability/by harold jeffreys. *International series of monographs on physics.*.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, 263–291.

Kellen, D., Mata, R., & Davis-Stober, C. P. (2017). Individual classification of strong risk attitudes: An application across lottery types and age groups. *Psychonomic Bulletin & Review*, 1–9.

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*(12), 6367–6379.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: a Bayesian approach. *Psychological methods*, *10*(4), 477–493.

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological methods*, *15*(3), 281.

Lesmes, L. A., Lu, Z.-L., Baek, J., & Albright, T. D. (2010). Bayesian adaptive estimation of the contrast sensitivity function: The quick csf method. *Journal of Vision*, *10*(3), 17.

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*(5), 331–347.

Love, J., Selker, R., Verhagen, J., Marsman, M., Gronau, Q. F., Jamil, T., … others (2015). Software to sharpen your stats. *APS Observer*, *28*(3).

Luce, R. D. (2000). *Utility of gains and losses: Measurement-theoretical and experimental approaches.* Psychology Press.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, *10*(4), 325–337.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods*, *7*(1), 19–40.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological bulletin*, *113*(1), 181–190.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (Vol. 1). Psychology Press.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological methods*, *17*(3), 313-335.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190–204.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.

Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for Cumulative Prospect Theory. *Journal of Mathematical Psychology*, *55*(1), 84–93.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in cognitive sciences*, *6*(10), 421–425.

Rouder, J. N., Engelhardt, C. R., McCabe, S., & Morey, R. D. (2016). Model comparison in ANOVA. *Psychonomic bulletin & review*, *23*(6), 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*(5), 356–374.

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

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic bulletin & review*, *22*(2), 391–407.

Soper, D. S. (2017). *Fisher's exact test calculator for a 2x3 contingency table [software].* `http://www.danielsoper.com/statcalc`.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*(4), 1004–1017.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Streiner, D. L. (2002). Breaking up is hard to do: the heartbreak of dichotomizing continuous data. *The Canadian Journal of Psychiatry*, *47*(3), 262–266.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the savage-dickey density ratio. *Journal of the American Statistical Association*, *90*(430), 614–618.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using akaike weights. *Psychonomic bulletin & review*, *11*(1), 192–196.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive psychology*, *60*(3), 158–189.

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the savage–dickey density ratio. *Computational Statistics & Data Analysis*, *54*(9), 2094–2102.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064.

Wiecki, T. V., Poland, J., & Frank, M. J. (2015). Model-based cognitive neuroscience approaches to computational psychiatry clustering and classification. *Clinical Psychological Science*, *3*(3), 378–399.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics*, *7*, 14.

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013). Heterogeneity of strategy use in the iowa gambling task: a comparison of win-stay/lose-shift and reinforcement learning models. *Psychonomic bulletin & review*, *20*(2), 364–371.

Young, M. E. (2016). The problem with categorical thinking by psychologists. *Behavioural processes*, *123*, 43–53.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological methods*, *14*(4), 301–322.

Table 1
Hypothetical AIC model-selection results from the experiment in Example 1.

