UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

The Bayesian Methodology of Sir Harold Jeffreys as a Practical Alternative to the P Value Hypothesis Test

Ly, A.; Stefan, A.; van Doorn, J.; Dablander, F.; van den Bergh, D.; Sarafoglou, A.; Kucharský, S.; Derks, K.; Gronau, Q.F.; Raj, A.; Boehm, U.; van Kesteren, E.-J.; Hinne, M.; Matzke, D.; Marsman, M.; Wagenmakers, E.-J.

Link to publication

ORIGINAL PAPER

The Bayesian Methodology of Sir Harold Jeffreys as a Practical Alternative to the *P* Value Hypothesis Test

Alexander Ly[1,2] · Angelika Stefan[1] · Johnny van Doorn[1] · Fabian Dablander[1] · Don van den Bergh[1] ·
Alexandra Sarafoglou[1] · Šimon Kucharský[1] · Koen Derks[3] · Quentin F. Gronau[1] · Akash Raj[1] · Udo Boehm[1] ·
Erik-Jan van Kesteren[1] · Max Hinne[4] · Dora Matzke[1] · Maarten Marsman[1] · Eric-Jan Wagenmakers[1] 🄳

Abstract
Despite an ongoing stream of lamentations, many empirical disciplines still treat the *p* value as the sole arbiter to separate the scientific wheat from the chaff. The continued reign of the *p* value is arguably due in part to a perceived lack of workable alternatives. In order to be workable, any alternative methodology must be (1) *relevant*: it has to address the practitioners' research question, which—for better or for worse—most often concerns the test of a hypothesis, and less often concerns the estimation of a parameter; (2) *available*: it must have a concrete implementation for practitioners' statistical workhorses such as the *t* test, regression, and ANOVA; and (3) *easy to use*: methods that demand practitioners switch to the theoreticians' programming tools will face an uphill struggle for adoption. The above desiderata are fulfilled by Harold Jeffreys's Bayes factor methodology as implemented in the open-source software JASP. We explain Jeffreys's methodology and showcase its practical relevance with two examples.

Keywords Bayes factor · Hypothesis testing · Replication · Statistical evidence · Statistical practice

Introduction

Frequently maligned and rarely praised, the practice of *p* value null hypothesis significance testing (NHST) nevertheless remains entrenched as the dominant method for drawing scientific conclusions from noisy data. The recent ASA policy statement on *p* values (Wasserstein and Lazar (2016); see also Wasserstein et al. (2019)) warns against common misinterpretations and abuses, but does not take a strong stance on what methodology should supplement or supplant the *p* value. Left without a concrete and workable alternative to test their hypotheses, practitioners arguably have little choice but to continue reporting and interpreting *p* values.

Promising alternatives to *p* value NHST do exist, however, and one of the most noteworthy is the Bayesian testing methodology advanced and promoted by the geophysicist Sir Harold Jeffreys (e.g., Etz and Wagenmakers 2017; Howie 2002; Jeffreys 1961; Ly et al. 2016a, b and Robert et al. 2009). We believe that Jeffreys's methodology fulfils three important desiderata. First, the methodology is grounded on the assumption that a general law (e.g., the null hypothesis) has a positive point mass, allowing practitioners to quantify the degree to which it is supported or undermined by the data. Second, Jeffreys implemented his conceptual framework for several run-of-the-mill statistical scenarios; to facilitate practical application, he also derived adequate approximations for unwieldy analytic solutions. Third, Jeffreys's tests and recent extensions are available in the open-source software program JASP (Jeffreys's Amazing Statistics Program; jasp-stats.org). JASP has an attractive interface that allows the user to obtain a reproducible and comprehensive Bayesian analysis with only a few mouse clicks.

The outline of this paper follows that of the above desiderata: first, we summarise Jeffreys's philosophy of scientific learning (developed together with his early collaborator Dorothy Wrinch), which makes the methodology

✉ Eric-Jan Wagenmakers
EJ.Wagenmakers@gmail.com

1 University of Amsterdam, Amsterdam, The Netherlands

2 Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

3 Nyenrode Business University, Breukelen, The Netherlands

4 Radboud University, Nijmegen, The Netherlands

relevant for practitioners; then, we describe how Jeffreys created concrete tests for common inferential tasks, making the methodology *available*; finally, we present two examples that demonstrate how JASP allows practitioners to obtain comprehensive Bayesian analyses with minimal effort, making the methodology *easy to use*.

