Informed Bayesian *t*-Tests

Quentin F. Gronau, Alexander Ly & Eric-Jan Wagenmakers

Taylor & Francis
Taylor & Francis Group

Informed Bayesian *t*-Tests

Quentin F. Gronau[a], Alexander Ly[a,b], and Eric-Jan Wagenmakers[a]

[a]Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands; [b]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

ABSTRACT

Across the empirical sciences, few statistical procedures rival the popularity of the frequentist *t*-test. In contrast, the Bayesian versions of the *t*-test have languished in obscurity. In recent years, however, the theoretical and practical advantages of the Bayesian *t*-test have become increasingly apparent and various Bayesian *t*-tests have been proposed, both objective ones (based on general desiderata) and subjective ones (based on expert knowledge). Here, we propose a flexible *t*-prior for standardized effect size that allows computation of the Bayes factor by evaluating a single numerical integral. This specification contains previous objective and subjective *t*-test Bayes factors as special cases. Furthermore, we propose two measures for informed prior distributions that quantify the departure from the objective Bayes factor desiderata of predictive matching and information consistency. We illustrate the use of informed prior distributions based on an expert prior elicitation effort. Supplementary materials for this article are available online.

1. Introduction

The *t*-test is designed to assess whether or not two means differ. The question is fundamental, and consequently the *t*-test has grown to be an inferential workhorse of the empirical sciences. The popularity of the *t*-test is underscored by considering the *p*-values published in eight major psychology journals from 1985 until 2013 (Nuijten et al. 2016); out of a total of 258,105 *p*-values, 26% tested the significance of a *t* statistic. For comparison, 4% of those *p*-values tested an *r* statistic, 4% a *z* statistic, 9% a χ^2 statistic, and 57% an *F* statistic. Similarly, Wetzels et al. (2011) found 855 *t*-tests reported in 252 psychology articles, for an average of about 3.4 *t*-tests per article.

The two-sample *t*-test typically assumes that the data are normally distributed with common standard deviation, that is, $Y_{1i} \sim \mathcal{N}(\mu + \frac{\sigma\delta}{2}, \sigma^2)$ and $Y_{2j} \sim \mathcal{N}(\mu - \frac{\sigma\delta}{2}, \sigma^2)$ for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. The parameter μ is interpreted as a grand mean, σ as the common standard deviation, and δ as the (standardized) effect size. A typical application involves a treatment group and a control group and the task is to infer whether or not the treatment has an effect. The null hypothesis of the treatment not being effective corresponds to $\mathcal{H}_0 : \delta = 0$ and implies that the population means of the two groups are the same, while the two-sided alternative $\mathcal{H}_1$ allows the effect size to vary freely, and implies that the population means of the two groups differ.

This article concerns the Bayesian *t*-test originally developed by Jeffreys (1948) in the one-sample setting, and recently extended to the two-sample set-up by Gönen et al. (2005) and, subsequently, Rouder et al. (2009). In his work on hypothesis testing, Jeffreys focused on the *Bayes factor* (Etz and Wagenmakers 2017; Kass and Raftery 1995; Ly, Verhagen, and Wagenmakers 2016a, 2016b; Robert, Chopin, and Rousseau 2009), the predictive updating factor that quantifies the change in relative beliefs about the hypotheses $\mathcal{H}_1$ and $\mathcal{H}_0$ based on observed data d (Wrinch and Jeffreys 1921, p. 387):

$$\underbrace{\frac{P(\mathcal{H}_1 \mid d)}{P(\mathcal{H}_0 \mid d)}}_{\text{Posterior odds}} = \underbrace{\frac{p(d \mid \mathcal{H}_1)}{p(d \mid \mathcal{H}_0)}}_{\text{BF}_{10}(d)} \underbrace{\frac{P(\mathcal{H}_1)}{P(\mathcal{H}_0)}}_{\text{Prior odds}} . \quad (1)$$

The Bayes factor is given by the ratio of the marginal likelihoods of $\mathcal{H}_1$ and $\mathcal{H}_0$ that are obtained by integrating out the model parameters with respect to the parameters' prior distribution. For the two-sample *t*-test, the null model $\mathcal{H}_0$ specifies two free parameters $\zeta = (\mu, \sigma)$, while the alternative has three, namely, $(\zeta, \delta) = (\mu, \sigma, \delta)$. Once the priors $\pi_0(\zeta)$ and $\pi_1(\zeta, \delta)$ are specified, the parameters of each model can be integrated out as follows

$$\text{BF}_{10}(d) = \frac{\int_\Delta \int_Z f(d \mid \delta, \zeta, \mathcal{H}_1) \, \pi_1(\delta, \zeta) \, d\zeta \, d\delta}{\int_Z f(d \mid \zeta, \mathcal{H}_0) \, \pi_0(\zeta) \, d\zeta}. \quad (2)$$

Equation (2) shows that the Bayes factor can be regarded as the ratio of two weighted averages where the weights correspond to the prior distribution for the parameters. Consequently, the choice of the prior distributions is crucial for the development

CONTACT Quentin F. Gronau ✉ Quentin.F.Gronau@gmail.com Department of Psychological Methods, University of Amsterdam, Nieuwe Achtergracht 129 B, 1018 WT Amsterdam, The Netherlands.

of a Bayes factor hypothesis test. Jeffreys (1961) elaborated on various procedures to select priors for a Bayes factor and the construction of his one-sample t-test became the norm in objective Bayesian analysis (e.g., Bayarri et al. 2012; Berger and Pericchi 2001; Liang et al. 2008). Jeffreys's Bayes factor for the two-sample t-test, however, was needlessly complicated and it was Gönen et al. (2005) who provided the desired simplification.