	Condition 1					**Condition 2**			
Sub ID	**AIC(A)**	**AIC(B)**	**AIC(C)**	**Preferred**	**Sub ID**	**AIC(A)**	**AIC(B)**	**AIC(C)**	**Preferred**
C1-1	15.86	14.95	**10.30**	C	C2-1	20.87	15.52	**10.14**	C
C1-2	15.73	**10.34**	14.66	B	C2-2	18.00	16.91	**10.10**	C
C1-3	**11.31**	14.67	11.91	A	C2-3	14.70	18.65	**10.23**	C
C1-4	18.42	14.76	**10.23**	C	C2-4	19.29	15.82	**10.13**	C
C1-5	17.88	18.97	**10.06**	C	C2-5	15.18	15.53	**10.30**	C
C1-6	19.79	15.22	**10.17**	C	C2-6	**10.20**	20.40	14.82	A
C1-7	16.01	16.78	**10.17**	C	C2-7	15.64	17.21	**10.18**	C
C1-8	**10.88**	19.05	12.14	A	C2-8	**10.25**	22.38	14.31	A
C1-9	17.77	14.99	**10.22**	C	C2-9	14.61	15.28	**10.38**	C
C1-10	19.66	**10.21**	14.82	B	C2-10	**10.15**	16.47	16.88	A
C1-11	15.06	**10.26**	16.27	B	C2-11	16.03	15.61	**10.23**	C
C1-12	17.69	14.69	**10.25**	C	C2-12	**10.14**	19.43	15.65	A
C1-13	16.74	15.53	**10.20**	C	C2-13	**10.15**	15.44	20.47	A
C1-14	16.73	14.80	**10.27**	C	C2-14	11.79	**11.42**	14.61	B
C1-15	15.22	19.31	**10.17**	C	C2-15	11.77	15.37	**11.31**	C
C1-16	**10.79**	16.08	12.56	A	C2-16	**10.15**	15.28	23.55	A
C1-17	16.76	14.69	**10.28**	C	C2-17	**10.22**	17.93	14.96	A
C1-18	16.58	17.10	**10.14**	C	C2-18	11.83	**11.03**	23.04	B
C1-19	18.01	19.02	**10.06**	C	C2-19	**10.18**	17.50	15.59	A
C1-20	**10.84**	14.67	12.81	A	C2-20	**10.20**	16.96	15.49	A
C1-21	16.22	16.62	**10.17**	C	C2-21	18.11	17.05	**10.10**	C
C1-22	16.66	16.99	**10.14**	C	C2-22	20.05	17.59	**10.06**	C
C1-23	**10.91**	21.02	12.03	A	C2-23	**10.16**	15.13	23.94	A
C1-24	19.23	**10.33**	13.91	B	C2-24	16.43	21.78	**10.09**	C
C1-25	18.31	18.77	**10.06**	C	C2-25	17.49	15.06	**10.22**	C
C1-26	22.59	19.61	**10.02**	C	C2-26	**10.11**	17.96	16.72	A
C1-27	15.93	15.96	**10.22**	C	C2-27	17.00	**10.98**	12.07	B
C1-28	15.57	14.71	**10.34**	C	C2-28	21.82	14.93	**10.18**	C
C1-29	15.64	15.19	**10.29**	C	C2-29	**10.07**	21.30	16.83	A
C1-30	16.00	**10.19**	16.46	B	C2-30	13.66	19.59	**10.37**	C

Table 2
Contingency table of dichotomized classifications in Example 1.

	Model A	**Model B**	**Model C**	**Total**
Condition 1	5	5	20	30
Condition 2	12	3	15	30
Total	17	8	35	60

Fisher's exact test: $p = 0.138$

Table 3

Classification probabilities for Example 1, derived from the AIC values in Table 1, using Equation 7.