First Desideratum: Relevance from Wrinch and Jeffreys's Philosophy of Scientific Learning

At the start of the twentieth century, Wrinch and Jeffreys (1919, 1921, 1923) set out to develop a theory of scientific learning that allows one to quantify the degree to which observed data corroborate or undermine a general law (e.g., the null hypothesis $\mathcal{H}_0$; within the context of Newton's equation, for instance, $\mathcal{H}_0$ may stipulate that the value of the gravitational constant equals $G = 6.674 \times 10^{-11} \frac{\text{m}^3}{\text{s}^2\text{kg}}$). Their theory demanded that (i) the general law is instantiated as a separate model $\mathcal{M}_0$ and (ii) the inference process is Bayesian. The first postulate is a manifestation of what may be called Jeffreys's razor: 'Variation is random until the contrary is shown; and new parameters in laws, when they are suggested, must be tested one at a time unless there is specific reason to the contrary.' (Jeffreys 1961, p. 342). The idea is shared with the p value, but Wrinch and Jeffreys also demanded the specification of an alternative model $\mathcal{M}_1$ that relaxes the general law (e.g., Newton's equation with G unrestricted, thus, an unknown parameter that has to be inferred from data).

The second postulate was the foundation for all of Jeffreys's work in statistics: 'The essence of the present theory is that no probability, direct, prior, or posterior, is simply a frequency.' (Jeffreys 1961, p. 401). This postulate implies that all unknowns, such as whether $\mathcal{M}_0$ or $\mathcal{M}_1$ holds true, should be formalised as probabilities or 'a reasonable degree of belief' (Jeffreys 1961, p. 401). By setting prior model probabilities $P(\mathcal{M}_0) \in (0, 1)$ and $P(\mathcal{M}_1) = 1 - P(\mathcal{M}_0)$, these can then be updated to posterior model probabilities $P(\mathcal{M}_j \mid d)$, $j = 0, 1$ in light of the data d using Bayes' rule. The ratio of these two quantities yields the crucial equation as follows:

$$\underbrace{\frac{P(\mathcal{M}_1 \mid d)}{P(\mathcal{M}_0 \mid d)}}_{\text{Posterior model odds}} = \underbrace{\frac{p(d \mid \mathcal{M}_1)}{p(d \mid \mathcal{M}_0)}}_{\text{BF}_{10}(d)} \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}}_{\text{prior model odds}} \tag{1}$$

where $p(d \mid \mathcal{M}_j) = \int f(d \mid \theta_j)\pi_j(\theta_j)\,d\theta_j$ is the marginal likelihood of $\mathcal{M}_j$ with respect to the prior density $\pi_j(\theta_j) = \pi(\theta_j \mid \mathcal{M}_j)$ on the free parameters within $\mathcal{M}_j$.

The Bayes Factor

The updating term $\text{BF}_{10}(d)$ is known as the *Bayes factor* and quantifies the change from prior to posterior model odds brought about by the observations d. The Bayes factor has an intuitive interpretation: $\text{BF}_{10}(d) = 7$ indicates that the data are 7 times more likely under $\mathcal{M}_1$ than under $\mathcal{M}_0$, whereas $\text{BF}_{10}(d) = 0.2$ indicates that the observations are 5 times more likely under $\mathcal{M}_0$ than under $\mathcal{M}_1$. The Bayes factor can be used to quantify the evidence against, but also in support of $\mathcal{M}_0$, and it only depends on the observed data, unlike the p value which also depends on more extreme—but not observed—outcomes of a statistic.[1]

The Bayes factor methodology can be generalised in several ways. Suppose that data d_1 are observed and two or *more* competing models are entertained. The posterior probability for model $\mathcal{M}_j$ can be obtained as follows:

$$P(\mathcal{M}_j \mid d_1) = \frac{\text{BF}_{j0}(d_1)P(\mathcal{M}_j)}{\sum_{j'} \text{BF}_{j'0}(d_1)P(\mathcal{M}_{j'})}, \tag{2}$$

where the sum in the denominator is over the set of all candidate models, and $\text{BF}_{00}(d_1) = 1$. If an all-or-none decision on the model space is required, we can use the $P(\mathcal{M}_j \mid d_1)$ to select a single model. Once a single model $\mathcal{M}_j$ is selected, the knowledge about its parameters is then given by the posterior density defined as follows:

$$\pi_j(\theta_j \mid d_1, \mathcal{M}_j) = \frac{f(d_1 \mid \theta_j)\pi(\theta_j \mid \mathcal{M}_j)}{\int f(d_1 \mid \theta_j)\pi(\theta_j \mid \mathcal{M}_j)\,d\theta_j}. \tag{3}$$