The innovation of Gönen et al. (2005) was to reparameterize the means of the two groups, μ_1 and μ_2, in terms of a grand mean and the effect size, as was introduced at the start of this section. Following Jeffreys, the second idea was to use a right Haar prior $\pi_0(\mu, \sigma) \propto \sigma^{-1}$ on the nuisance parameters, the parameters common to both the null and the alternative model (Bayarri et al. 2012; Berger, Pericchi, and Varshavsky 1998; Severini, Mukerjee, and Ghosh 2002). Using this prior choice, the marginal likelihood of the null model—the denominator of the Bayes factor $\mathrm{BF}_{10}(d)$—is proportional to the density of a standard t-distribution evaluated at the observed t-value. The third idea was to decompose the prior under the alternative hypothesis into a product of the prior used under the null hypothesis, and a test-relevant prior on the (standardized) effect size, that is, $\pi_1(\mu, \sigma, \delta) = \pi_0(\mu, \sigma)\pi(\delta)$. Finally, Gönen et al. (2005) showed that a normal prior $\delta \sim \mathcal{N}(\mu_\delta, g)$ on the effect size yields a Bayes factor for the two-sample t-test, that is, easily calculated:

$$\mathrm{BF}_{10}(d; \mu_\delta, g) = \frac{\frac{1}{\sqrt{1+n_\delta g}} T_\nu\left(\frac{t}{\sqrt{1+n_\delta g}}; \sqrt{\frac{n_\delta}{1+n_\delta g}}\mu_\delta\right)}{T_\nu(t)}, \quad (3)$$

where $\frac{1}{b} T_\nu\left(\frac{t}{b}; a\right)$ denotes the density of a t-distribution with ν degrees of freedom, noncentrality parameter a and scale b, $T_\nu(t) = T_\nu(t; 0)$ denotes the density of a standard t-distribution, and d refers to the data consisting of degrees of freedom $\nu = n_1 + n_2 - 2$, the observed t-value $t = \sqrt{n_\delta}(\bar{y}_1 - \bar{y}_2)/s_p$, where $n_\delta = (1/n_1 + 1/n_2)^{-1}$ is the effective sample size, and $\nu s_p^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2$ the pooled sums of squares. In fact, the Bayes factors for the two-sample t-test discussed here also cover the one-sample case, by (1) replacing the effective sample size by the sample size n; (2) replacing the degrees of freedom ν by $n-1$; and (3) replacing the two-sample t-value by its one sample equivalent $t = \sqrt{n}\bar{y}/s_y$, where $\nu s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2$. This means that practitioners who can calculate a classical t-test can also easily conduct a Bayesian two-sample t-test: they only need to choose the hyperparameter μ_δ corresponding to the effect size prior mean and the hyperparameter g corresponding to the prior variance. For brevity, we refer to the latter choice $\delta \sim \mathcal{N}(\mu_\delta, g)$ as a g-prior on δ, since it resembles the priors Zellner (1986) proposed in the regression framework. When $\mu_\delta = 0$, the normal g-prior on δ translates to Zellner's g-prior on the mean difference $(\mu_1 - \mu_2) \sim \mathcal{N}(0, g\sigma^2)$.

Later Bayes factors for the two-sample t-test proposed by Rouder et al. (2009) and Wang and Liu (2016) retained the first three ideas: the parameterization in terms of the grand mean and effect size, the use of the right Haar prior on the nuisance parameters $\pi_0(\mu, \sigma) \propto \sigma^{-1}$, and the decomposition $\pi_1(\mu, \sigma, \delta) = \pi_0(\mu, \sigma)\pi(\delta)$, but they differ in the choice of the test relevant prior $\pi(\delta)$. Wang and Liu (2016) noted that the Bayes factors of Gönen et al. (2005) are *information*

inconsistent, which implies that the Bayes factor in favor of the alternative does not go to infinity when the observed t-value increases indefinitely. To make the Bayes factor information consistent, Wang and Liu (2016) instead proposed to assign g a Pearson Type VI/beta prime hyper-prior distribution (see also Maruyama and George 2011, for this proposal in the regression context). Inspired by the developments of Liang et al. (2008) in the regression framework, Rouder et al. (2009) proposed to replace the normal prior on δ by a Cauchy prior $\pi(\delta) = \mathrm{Cauchy}(\delta; 0, \gamma)$, a choice that resembles that of Jeffreys (1948) proposition for the one-sample t-test with prior scale $\gamma = 1$. In their response to Wang and Liu (2016), Gönen et al. (2019) stressed the relevance of a subjective prior specification and noted that the Bayes factors proposed by Rouder et al. (2009) and Wang and Liu (2016) are not flexible enough to incorporate available expert knowledge, since these objective Bayes factors are based on priors that are centered at zero. Here—without taking sides in the discussion between objective and subjective inference—we present a generalized form of the Bayes factor developed by Rouder et al. (2009) that allows the prior specification to be informed by substantive domain knowledge. The remainder of this article is organized as follows: Section 2 presents the proposed Bayes factor and two measures for quantifying the departure from Jeffreys's desiderata of predictive matching and information consistency. Section 3 demonstrates, using a concrete example, how the proposed Bayes factor can be used in practice to incorporate expert knowledge based on a prior elicitation effort. The article ends with concluding comments.