Condition 1					**Condition 2**				
Sub ID	**p(A)**	**p(B)**	**p(C)**	**Highest**	**Sub ID**	**p(A)**	**p(B)**	**p(C)**	**Highest**
C1-1	0.053	0.084	0.863	C	C2-1	0.004	0.063	0.932	C
C1-2	0.057	0.846	0.097	B	C2-2	0.018	0.032	0.950	C
C1-3	0.519	0.097	0.384	A	C2-3	0.095	0.013	0.891	C
C1-4	0.015	0.093	0.892	C	C2-4	0.010	0.054	0.936	C
C1-5	0.019	0.011	0.969	C	C2-5	0.075	0.063	0.862	C
C1-6	0.007	0.074	0.919	C	C2-6	0.904	0.006	0.090	A
C1-7	0.050	0.034	0.917	C	C2-7	0.060	0.027	0.913	C
C1-8	0.646	0.011	0.344	A	C2-8	0.882	0.002	0.116	A
C1-9	0.021	0.083	0.897	C	C2-9	0.100	0.072	0.829	C
C1-10	0.008	0.902	0.090	B	C2-10	0.929	0.039	0.032	A
C1-11	0.080	0.877	0.043	B	C2-11	0.049	0.061	0.890	C
C1-12	0.021	0.096	0.883	C	C2-12	0.932	0.009	0.059	A
C1-13	0.034	0.063	0.903	C	C2-13	0.929	0.066	0.005	A
C1-14	0.035	0.091	0.875	C	C2-14	0.409	0.491	0.100	B
C1-15	0.073	0.010	0.917	C	C2-15	0.412	0.068	0.520	C
C1-16	0.674	0.048	0.278	A	C2-16	0.928	0.071	0.001	A
C1-17	0.034	0.096	0.870	C	C2-17	0.897	0.019	0.084	A
C1-18	0.037	0.029	0.934	C	C2-18	0.401	0.598	0.001	B
C1-19	0.018	0.011	0.971	C	C2-19	0.915	0.024	0.061	A
C1-20	0.658	0.097	0.246	A	C2-20	0.905	0.031	0.064	A
C1-21	0.045	0.037	0.919	C	C2-21	0.017	0.029	0.953	C
C1-22	0.036	0.030	0.934	C	C2-22	0.007	0.022	0.971	C
C1-23	0.634	0.004	0.362	A	C2-23	0.922	0.078	0.000	A
C1-24	0.010	0.849	0.141	B	C2-24	0.040	0.003	0.957	C
C1-25	0.016	0.012	0.972	C	C2-25	0.024	0.080	0.897	C
C1-26	0.002	0.008	0.990	C	C2-26	0.947	0.019	0.035	A
C1-27	0.052	0.051	0.898	C	C2-27	0.030	0.614	0.356	B
C1-28	0.062	0.095	0.844	C	C2-28	0.003	0.085	0.912	C
C1-29	0.060	0.075	0.866	C	C2-29	0.964	0.004	0.033	A
C1-30	0.050	0.911	0.040	B	C2-30	0.160	0.008	0.832	C
$\boldsymbol{\alpha'_1}$	5.024	6.721	21.255		$\boldsymbol{\alpha'_2}$	13.966	3.750	15.283	

$BF_{e0} = 3.125$

Table 4

Re-analysis of Kellen et al. (2017)

Panel A: Global analysis showing posterior α' parameters corresponding to each classification (A-G) in each group (Young, Old), and the Bayes factor against a treatment effect.

	A	B	C	D	E	F	G
Young	1.70	5.34	3.99	16.67	1.17	4.73	3.41
Old	10.01	6.59	2.61	7.17	1.13	3.24	6.26

$BF = 0.157$

Panel B: Posterior Dirichlet parameters ($\boldsymbol{\alpha}'$) and Bayes factor against a treatment effect using the submodel with Classifications B, C, E, F, and G collapsed into a single classification, denoted "Other."

	A	D	Other
Young	1.70	16.67	14.63
Old	10.01	7.17	15.83

$BF = 0.026$

Table 5

Type-1 error rates in the simulation study. Percentages represent how often the test met the corresponding decision criterion for inferring a treatment effect.

Panel A: Type 1 Error Rate with Fisher's Exact Test

	Sample Size		
Decision Criterion	**n=10**	**n=30**	**n=100**
$p < .01$	0.1%	0.3%	0.7%
$p < 0.05$	0.7%	2.5%	3.5%
$p < .10$	2.5%	5.4%	7.4%

Panel B: Type-1 Error Rate with Bayesian test

	Sample Size		
Decision Criterion	**n=10**	**n=30**	**n=100**
Strong Evidence ($BF_{e0} > 10$)	0.7%	0.0%	0.0%
Substantial Evidence ($BF_{e0} > 3.0$)	2.5%	0.0%	0.0%
Barely worth mentioning $BF_{e0} > 1.0$)	16.1%	0.0%	0.0%

Panel C: Rate of Correct Inference (Infer No Effect)

	Sample Size		
Decision Criterion	**n=10**	**n=30**	**n=100**
Strong Evidence ($BF_{e0} < 0.10$)	0%	0%	0%
Substantial Evidence ($BF_{e0} < 0.33$)	14%	84%	100%
Barely worth mentioning ($BF_{e0} < 1.0$)	84%	100%	100%

Table 6

Type-2 error rates in the simulation study. Percentages represent how often the test failed *to meet the corresponding decision criterion for inferring a treatment effect.*

Panel A: Type-2 Error Rate with Fisher's Exact Test

		Sample Size	
Decision Criterion	**n=20**	**n=30**	**n=40**
p < .01	95%	84%	58%
p < 0.05	72%	54%	26%
p < .10	72%	35%	13%