Alternatively, we may forego selection, retain the uncertainty across the model space and use Bayesian model averaging (BMA; e.g., Raftery et al. (1997)). To simplify matters, suppose that there are only two models and $\mathcal{M}_0$ has parameters θ_0, whereas $\mathcal{M}_1$ has parameters $\theta_1 = (\delta, \theta_0)$. We refer to θ_0 as the common parameters and refer to δ as the test-relevant parameter. The model-averaged posterior density of the common parameters θ_0 is as follows:

$$\pi(\theta_0 \mid d_1) = \sum_{j'} \pi(\theta_0 \mid d_1, \mathcal{M}_{j'})P(\mathcal{M}_{j'} \mid d_1). \tag{4}$$

The model-averaged posterior of the test-relevant parameter δ is a convex mixture of a point mass at zero and the posterior density within $\mathcal{M}_1$, that is,

$$\pi(\delta \mid d_1) = \mathbf{1}\{\delta = 0\}\,P(\mathcal{M}_0 \mid d_1) + \pi(\delta \mid d_1, \mathcal{M}_1)P(\mathcal{M}_1 \mid d_1). \tag{5}$$

Another natural generalisation of Jeffreys's Bayes factor framework concerns sequential analysis and evidence

---

[1] As famously stated by Jeffreys (1961, p. 385): 'What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure.'

updating. For instance, suppose that an original experiment yields data d_1, and a direct replication attempt yields data d_2. One may then compute a so-called replication Bayes factor (Ly et al. 2019) as follows:

$$\mathrm{BF}_{10}(d_2 \mid d_1) = \frac{\mathrm{BF}_{10}(d_1, d_2)}{\mathrm{BF}_{10}(d_1)}, \qquad (6)$$

whenever d_1 and d_2 are exchangeable. Note that $\mathrm{BF}_{10}(d_2 \mid d_1)$ quantifies the *additional* evidence for $\mathcal{M}_1$ over $\mathcal{M}_0$ provided by the data d_2 of the replication attempt given d_1. The total Bayes factor $\mathrm{BF}_{10}(d_1, d_2)$ is practical for collaborative learning, whereas the replication Bayes factor quantifies the individual contribution of the replication attempt.

Bayes Factor Rationale

For a deeper appreciation of the fact that scientific learning (i.e., the conformation and disconfirmation of a general law) requires both of Wrinch and Jeffreys's postulates, we illustrate their main argument with a recent example. Pisa et al. (2015) considered the hypothesis that Alzheimer's disease has a microbial aetiology; specifically, the hypothesis states that all Alzheimer's patients have fungal infections in their brains. To provide evidence for this $\mathcal{H}_0$, they investigated a sample of deceased patients. Let y denote the observed number of Alzheimer's patients with fungal infections in their brains out of a total number n who were examined. Here, $\mathcal{M}_0$ implies that $Y \sim \mathrm{Bin}(n, \vartheta)$ with the restriction $\mathcal{H}_0 : \vartheta = 1$, whereas $\mathcal{H}_1 : \vartheta \in (0, 1)$.

As $\mathcal{M}_0$ does not have any free parameters, only $\pi_1(\vartheta)$ needs to be specified in order to obtain a Bayes factor. Following Wrinch and Jeffreys, let $\pi_1(\vartheta) = 1$ on $(0, 1)$. A direct computation shows that $\mathrm{BF}_{01}(d) = (n+1)\mathbf{1}\{y = n\}$, where $\mathbf{1}\{y = n\}$ is one whenever $y = n$, and zero otherwise. For the data $d_1 : y_1 = n_1 = 10$ of Pisa et al. (2015), this yields $\mathrm{BF}_{01}(d_1) = 11$, thus evidence in favour of their proposed null hypothesis.

Note that with $\mathrm{BF}_{01}(d) = (n + 1)\mathbf{1}\{y = n\}$, the evidence for $\mathcal{M}_0$ continues to grow with n as long as only confirming cases are observed. Upon encountering a single disconfirming case, however, we have $\mathrm{BF}_{01}(d) = 0$ and $P(\mathcal{M}_0 \mid d) = 0$, meaning that $\mathcal{M}_0$ is falsified or 'irrevocably exploded' (Polya 1954, p. 6), regardless of the number of previously observed successes.