2. Theory

We use the framework of Gönen et al. (2005) and extend the priors proposed by Rouder et al. (2009) to allow for more informed Bayesian t-tests. We exploit the fact that, with $\pi_0(\mu, \sigma) \propto \sigma^{-1}$, the Bayes factor can be written as (a derivation is provided in the online appendix, supplementary materials [Theorems A.1, A.2, and the associated corollaries]):

$$\mathrm{BF}_{10}(d) = \frac{\int T_\nu(t \mid \sqrt{n_\delta}\delta)\pi(\delta)\mathrm{d}\delta}{T_\nu(t)}, \quad (4)$$

where $T_\nu(t \mid a)$ denotes the density of a t-distribution with ν degrees of freedom and noncentrality parameter a. The numerator can be easily evaluated using numerical integration. Consequently, Equation (4) shows that researchers can easily obtain a Bayes factor based on any proper prior for the standardized effect size δ by inserting the prior density of interest for $\pi(\delta)$.

We propose the use of a flexible t-prior for δ, that is, $\pi(\delta) = \frac{1}{\gamma} T_\kappa\left(\frac{\delta - \mu_\delta}{\gamma}\right)$, allowing practitioners to incorporate expert knowledge about standardized effect size by specifying a location hyperparameter μ_δ, a scale hyperparameter γ, and a degrees of freedom hyperparameter κ. The resulting Bayes factor is given by:

$$\mathrm{BF}_{10}(d; \mu_\delta, \gamma, \kappa) = \frac{\int T_\nu(t \mid \sqrt{n_\delta}\delta)\frac{1}{\gamma} T_\kappa\left(\frac{\delta - \mu_\delta}{\gamma}\right)\mathrm{d}\delta}{T_\nu(t)}, \quad (5)$$

where the integral in the numerator can be easily calculated using free software packages such as R (R Core Team 2016).

We believe that the proposed Bayes factor based on a t-prior for effect size has a number of advantages. First, similar to the Bayes factor proposed by Gönen et al. (2005)—which is a special case obtained by taking $\gamma = \sqrt{g}$ and $\kappa \to \infty$– it allows researchers, if desired, to incorporate existing expert knowledge about effect size into the prior specification furthering cumulative scientific learning. Second, this class of priors contains the Cauchy prior of Rouder et al. (2009) as a special case (obtained by setting $\kappa = 1$, $\mu_\delta = 0$). Therefore, using the same expression, researchers can incorporate expert prior knowledge or they can use an objective default prior. Third, this set-up allows researchers to quantify the departure from Jeffreys's predictive matching and information consistency desiderata based on departure measures proposed below. This enables a more formal assessment of differences between objective and subjective prior choices and may benefit the dialog between objective and subjective Bayesians (see, e.g., Wang and Liu 2016; Gönen et al. 2019).

2.1. Two Measures for the Departure From Jeffreys's Desiderata

2.1.1. Predictive Matching
Jeffreys considered two desiderata for prior choice. The first desideratum, *predictive matching*, states that the Bayes factor should be perfectly indifferent (i.e., $\text{BF}_{10}(d) = 1$) in case the data are completely uninformative. Recall that the alternative model has three free parameters; it is therefore natural to require at least three observations before conclusions can be drawn. Consequently, Jeffreys required a Bayes factor of 1 for any dataset of size smaller or equal to 2, thus, for $\nu = 0$. As apparent from Equation (1), this requirement guarantees the posterior model odds to be the same as the prior model odds for completely uninformative datasets. For instance, the dataset $d_{\nu < \min}$ consisting of only one observation in each group $n_1 = n_2 = 1$ automatically has zero sums of squares, that is, $\nu s_p^2 = 0$. If $\bar{y}_1 \neq \bar{y}_2$ the associated t-value would then be unbounded. Let $f(d \mid \delta)$ denote the *reduced* likelihood (i.e., the likelihood with the nuisance parameters integrated out): $f(d \mid \delta) = \int \int f(d \mid \mu, \sigma, \delta) \sigma^{-1} \mathrm{d}\mu \mathrm{d}\sigma$. Using a lemma distilled from the Bateman project (Bateman et al. 1954, 1953; Ly, Marsman, and Wagenmakers 2018), straightforward but tedious computations show that $f(d \mid \delta)$ is proportional to the density of a t-distribution with ν degrees of freedom and noncentrality parameter $\sqrt{n_\delta}\delta$ (see Theorem A.2 in the online appendix, supplementary materials for details). To convey that nothing is learned from the dataset $d_{\nu < \min}$, Jeffreys chose $\pi(\delta)$ such that

$$p(d_{\nu < \min} \mid \mathcal{H}_0) = p(d_{\nu < \min} \mid \mathcal{H}_1) = \int f(d_{\nu < \min} \mid \delta) \pi(\delta) \mathrm{d}\delta. \quad (6)$$