Panel B: Type-2 Error Rate with Bayesian Test

		Sample Size	
Decision Criterion	**n=20**	**n=30**	**n=40**
Strong Evidence ($BF_{e0} > 10.0$)	99%	52%	3.0%
Substantial Evidence ($BF_{e0} > 3.0$)	51%	2.0%	0.0%
Barely worth mentioning $BF_{e0} > 1.0$)	0.0%	0.0%	0.0%

Table 7

First set of hypothetical data in the snack experiment, analyzed using both classical and Bayesian tests. The Bayes factor is for the null model (no effect) versus the encompassing model, so a value less than 1.0 indicates stronger evidence for a treatment effect.

Panel A: Contingency Table

	Healthy	**Decadent**	**Total**
Intervention	11	19	30
Control	20	10	30
Total	31	29	60

Panel B: Classical test results

Test	**p-value**
Z	0.0201
Chi Squared	0.0148
Fisher's Exact	0.0379

Panel C: Bayesian test result

$\mathbf{BF_{0e}}$	0.23

Table 8

Bayes factor against a treatment effect (BF_{0e}) for different combinations of sample size (N_k) and classification mixture in each condition.

	Proportion										
N_k	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
10	5.77	3.03	2.41	2.13	2.02	1.97	2.02	2.14	2.41	3.04	5.74
20	10.75	4.25	3.29	2.90	2.72	2.67	2.72	2.89	3.28	4.25	10.78
30	15.71	5.19	3.96	3.48	3.27	3.20	3.28	3.49	3.98	5.20	15.73
40	20.71	5.96	4.55	3.98	3.74	3.68	3.73	3.99	4.56	5.97	20.68
50	25.68	6.71	5.05	4.46	4.15	4.09	4.18	4.43	5.04	6.68	25.54
60	30.43	7.35	5.51	4.84	4.55	4.45	4.53	4.85	5.53	7.33	30.81
70	36.12	7.89	5.99	5.22	4.91	4.79	4.90	5.19	5.95	7.88	35.55
80	40.78	8.46	6.38	5.58	5.21	5.12	5.24	5.55	6.37	8.40	40.86
90	45.66	8.97	6.75	5.92	5.52	5.40	5.51	5.88	6.78	8.93	45.43
100	51.43	9.47	7.08	6.22	5.80	5.71	5.80	6.21	7.12	9.36	50.93

Table 9

Results and analysis in hypothetical Example 4.

Panel A: Classification probabilities

	Condition 1			**Condition 2**		
Block 1	**A**	**B**	**D**	**A**	**B**	**D**
	0.1	0.1	0.8	0.8	0.1	0.1
	0.1	0.1	0.8	0.8	0.1	0.1
	0.1	0.1	0.8	0.8	0.1	0.1
	0.1	0.1	0.8	0.8	0.1	0.1
	0.1	0.1	0.8	0.8	0.1	0.1
	0.1	0.1	0.8	0.8	0.1	0.1
	0.1	0.1	0.8	0.8	0.1	0.1
	0.1	0.1	0.8	0.8	0.1	0.1
	0.1	0.1	0.8	0.8	0.1	0.1
	0.1	0.1	0.8	0.8	0.1	0.1
Block 2	**A**	**B**	**D**	**A**	**B**	**D**
	0.7	0.2	0.1	0.2	0.1	0.7
	0.7	0.2	0.1	0.2	0.1	0.7
	0.7	0.2	0.1	0.2	0.1	0.7
	0.7	0.2	0.1	0.2	0.1	0.7
	0.7	0.2	0.1	0.2	0.1	0.7
	0.7	0.2	0.1	0.2	0.1	0.7
	0.7	0.2	0.1	0.2	0.1	0.7
	0.7	0.2	0.1	0.2	0.1	0.7
	0.7	0.2	0.1	0.2	0.1	0.7
	0.7	0.2	0.1	0.2	0.1	0.7

Panel B: Analysis results

Model	**Interpretation**	**Bayes factor versus H_0**
H_0	No treatment effects	1.00
H_1	Effect of Condition but not block	0.21
H_2	Effect of Block but not treatment	0.19
H_e	Effects of both treatment and block	130