Bayes Factors Versus Bayesian Estimation

According to Wrinch and Jeffreys, a separate $\mathcal{M}_0$ is necessary for falsification and sensible generalisations of $\mathcal{H}_0$. This meant a radical departure from the Bayesian estimation procedures that had been standard practice ever since the time of Laplace. In the Laplace approach, the problem is treated as one of Bayesian estimation, and the uniform prior $\pi_1(\vartheta)$ under $\mathcal{M}_1$ implies that the prior probability of $\mathcal{H}_0 : \vartheta = 1$ is zero. For $d : y = n = 1,000$, $\pi_1(\vartheta \mid d)$ would then have a median of $\tilde{\vartheta} = 0.999$ and an approximate central 95% credible interval of $[0.996, 1.000]$. The introduction of a single failure, which 'irrevocably explodes' the null hypothesis, only leads to a small shift in the posterior distribution, resulting in a $\pi_1(\vartheta \mid d)$ with median $\tilde{\vartheta} = 0.998$ and an approximate central 95% credible interval of $[0.994, 1.000]$.

Furthermore, without a separate $\mathcal{M}_0$, the predictions for new data d_2 given d_1 only involve the posterior predictive under $\mathcal{M}_1$, that is, $p(d_2 \mid d_1, \mathcal{M}_1) = \int f(d_2 \mid \theta_1)\pi(\theta_1 \mid d_1, \mathcal{M}_1)\, \mathrm{d}\theta_1$. For example, when $d_1 : y_1 = n_1 = 10$, then a new sample $d_2 : y_2 = n_2$ consistent with $\mathcal{M}_0$ is predicted to occur with chance $p(d_2 \mid d_1, \mathcal{M}_1) = 11/(11 + n_2)$; when $n_2 = 11$, this chance is 50%; when $n_2 \to \infty$, this chance goes to 0. In other words, in the Laplace estimation framework, the proposition that fungal infections are *not* a necessary and sufficient cause of Alzheimer's disease is a foregone conclusion. Jeffreys (1980) felt such generalisations violated common sense.[2]

Generalisations based on the Wrinch and Jeffreys posterior model probabilities $P(\mathcal{M}_j \mid d_1)$ are very different from the ones that follow from the Laplace estimation approach. Suppose that Pisa et al. (2015) set $P(\mathcal{M}_0) = P(\mathcal{M}_1) = 1/2$, then $P(\mathcal{M}_0 \mid d_1) = \frac{n_1 + 1}{n_1 + 2} = \frac{11}{12} \approx 0.917$ and $P(\mathcal{M}_1 \mid d_1) \approx 0.083$. If $\mathcal{M}_0$ is selected—and the uncertainty of this selection, $P(\mathcal{M}_1 \mid d_1) = 0.083$, is ignored—then $d_2 : y_2 = n_2 = 11$ is predicted to occur with 100% chance, which is a quite confident and perhaps risky generalisation of d_1. On the other hand, the Bayesian model–averaged prediction is defined as follows:

$$p(d_2 \mid d_1) = \sum_{j'} p(d_2 \mid d_1, \mathcal{M}_{j'}) P(\mathcal{M}_{j'} \mid d_1), \qquad (7)$$

and for $d_2 : y_2 = n_2 = 11$ given $d_1 : y_1 = n_1 = 10$, this yields a chance of 95.7%.

Bayes Factors Versus the *P* value

The p value takes $\mathcal{M}_0$ seriously, but does not assign it a prior model probability; also, the computation of the p value does not involve the specification of $\mathcal{M}_1$. Consequently, p values do not allow probabilistic statements about $\mathcal{M}_0$. For the Alzheimer example with $y = n$, the p value equals 1 regardless of the number of confirmatory cases n, which illustrates that the p value cannot be used to quantify

---

[2] 'A general rule would never acquire a high probability until nearly the whole of the class had been inspected. We could never be reasonably sure that apple trees would always bear apples (if anything). The result is preposterous, and started the work of Wrinch and myself in 1919-1923.' (Jeffreys 1980, p. 452)

evidence in support of $\mathcal{M}_0$. It is also unclear how p values could be used to average predictions. The p value does, however, allow for falsification as the observation of a single failure yields $p = 0 < \alpha$ for any $\alpha > 0$.

Bayes Factor Relevance

The philosophy of scientific learning as articulated by Wrinch and Jeffreys is relevant to empirical researchers in two main ways. First of all, the philosophy allows one to test hypotheses. In contrast, some statistical reformers have argued that empirical researchers should abandon testing altogether and focus instead on estimating the size of effects (e.g., Cumming (2014)). We believe the estimation-only approach cannot address the primary goal that many researchers have, which is—for better or for worse—to convince a sceptical audience that their effects are real. Before interpreting the size of effects, one needs to be convinced that the effects exist at all. The estimation-only approach violates Jeffreys's razor, leads to questionable generalisations, and starts out by assuming the very thing that many researchers seek to establish: the presence of an effect (Haaf et al. 2019).