As $\nu s_p^2 = 0$, $n_\delta = 1/2$, and $\bar{y}_1 \neq \bar{y}_2$, we obtain

$$(2|\bar{y}_1 - \bar{y}_2|)^{-1} = \int (2|\bar{y}_1 - \bar{y}_2|)^{-1}$$
$$\times \left[1 + \text{sign}(\bar{y}_1 - \bar{y}_2)\text{Erf}(\tfrac{\delta}{2})\right]\pi(\delta)\mathrm{d}\delta, \quad (7)$$

where $\text{sign}(z)$ is one when z is positive, minus one when z is negative, and zero otherwise (see Corollaries A.1.3 and A.2.1 in the online appendix, supplementary materials). $\text{Erf}(z) =$

$\frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} \mathrm{d}u$ is the error function, an odd function of z. Note that the requirement Equation (7) is fulfilled if a proper symmetric prior is used for δ. Based on Equation (7) we define the (two-sided) departure of any proper prior with respect to Jeffreys's predictive matching criterion as

$$D(\pi, \text{Pred} \mid d_{\nu < \min}) = \int \text{sign}(\bar{y}_1 - \bar{y}_2)\text{Erf}(\tfrac{\delta}{2})\pi(\delta)\mathrm{d}\delta, \quad (8)$$

and note that $\text{BF}_{10}(d_{\nu < \min}) = 1 + D(\pi, \text{Pred} \mid d_{\nu < \min})$. For instance, a t-prior located at $\mu_\delta = 0.350$, with scale $\gamma = 0.102$ and $\kappa = 3$ degrees of freedom, as used later on in the example, has a departure of the predictive matching criterion of 0.0198 when $\bar{y}_1 > \bar{y}_2$. In other words, for completely uninformative datasets with $\bar{y}_1 < \bar{y}_2$ the Bayes factor will be $\text{BF}_{10}(d_{\nu < \min}) \approx 0.98$, while if $\bar{y}_1 > \bar{y}_2$ the Bayes factor would be $\text{BF}_{10}(d_{\nu < \min}) \approx 1.02$, instead.

2.1.2. Information Consistency
The second desideratum, *information consistency*, states that the Bayes factor should provide infinite support for the alternative in case the data are overwhelmingly informative (Bayarri et al. 2012; Jeffreys 1942). An overwhelmingly informative dataset for the two-sample t-test is denoted by $d_{\text{info},\nu}$ with $\nu \geq 1$, effective sample size $n_\delta > 1/2$ (this condition implies that there is at least one observation per group), a (pooled) sums of squares $\nu s_p^2 = 0$, and an observed mean difference $\bar{y}_1 - \bar{y}_2 \neq 0$, thus, an unbounded t-value. For such an overwhelmingly informative dataset $d_{\text{info},\nu}$ to provide infinite support for the alternative, Jeffreys required that $p(d_{\text{info},\nu} \mid \mathcal{H}_0)$ is bounded and that $\pi(\delta)$ is chosen such that $\int f(d_{\text{info},\nu} \mid \delta)\pi(\delta)\mathrm{d}\delta$ diverges. With $\nu s_p^2 = 0$ and $\bar{y}_1 \neq \bar{y}_2$ the marginal likelihood of the null model becomes

$$p(d_{\text{info},\nu} \mid \mathcal{H}_0) = \frac{\Gamma(\frac{\nu+1}{2})}{2\pi^{\frac{\nu+1}{2}}\sqrt{\nu+2}}\left(n_\delta(\bar{y}_1 - \bar{y}_2)^2\right)^{-\frac{\nu+1}{2}}, \quad (9)$$

which is indeed bounded (see Corollary A.1.3 in the online appendix, supplementary materials). In Corollary A.2.2 of the online appendix (supplementary materials) it is shown that for δ large, the reduced likelihood $f(d_{\text{info},\nu} \mid \delta)$ with $\nu s_p^2 = 0$ behaves like a polynomial with leading order ν, that is,

$$f(d_{\text{info},\nu} \mid \delta) \sim \delta^\nu. \quad (10)$$

To guarantee for degrees of freedom ν that $\int f(d_{\text{info},\nu} \mid \delta)\pi(\delta)\mathrm{d}\delta$ diverges, it suffices to take a prior that does not have the νth moment. As information consistency should hold for all $\nu \geq 1$, this implies that $\pi(\delta)$ should be chosen such that it does not have a first moment. Based on the condition that the marginal likelihood should already diverge for $\nu = 1$, we define the departure of Jeffreys's information consistency criterion as

$$D(\pi, \text{InfoConsist})$$
$$= \arg\min\left\{\nu \in \mathbb{N} : \int f(d_{\text{info},\nu} \mid \delta)\pi(\delta)\mathrm{d}\delta \notin \mathbb{R}\right\} - 1. \quad (11)$$

If $\pi(\delta)$ is taken to be a t-prior with κ degrees of freedom the departure from Jeffreys's information consistency criterion is $\kappa - 1$, since a t-distribution has $\kappa - 1$ moments. For instance, a t-prior with $\kappa = 3$ degrees of freedom has only two moments

and, therefore, misses the information consistency by two samples. This means that the Bayes factor only goes to infinity for overwhelmingly informative data when $\nu \geq 3$. Therefore, an informed t-prior with degrees of freedom larger than one requires more observations to be "convinced" by the data than does an objective prior with degrees of freedom equal to 1.