Cavagnaro FINAL fig 1.png

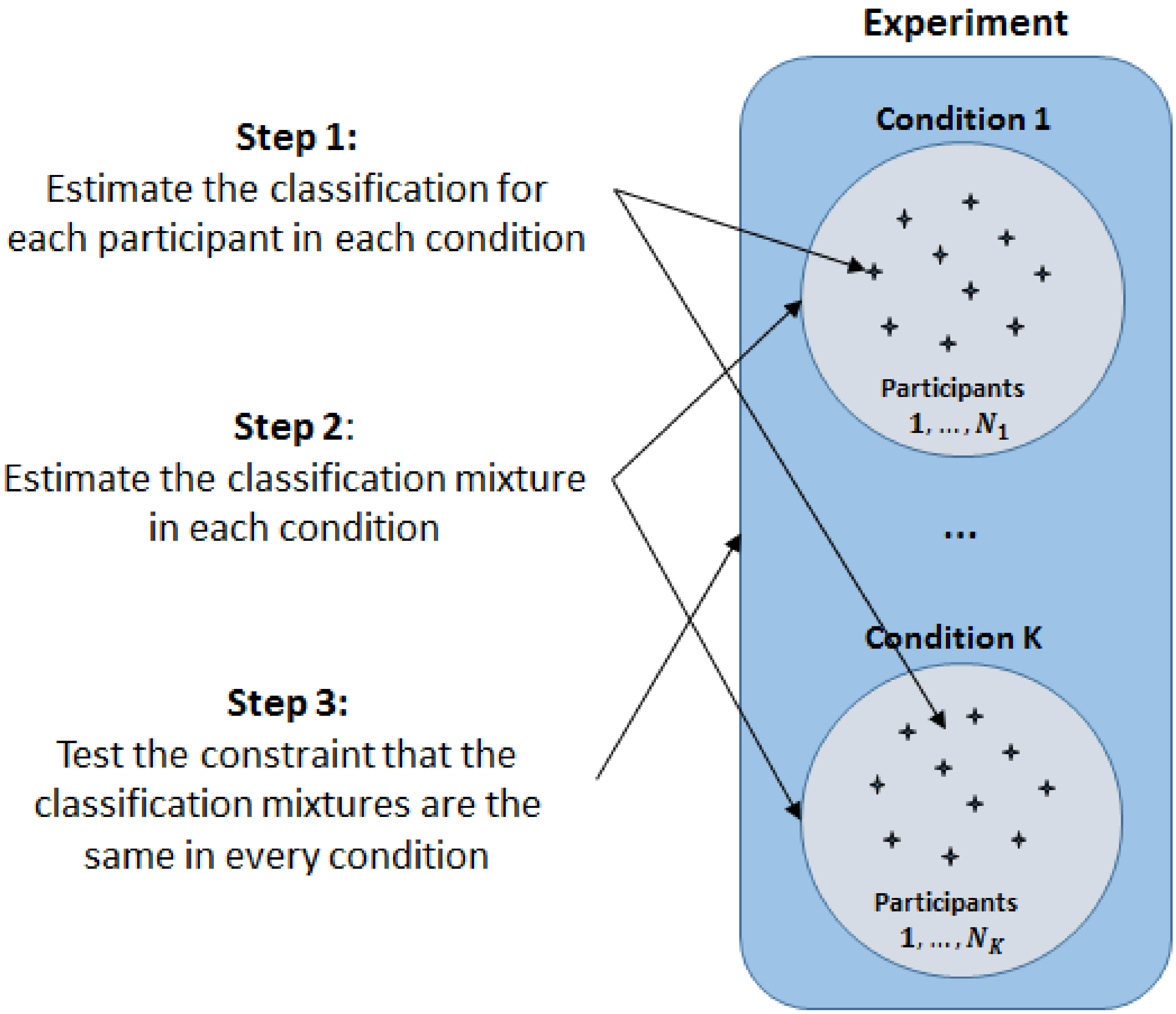
Experiment
Condition 1
Participants
$1, \ldots, N_1$
...
Condition K
Participants
$1, \ldots, N_K$
Step 1:
Estimate the classification for each participant in each condition
Step 2:
Estimate the classification mixture in each condition
Step 3:
Test the constraint that the classification mixtures are the same in every condition

Figure 1. Flowchart of the three steps in the method. The diagram on the right depicts the organization of an experiment consisting of K conditions, with N_k participants assigned to condition k for $k = 1, \ldots, K$. Arrows from the steps on the left to the diagram on the right indicate the level of generality at which each step operates: Step 1 on individual participants, Step 2 on conditions, and Step 3 on the entire experiment.