Wrinch and Jeffreys's philosophy of scientific learning is also relevant because it allows practitioners to quantify evidence—the degree to which the data support or undermine the hypothesis of interest. The Bayesian formalism allows researchers to update their knowledge in a coherent fashion, which allows a straightforward extension to techniques such as BMA and sequential analysis.

Second Desideratum: Availability for Common Statistical Scenarios

In Chapter V of 'Theory of Probability' (first edition 1939, second edition 1948, third edition 1961), Jeffreys demonstrates how his philosophy of scientific learning can be implemented for common statistical scenarios. In all scenarios considered, Jeffreys was confronted with a nested model comparison $\mathcal{M}_0 \subset \mathcal{M}_1$, and he set out to choose a pair of prior densities (π_0, π_1) from which to construct a Bayes factor, that is,

$$\text{BF}_{10}(d) = \frac{\int \int f(d \mid \theta_0, \delta) \pi_1(\theta_0, \delta) \mathrm{d}\theta_0 \mathrm{d}\delta}{\int f(d \mid \theta_0, \delta = 0) \pi_0(\theta_0) \mathrm{d}\theta_0}. \qquad (8)$$

The main challenge is that general 'noninformative' choices lead to Bayes factors with undesirable properties. Specifically, prior densities of arbitrary width yield Bayes factors that favour $\mathcal{M}_0$ irrespective of the observed data. Furthermore, the prior density on the test-relevant parameter

must be proper, as improper priors contain suppressed normalisation constants and lead to Bayes factors of arbitrary value.

To overcome these problems for his one-sample t test (e.g., Jeffreys (1961); p. 268–274; Ly et al. 2016a, b), Jeffreys used invariance principles and two desiderata to select π_0 and π_1. The desideratum of *predictive matching* states that the Bayes factor ought to be perfectly indifferent, i.e., $\text{BF}_{10}(d) = 1$, in case n is too small; the desideratum of *information consistency* implies falsification of $\mathcal{M}_0$, as it requires $\text{BF}_{10}(d) = \infty$ in case data of sufficient size are overwhelmingly informative against $\mathcal{H}_0$.

The same two desiderata that Jeffreys formulated for his one-sample t test were extended and used to construct Bayes factors for correlations (e.g., Ly et al. (2018b)), linear regression (e.g., Bayarri et al. (2012), Liang et al. (2008), and Zellner and Siow (1980)), ANOVA (e.g., Rouder et al. (2012)), and generalised linear models (e.g., Li and Clyde (2018))—scenarios that are highly relevant to contemporary empirical researchers.

In addition, recent results show that these (group invariant) Bayes factors are robust under optional stopping (Hendriksen et al. 2018; Grünwald et al. 2019). Hence, the data collection can be stopped when the result is sufficiently compelling or when the available resources have been depleted, without increasing the chance of a false discovery. Furthermore, it was shown that $\text{BF}_{j0}(d) = \mathcal{O}_P(n^{-1/2})$ for data generated under $\mathcal{M}_0$, and $\text{BF}_{j0}(d) = \mathcal{O}_P(e^{nD(\theta_j^* \| \mathcal{M}_0)})$ for data generated under $\mathcal{M}_j$ for $j \neq 0$, where $D(\theta_j^* \| \mathcal{M}_0)$ is the Kullback–Leibler projection of the true θ_j^* parameter value onto the null model (e.g., Johnson and Rossell (2010)). Hence, if $\delta \neq 0$, it will be detected relatively quickly.

Throughout his own work, Jeffreys devised ingenious approximations to facilitate the use of his Bayes factors in empirical research. But, despite the tests being relevant and readily available, empirical researchers for some years largely ignored Jeffreys's Bayes factors. Their lack of popularity can be explained not only by Jeffreys's unusual notation and compact writing style but also by the fact that during his lifetime Bayesian approaches were not yet mainstream.