2.1.3. Practical Value of the Proposed Departure Measures

The departure measures introduced above can be used to issue recommendations for researchers who would like to incorporate expert knowledge into the prior specification, but would also like to retain Jeffreys's desiderata as much as possible. For the proposed t-prior, we recommend that researchers who would like to retain information consistency choose $\kappa \in (0, 1]$. For instance, setting $\kappa = 1$ results in a Cauchy prior. Note that, crucially, information consistency still holds if this Cauchy prior is centered on a value other than zero which enables one to incorporate expert knowledge about effect size by shifting the prior away from zero. Researchers who want to retain predictive matching should specify the prior to be centered on zero (i.e., $\mu_\delta = 0$); however, the scale parameter γ and the degrees of freedom κ can be chosen freely. Next, we demonstrate with an example how the proposed Bayes factor can be used in practice. The example features a prior elicitation effort (e.g., Kadane and Wolfson 1998) highlighting the practical feasibility of specifying an informed prior based on expert knowledge.

3. Practice

The *facial feedback hypothesis* states that affective responses can be influenced by one's facial expression even when that facial expression is not the result of an emotional experience. In a seminal study, Strack, Martin, and Stepper (1988) found that participants who held a pen between their teeth (inducing a facial expression similar to a smile) rated cartoons as more funny on a 10-point Likert scale ranging from 0 to 9 than participants who held a pen with their lips (inducing a facial expression similar to a pout).

In a recently published Registered Replication Report (Wagenmakers et al. 2016), 17 labs worldwide attempted to replicate this finding using a preregistered and independently vetted protocol. A classical random-effects meta-analysis yielded an estimate of the mean difference between the "smile" and "pout" condition equal to 0.03 [95% CI: −0.11, 0.16]. Furthermore, one-sided default Bayesian unpaired t-tests (using a zero-centered Cauchy prior with scale $1/\sqrt{2}$ for effect size, the current standard in the field of psychology; see Morey and Rouder 2015) revealed that for all 17 studies, the Bayes factor indicated evidence in favor of the null hypothesis and for 13 out of the 17 studies, the Bayes factor in favor of the null was larger than 3. Overall, the authors concluded that "the results were inconsistent with the original result" (Wagenmakers et al. 2016, p. 924).

Here, we present an informed reanalysis of the data of one of the labs based on a prior elicitation effort with Dr. Suzanne Oosterwijk, a social psychologist at the University of Amsterdam with considerable expertise in this domain. The results for the other labs can be found in online appendix C (supplementary materials).

3.1. Prior Elicitation

Before commencing the elicitation process, we asked our expert to ignore the knowledge about the failed replication of Strack, Martin, and Stepper (1988). Next, we stressed that the goal of the elicitation effort was to obtain an informed prior distribution for δ *under the alternative hypothesis* $\mathcal{H}_1$, that is, under the assumption that the effect is present. This was important to prevent unwittingly eliciting a prior, that is, a mixture between a point mass at zero and the distribution of interest. Then, we proceeded in steps of increasing sophistication. First, together with the expert we decided that the theory specified a direction, implying a one-sided hypothesis test. Next, we asked the expert to provide a value for the median of the effect size: this yielded a value of 0.35. Subsequently, we asked for values for the 33% and 66% percentile of the prior distribution for the effect size: this yielded values of 33%-tile = 0.25 and 66%-tile = 0.45. To finesse and validate the specified prior distribution we used the MATCH Uncertainty Elicitation Tool (*http://optics.eee.nottingham.ac.uk/ match/uncertainty.php*; see also online appendix B, supplementary materials), a web application that allows one to elicit probability distributions from experts (Morris, Oakley, and Crowe 2014). Furthermore, we used R's (R Core Team 2016) plotting capabilities for eliciting the prior number of degrees of freedom. The complete elicitation effort took approximately one hour and resulted in a t-distribution with location 0.350, scale 0.102, and 3 degrees of freedom. As shown in the theory part, this prior choice has a departure from the predictive matching criterion of ± 0.0198 and misses information consistency by two samples. It should be emphasized, however, that the goal of this prior elicitation was to construct a prior that truly reflects the expert's knowledge without being constrained by considerations about Bayes factor desiderata. Alternatively, in an elicitation effort that puts more emphasis on these desiderata, one could, for instance, fix the degrees of freedom to one and let the expert only choose the location and scale.

3.2. Reanalysis of the Oosterwijk Replication Study

Having elicited an informed prior distribution for δ under the alternative hypothesis, we now turn to a detailed reanalysis of the facial feedback replication attempt from Dr. Oosterwijk's lab at the University of Amsterdam. This dataset features 53 participants in the "smile" condition with an average funniness rating of 4.63 (SD = 1.48), and 57 participants in the "pout" condition with an average funniness rating of 4.87 (SD =1.32); consequently, the observed t-statistic is $t(108) = -0.90$.