Cavagnaro FINAL fig 2.png

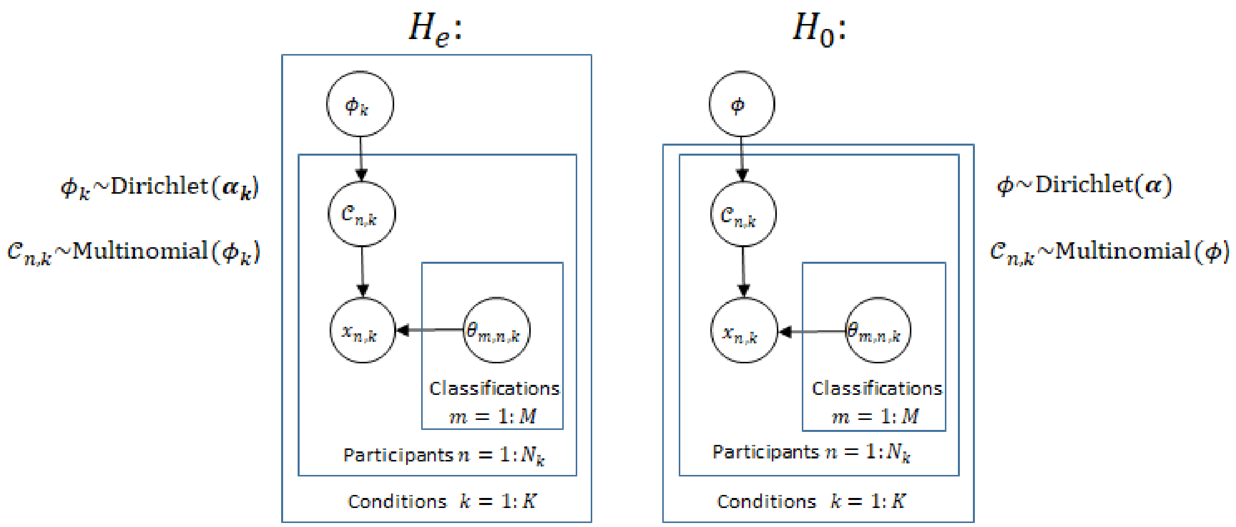
H_e:
H_0:
ϕ_k
$c_{n,k}$
$x_{n,k}$
$\theta_{m,n,k}$
Classifications $m = 1{:}M$
Participants $n = 1{:}N_k$
Conditions $k = 1{:}K$
ϕ_k~Dirichlet(α_k)
$c_{n,k}$~Multinomial(ϕ_k)
ϕ
$c_{n,k}$
$x_{n,k}$
$\theta_{m,n,k}$
Classifications $m = 1{:}M$
Participants $n = 1{:}N_k$
Conditions $k = 1{:}K$
ϕ~Dirichlet(α)
$c_{n,k}$~Multinomial(ϕ)

Figure 2. Graphical representation of the null and encompassing models. The difference between the two is that the encompassing model (H_e, left) has a separate Multinomial parameter for each condition, while the null model (H_0, right) has a single multinomial parameter that is common to every condition. In both models, the distributions of $x_{n,k}$ and $\theta_{m,n,k}$ are governed by likelihood function and prior, respectively, in the classifications supplied by the user for the particular study.