Third Desideratum: Ease of Use Through JASP

In order to make Bayesian statistics more accessible to practitioners, we have developed `JASP`, an open-source software package with an intuitive graphical user interface. Built on `R` (R Development Core Team 2004), `JASP` allows practitioners to conduct comprehensive Bayesian analyses with minimal effort. `JASP` also offers frequentist
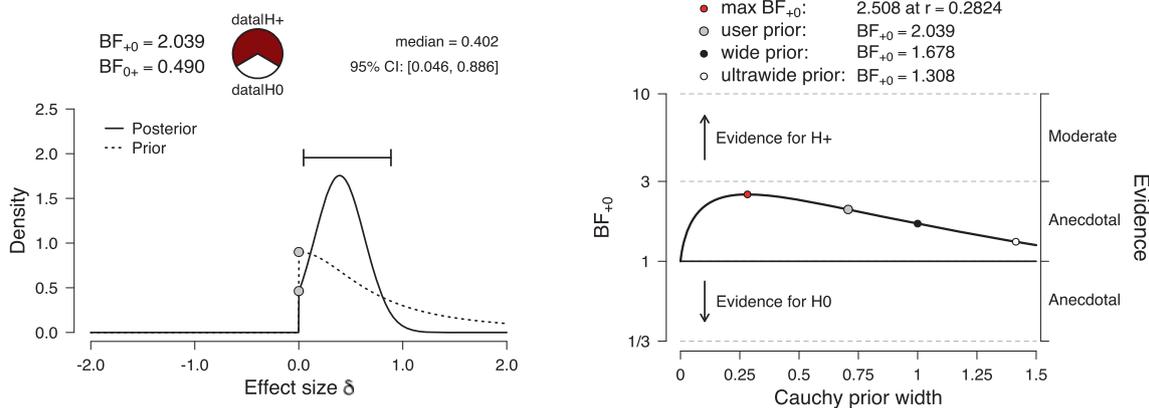
Fig. 1 A Bayes factor reanalysis of Strack et al. (1988). Left panel: The data show anecdotal evidence in favour of the alternative hypothesis. Right panel: Robustness check for different specifications of the width of the Cauchy prior on δ. Figures from JASP

methods, making it straightforward to check whether the two paradigms support the same conclusion.

One of JASP's design principles is that the Bayesian analyses come with default prior specifications that yield reference-style outcomes. For example, the default settings for most tests in JASP define Bayes factors that adhere to Jeffreys's method of learning, that is, they are predictively matched and information consistent. The following two examples demonstrate how JASP allows a straightforward application of Jeffreys's philosophy of scientific learning in practical situations.[3]

Example 1: Bayesian Independent Samples *t* Tests

The *t* test is designed to address the fundamental question of whether two means differ. Below, we first outline the default Bayes factor *t* test (see also Gönen et al. (2005) and Rouder et al. (2009)) and then use it to reanalyse a seminal psychological experiment as well as a recent replication attempt.

Bayesian Background

For the comparison of two sample means, the kth sample, $k = 1, 2$, is assumed to be normally distributed as $\mathcal{N}(\mu_k, \sigma^2)$, where $\mu_k = \mu_g + (-1)^{k-1} \sigma \delta / 2$ and where μ_g is interpreted as a grand mean and δ as the effect size. Hence, for $\mathcal{M}_0$, we have $\theta_0 = (\mu_g, \sigma)$, whereas $\mathcal{M}_1$ has one more parameter; $\theta_1 = (\delta, \theta_0)$. In this case, Jeffreys's invariance principle leads to the right Haar prior $\pi_0(\theta_0) \propto \sigma^{-1}$ on the common parameters and $\pi_1(\delta, \theta_0) = \pi_1(\delta)\pi_0(\theta_0)$. Predictive matching requires $\pi_1(\delta)$ to be symmetric around zero, whereas information consistency requires that $\pi_1(\delta)$ is without moments.

Following Jeffreys, $\pi_1(\delta)$ is assigned a zero-centred Cauchy prior, but with default width $\kappa = 1/\sqrt{2}$ (Rouder et al. 2012). This default Bayes factor is recommended at the start of an investigation, as it is invariant (a) under permutation of the group membership label k and (b) under location-scale transformation of the units of measurements, for instance, from Celsius to Fahrenheit.

Default Bayesian Analysis of an Original Study

To demonstrate this default Bayes factor in practice, we first reanalyse a seminal study on the *facial feedback* hypothesis of emotion. Specifically, Strack et al. (1988) reported that people rate cartoons as more amusing when they have to hold a pen between their teeth (the 'smile' condition) than when they have to hold a pen with their lips (the 'pout' condition).