The alternative hypothesis is directional, that is, the teeth condition is predicted to result in relatively high funniness ratings, not relatively low funniness ratings. To respect the directional nature of the alternative hypothesis the two-sided informed t-test outlined above requires an adjustment. Specifically, the Bayes factor that compares an alternative hypothesis that only allows for positive effect size values to the null hypothesis can be computed via a simply identity that exploits the

transitive nature of the Bayes factor (Morey and Wagenmakers 2014):

$$\mathrm{BF}_{+0}(d) = \underbrace{\frac{p(d \mid \mathcal{H}_+)}{p(d \mid \mathcal{H}_1)}}_{\mathrm{BF}_{+1}(d)} \underbrace{\frac{p(d \mid \mathcal{H}_1)}{p(d \mid \mathcal{H}_0)}}_{\mathrm{BF}_{10}(d)} = \mathrm{BF}_{+1}(d)\mathrm{BF}_{10}(d). \quad (12)$$

We already showed how to obtain $\mathrm{BF}_{10}(d)$, that is, the Bayes factor for the two-sided test of an informed alternative hypothesis; the correction term $\mathrm{BF}_{+1}(d)$ can be obtained by simply dividing the posterior mass for δ larger than zero by the prior mass for δ larger than zero. The expression for the marginal posterior distribution for δ is provided in Corollary A.2.3 in the online appendix (supplementary materials). Using this expression, numerical integration can be used to obtain the desired posterior mass. The Bayes factor hypothesis test that we report will respect the directional nature of the facial feedback hypothesis and include the correction term from Equation (12).

Figure 1 shows the results of the reanalysis of the data from the Oosterwijk lab. The displayed prior and posterior distribution do not impose the directional constraint. The one-sided Bayes factor based on the informed prior equals $\mathrm{BF}_{0+}(d; 0.350, 0.102, 3) = 11.5$, indicating that the data are about twelve times more likely under the null hypothesis than under the one-sided alternative hypothesis.

For comparison, Figure 2 displays the results based on the default one-sided zero-centered Cauchy distribution with scale $1/\sqrt{2}$. The one-sided default Bayes factor equals $\mathrm{BF}_{0+}(d; 0, 1/\sqrt{2}, 1) = 8.7$, indicating that the data are about 9 times more likely under the null hypothesis than under the one-sided default alternative hypothesis. Hence, both the informed and the default Bayes factor yield the same qualitative conclusion, that is, evidence for the null hypothesis. However, the unrestricted posterior distributions differ noticeably between the informed and the default analysis: the posterior median based on the informed prior specification is positive and equal to 0.153 (95% credible interval: [−0.264, 0.390]) whereas
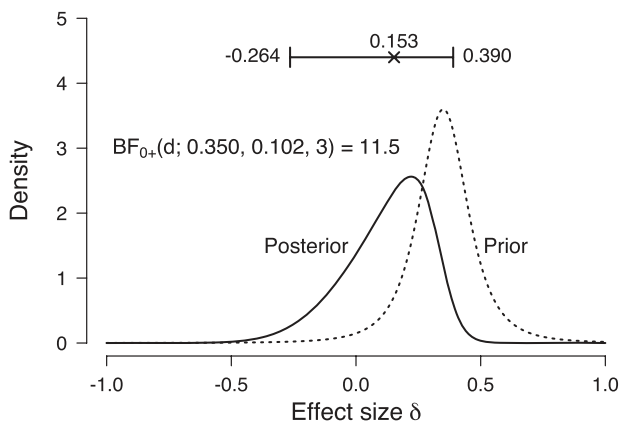

Figure 1. Results of an informed reanalysis of the facial feedback hypothesis replication data from the Oosterwijk lab. The dotted line corresponds to the elicited $\frac{1}{0.102} T_3\left(\frac{\delta - 0.350}{0.102}\right)$ prior distribution. The solid line corresponds to the associated posterior distribution, with a 95% credible interval and the posterior median displayed on top. The Bayes factor in favor of the null hypothesis over the one-sided informed alternative hypothesis equals $\mathrm{BF}_{0+}(d; 0.350, 0.102, 3) = 11.5$. Figure available at *https://tinyurl.com/mk7uaxm* under CC license *https://creativecommons.org/licenses/by/2.0/*.
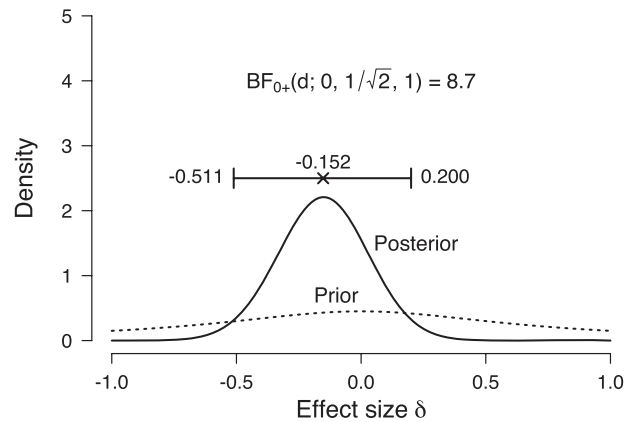

Figure 2. Results of the default analysis of the facial feedback hypothesis replication data from the Oosterwijk lab. The dotted line corresponds to the default Cauchy prior distribution with scale parameter $1/\sqrt{2}$. The solid line corresponds to the associated posterior distribution, with a 95% credible interval and the posterior median displayed on top. The Bayes factor in favor of the null hypothesis over the one-sided default alternative hypothesis equals $\mathrm{BF}_{0+}(d; 0, 1/\sqrt{2}, 1) = 8.7$. Figure available at *https://tinyurl.com/mgs28ob* under CC license *https://creativecommons.org/licenses/by/2.0/*.

the posterior median based on the default prior distribution is equal to −0.152 (95% credible interval: [−0.511, 0.200]).