Cavagnaro FINAL fig 3.png

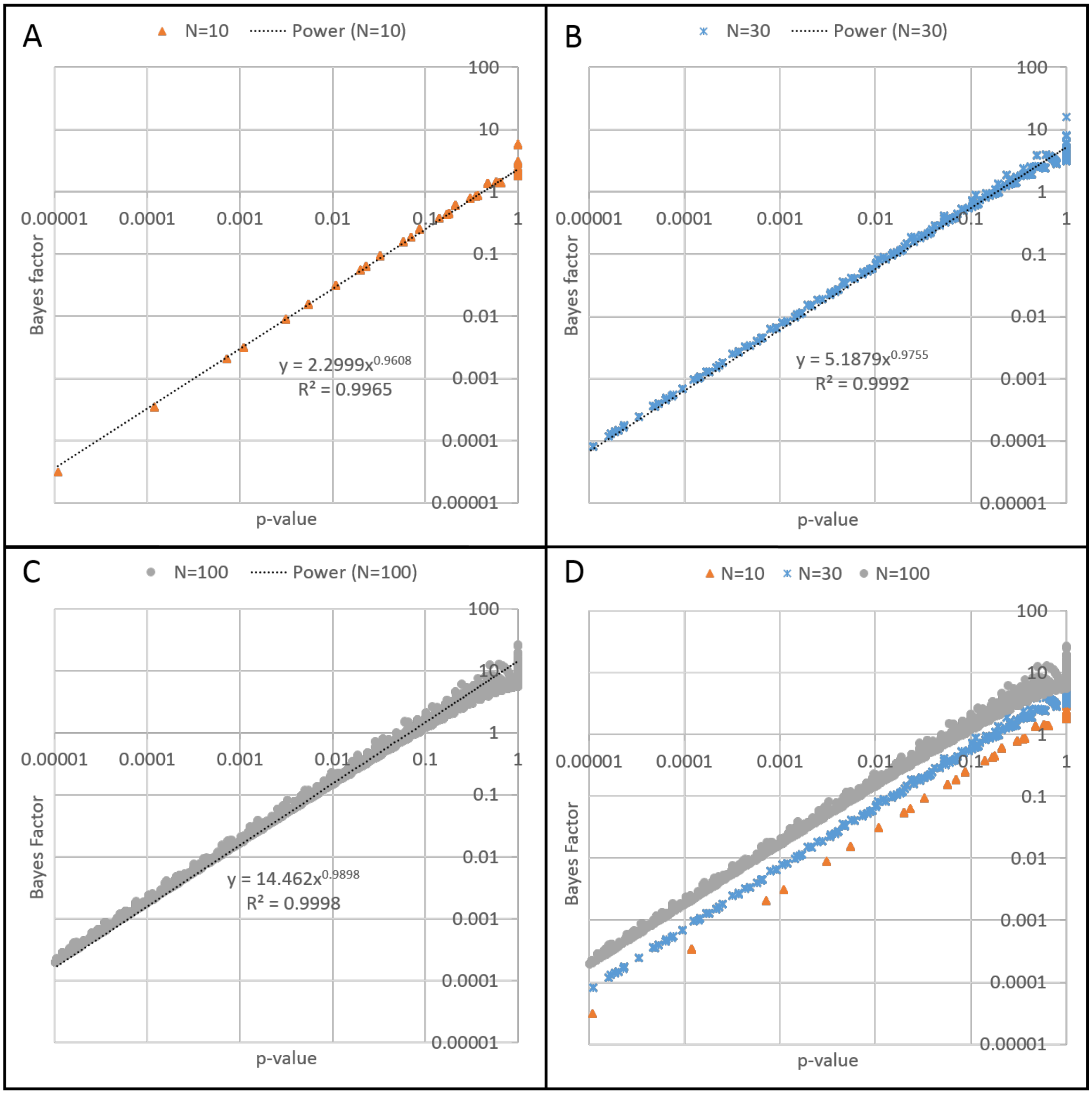

Figure 3. Depiction of the relationship between the Fisher's exact p value and the Bayes factor on 10,000 simulated 2x2 contingency tables, with either n=10 (Panel A), n=30 (Panel B), or n=100 (Panel C) observations in each row. A Bayes factor greater than 1.0 indicates evidence *against* an effect (i.e., in favor of the null hypothesis). Both axes are shown in logarithmic scale to emphasize the near-perfect linear relationship in log-log coordinates, which indicates a power relationship. Both axes are truncated below at 0.00001. Panel D overlays the simulations for all three sample sizes to highlight the fact that the same p-value may correspond to a different Bayes factor depending on the sample size. In general, the same p-value with a smaller sample size indicates *stronger* evidence of an effect.