Unfortunately, the raw data from this study are no longer available. However, the Summary Stats module in JASP affords a comprehensive Bayesian reanalysis using only the test statistics and the sample sizes obtained from the original article (Ly et al. 2018a).[4] According to Strack et al. (1988)'s working hypothesis, participants in the smile condition are expected to judge the cartoons to be more funny— not less funny—than participants in the pout condition. The directional nature of the alternative hypothesis is acknowledged by formalising it as $\mathcal{H}_+ : \delta > 0$, which we pitted against $\mathcal{H}_0 : \delta = 0$. Entering $d_1 : t = 1.85, n_1 = n_2 = 32$ into JASP immediately yields $\mathrm{BF}_{+0}(d_1) = 2.039$, which represents weak evidence for Strack et al. (1988)'s working hypothesis. A single tick mark in JASP produces the left panel plot from Fig. 1. This plot shows the prior

---

[3]Data sets and analysis scripts can be found at osf.io/7b6ws/.

[4]The original experiment also featured a neutral condition; the required summary statistics were obtained by assuming homogeneous variance and homogeneous sample size across the three experimental conditions.

and posterior distributions for effect size under $\mathcal{H}_+$. The posterior is relatively wide, indicating that there remains considerable uncertainty about the size of the effect. In addition, the two grey dots at $\delta = 0$ provide a visual representation of the Bayes factor as the ratio of the prior and posterior density under $\mathcal{M}_+$ evaluated at the restriction $\mathcal{H}_0 : \delta = 0$ (Dickey 1971). The probability wheel on top confirms that the evidence for $\mathcal{M}_+$ is rather meagre. The area of the red surface equals $P(\mathcal{M}_+ \mid d_1)$ under the assumption that $P(\mathcal{M}_0) = P(\mathcal{M}_+) = 1/2$.

Another tick mark in JASP produces the right panel plot from Fig. 1. This plot shows the robustness of the Bayes factor to the width κ of the Cauchy prior on effect size. In this case, reasonable changes in κ lead to only modest changes in the resulting Bayes factor. The classification scheme on the right axis is based on Jeffreys (1961, Appendix B) and provides a rough guideline for interpreting the strength of evidence.

In sum, from a relatively comprehensive Bayesian reanalysis, we conclude that (a) the data from the original experiment provide weak evidence in favour of Strack et al. (1988)'s working hypothesis; (b) this inference is robust to reasonable changes in the width of the Cauchy prior on effect size; (c) there remains considerable uncertainty about the size of the effect, assuming that it is positive. The entire analysis is obtained from JASP by typing three numbers and ticking two boxes.

Informed Bayesian Analysis of a Replication Attempt

Wagenmakers et al. (2016) recently conducted a multi-lab direct replication of the original facial feedback study. We use data from the Wagenmakers' lab and compare the results of a default analysis to an informed analysis based on expert knowledge and to an analysis that conditions on the original data.

First, the same default analysis as described in the previous section yields $\mathrm{BF}_{+0}(d_2) = 0.356$, thus, weak evidence in favour of $\mathcal{M}_0$. Second, Gronau et al. (2017) elicited an informed prior density from Suzanne Oosterwijk, an expert in the field. This elicitation effort resulted in a positive-only t-prior on δ with location 0.350, scale 0.102, and 3 degrees of freedom. The left panel of Fig. 2 shows that $\mathrm{BF}_{+0}^{\mathrm{info}}(d_2) = 0.597$, which is weak evidence in favour of $\mathcal{M}_0$. The right panel of Fig. 2 shows the evidential flow, that is, the progression of the Bayes factor as the experimental data accumulate.

Third, the test may also be informed by using $\pi_+(\delta \mid d_1)$ as a prior, which is equivalent to conditioning on the original data set, when d_1 and d_2 are exchangeable. As the total Bayes factor is $\mathrm{BF}_{+0}(d_1, d_2) = 1.400$, this leads to $\mathrm{BF}_{+0}(d_2 \mid d_1) = 0.686$, which is again weak evidence in favour of the null hypothesis.

We can also examine the maximum evidence—that is, the evidence obtained when a proponent of the effect were to cherry-pick a 'prior' after data inspection in order to achieve the highest possible Bayes factor (Edwards et al. 1963). This Bayes factor is obtained using an oracle prior on δ that assigns all prior mass to the maximum likelihood estimate, approximately $\hat{d} = 0.4625$, which leads to $\mathrm{BF}_{+0}^{\mathrm{max}}(d_1) = 5.26$. Although this cheating strategy leads to more evidence in favour of $\mathcal{M}_1$ given original data d_1, in light of new data d_2, it leads to a larger decrease, as $\mathrm{BF}_{+0}^{\mathrm{max}}(d_1, d_2) = 1.97$, and, therefore, $\mathrm{BF}_{+0}^{\mathrm{max}}(d_2 \mid d_1) = 0.37$, which is more evidence for $\mathcal{M}_0$ than provided by the default replication Bayes factor $\mathrm{BF}_{+0}(d_2 \mid d_1) = 0.69$.