4. Concluding Comments

The comparison between two means is a quintessential inference problem. Originally developed by Jeffreys (1948) in the one-sample setting, the Bayesian t-test has recently been extended to the two-sample set-up by Gönen et al. (2005) and, subsequently, by Rouder et al. (2009) and Wang and Liu (2016). Here, we showed that practitioners can easily and intuitively use a generalized version of the Bayes factor by Rouder et al. (2009) to inform their two-sample Bayesian t-tests. We used the framework of Gönen et al. (2005) and extended the priors by Rouder et al. (2009) to allow for more informed Bayesian t-tests that can incorporate expert knowledge by using a flexible t-prior. An advantage of the flexible t-prior is that it contains the objective default prior by Rouder et al. (2009) as a special case and the subjective prior proposed by Gönen et al. (2005) as a limiting case. Therefore, practitioners can use the same formula to compute subjective and objective Bayesian t-tests. To encourage its adoption in applied work, we have implemented the proposed Bayesian t-test set-up in the open-source statistical program JASP (JASP Team 2018, *jasp-stats.org*). In the theoretical part of this article, we investigated theoretical properties of the informed t-prior. Specifically, we discussed popular Bayes factor desiderata and proposed measures to quantify the deviation of an informed t-test from its objective counterpart. In the practical part of the article, we illustrated the use of the informed Bayes factor with an example. Similar to the prior proposed by Gönen et al. (2005), the flexible t-prior may encourage the use of prior distributions that better represent the predictions from the hypothesis under test, allowing more meaningful conclusions to be drawn from the same data (Rouder et al. 2016a, 2016b).

Other choices than a t-prior for effect size are conceivable. Equation (3) shows that one can obtain a Bayes factor for any scale-mixture of normals by integrating Equation (3) with

respect to a prior on g (see Theorem A.3 in the online appendix, supplementary materials; for possible choices see, e.g., Liang et al. 2008 and Bayarri et al. 2012). This also includes the prior proposed by Wang and Liu (2016) and highlights that it is straightforward to extend this prior to include a location parameter that can be specified based on expert knowledge. In fact, the expressions for the Bayes factor that we presented make it relatively straightforward to use *any* proper prior on standardized effect size (see Equation (4)). The proposed departure measures can then be used to investigate information consistency and predictive matching for different choices.

In this article, we focused on the Bayes factor as the inferential tool for quantifying the relative evidence for competing hypotheses based on observed data. However, it could be argued that a complete Bayesian analysis requires one to also specify the prior plausibilities of the competing hypotheses. This is of particular importance in situations where unlikely hypotheses are tested or when multiple comparisons are considered (Scott and Berger 2010). Although specifying the prior plausibilities of the competing hypotheses may not be trivial, once this has been achieved, the Bayes factor can be simply multiplied by the prior odds to obtain the posterior odds of interest.

Supplementary Materials

Online Appendix: Informed Bayesian *t*-Tests: Derivations, details about prior elicitation, and additional analyses. (pdf)

Acknowledgments

Funding

References

Bateman, H., Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F. G., Bertin, D., Fulks, W. B., Harvey, A. R., Thomsen, D. L. Jr, Weber, M. A., Whitney, E. L., and Stampfel, R. (1953), *Higher Transcendental Functions* (Vol. 2), New York: McGraw-Hill. [139]

——— (1954), *Tables of Integral Transforms* (Vol. 1), New York: McGraw-Hill. [139]

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012), "Criteria for Bayesian Model Choice With Application to Variable Selection," *The Annals of Statistics*, 40, 1550–1577. [138,139,142]

Berger, J. O., and Pericchi, L. R. (2001), "Objective Bayesian Methods for Model Selection: Introduction and Comparison" (with discussion), in *Model Selection*, ed. P. Lahiri (Institute of Mathematical Statistics Lecture Notes—Monograph Series, Vol. 38), Beachwood, OH: Institute of Mathematical Statistics, pp. 135–207. [138]

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998), "Bayes Factors and Marginal Distributions in Invariant Situations," *Sankhyā: The Indian Journal of Statistics*, Series A, 60, 307–321. [138]

Etz, A., and Wagenmakers, E.-J. (2017), "J. B. S. Haldane's Contribution to the Bayes Factor Hypothesis Test," *Statistical Science*, 32, 313–329. [137]

Gönen, M., Johnson, W. O., Lu, Y., and Westfall, P. H. (2005), "The Bayesian Two-Sample *t*-Test," *The American Statistician*, 59, 252–257. [137,138,139,141]

——— (2019), "Comparing Objective and Subjective Bayes Factors for the Two-Sample Comparison: The Classification Theorem in Action," *The American Statistician*, 73, 22–31. [138,139]

JASP Team (2018), "JASP (Version 0.9) [Computer Software]." [141]

Jeffreys, H. (1942), "On the Significance Tests for the Introduction of New Functions to Represent Measures," *Proceedings of the Royal Society of London*, Series A, 180, 256–268. [139]

——— (1948), *Theory of Probability* (2nd ed.). Oxford: Oxford University Press. [137,138,141]

——— (1961), *Theory of Probability* (3rd ed.), Oxford: Oxford University Press. [138]

Kadane, J. B., and Wolfson, L. J. (1998), "Experiences in Elicitation," *Journal of the Royal Statistical Society*, Series D, 47, 3–19. [140]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [137]

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of *g* Priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103, 410–423. [138,142]

Ly, A., Marsman, M., and Wagenmakers, E.-J. (2018), "Analytic Posteriors for Pearson's Correlation Coefficient," *Statistica Neerlandica*, 72, 4–13. [139]

Ly, A., Verhagen, A. J., and Wagenmakers, E.-J. (2016a), "An Evaluation of Alternative Methods for Testing Hypotheses, From the Perspective of Harold Jeffreys," *Journal of Mathematical Psychology*, 72, 43–55. [137]

——— (2016b), "Harold Jeffreys's Default Bayes Factor Hypothesis Tests: Explanation, Extension, and Application in Psychology," *Journal of Mathematical Psychology*, 72, 19–32. [137]

Maruyama, Y., and George, E. I. (2011), "Fully Bayes Factors With a Generalized *g*-Prior," *The Annals of Statistics*, 39, 2740–2765. [138]

Morey, R. D., and Rouder, J. N. (2015), BayesFactor 0.9.11-1, Comprehensive R Archive Network. [140]

Morey, R. D., and Wagenmakers, E.-J. (2014), "Simple Relation Between Bayesian Order-Restricted and Point-Null Hypothesis Tests," *Statistics and Probability Letters*, 92, 121–124. [141]

Morris, D. E., Oakley, J. E., and Crowe, J. A. (2014), "A Web-Based Tool for Eliciting Probability Distributions From Experts," *Environmental Modelling & Software*, 52, 1–4. [140]

Nuijten, M. B., Hartgerink, C. H., Assen, M. A., Epskamp, S., and Wicherts, J. M. (2016), "The Prevalence of Statistical Reporting Errors in Psychology (1985–2013)," *Behavior Research Methods*, 48, 1205–1226. [137,142]

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [138,140]

Robert, C. P., Chopin, N., and Rousseau, J. (2009), "Harold Jeffreys's Theory of Probability Revisited," *Statistical Science*, 24, 141–172. [137]

Rouder, J. N., Morey, R. D., Verhagen, A. J., Province, J. M., and Wagenmakers, E.-J. (2016a), "Is There a Free Lunch in Inference?," *Topics in Cognitive Science*, 8, 520–547. [141]

Rouder, J. N., Morey, R. D., and Wagenmakers, E.-J. (2016b), "The Interplay Between Subjectivity, Statistical Practice, and Psychological Science," *Collabra*, 2, 1–12. [141]

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009), "Bayesian *t*-Tests for Accepting and Rejecting the Null Hypothesis," *Psychonomic Bulletin & Review*, 16, 225–237. [137,138,139,141]

Scott, J. G., and Berger, J. O. (2010), "Bayes and Empirical—Bayes Multiplicity Adjustment in the Variable—Selection Problem," *The Annals of Statistics*, 38, 2587–2619. [142]

Severini, T. A., Mukerjee, R., and Ghosh, M. (2002), "On an Exact Probability Matching Property of Right-Invariant Priors," *Biometrika*, 89, 952–957. [138]

Strack, F., Martin, L. L., and Stepper, S. (1988), "Inhibiting and Facilitating Conditions of the Human Smile: A Nonobtrusive Test of the Facial Feedback Hypothesis," *Journal of Personality and Social Psychology*, 54, 768–777. [140]

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin–Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin–Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., Dijkstra, K., Fischer, A. H., Foroni, F., Hess, U., Holmes, K. J., Jones, J. L. H., Klein, O., Koch, C., Korb, S., Lewinski, P., Liao, J. D., Lund, S., Lupiáñez, J., Lynott, D., Nance, C. N., Oosterwijk, S., Özdoğru, A. A., Pacheco–Unguetti, A. P., Pearson, B., Powis, C., Riding, S., Roberts, T.-A., Rumiati, R. I., Senden, M., Shea–Shumsky, N. B., Sobocko, K., Soto, J. A., Steiner, T. G., Talarico, J. M., van Allen, Z. M., Vandekerckhove, M., Wainwright, B., Wayand, J. F., Zeelenberg, R., Zetzer, E. E., and Zwaan, R. A. (2016), "Registered Replication Report: Strack, Martin, & Stepper (1988)," *Perspectives on Psychological Science*, 11, 917–928. [140]

Wang, M., and Liu, G. (2016), "A Simple Two-Sample Bayesian *t*-Test for Hypothesis Testing," *The American Statistician*, 70, 195–201. [138,139,141,142]

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., and Wagenmakers, E.-J. (2011), "Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 *t*-Tests," *Perspectives on Psychological Science*, 6, 291–298. [137]

Wrinch, D., and Jeffreys, H. (1921), "On Certain Fundamental Principles of Scientific Inquiry," *Philosophical Magazine*, 42, 369–390. [137]

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions," *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, 6, 233–243. [138]