The above results are summarised in Table 1. None of the analyses provide compelling evidence for the hypothesis
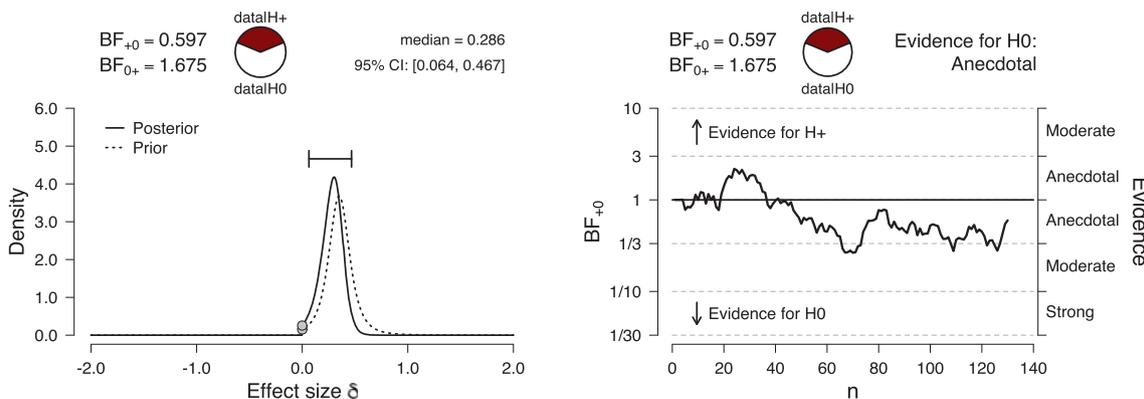

Fig. 2 An informed Bayesian analysis of the Wagenmakers replication data reported in Wagenmakers et al. (2016). Left panel: The Bayes factor indicates anecdotal evidence in favour of the null hypothesis. Right panel: Sequential monitoring of the Bayes factor as experimental data accumulate. Figures from JASP

Table 1 Bayes factors for the original study by Strack et al. (1988) and the replication attempt by Wagenmakers et al. (2016)

Bayes factor	Effect size prior $\pi_+(\delta)$	Data
$\text{BF}_{+0}(d_1) = 2.04$	Cauchy$(0, 1/\sqrt{2})\mathbf{1}\{\delta \geq 0\}$	Original
$\text{BF}_{+0}(d_2) = 0.36$	Cauchy$(0, 1/\sqrt{2})\mathbf{1}\{\delta \geq 0\}$	Replication
$\text{BF}_{+0}^{\text{info}}(d_2) = 0.60$	$t(0.350, 0.102, \text{df} = 3)\mathbf{1}\{\delta \geq 0\}$	Replication
$\text{BF}_{+0}(d_1, d_2) = 1.40$	Cauchy$(0, 1/\sqrt{2})\mathbf{1}\{\delta \geq 0\}$	Combined
$\text{BF}_{+0}(d_2 \mid d_1) = 0.69$	$\pi_+(\delta \mid d_1)$	Replication
$\text{BF}_{+0}^{\text{max}}(d_2 \mid d_1) = 0.37$	$\mathbf{1}\{\delta = 0.4625\}$	Replication

that cartoons are judged to be more funny in the smile condition than in the pout condition. The analyses that involve only the replication attempt all show weak evidence in favour of the null hypothesis. The analyses can be executed in JASP with minimal effort.

Example 2: Bayesian Linear Regression

Linear regression allows researchers to examine the association between a criterion variable and several possible predictors or covariates. Whereas the t test featured only two models, $\mathcal{M}_0$ and $\mathcal{M}_1$, linear regression features 2^p models, where p is the total number of mean-centred potential predictors. With a model space this large, it is unlikely that any single model will dominate; consequently, a Bayesian model averaging (BMA) approach is advisable.

Bayesian Background

Each regression model can be identified with an indicator vector $\gamma \in \{0, 1\}^p$, where $\gamma_k = 1$ implies that the kth predictor is included in the model, whereas $\gamma_k = 0$ implies that it is excluded. Note that $|\gamma| = \sum_{k=1}^p \gamma_k$ defines the number of predictors included in $\mathcal{M}_\gamma$. We then have the following:

$$\mathcal{M}_\gamma : Y \sim \mathcal{N}(\alpha \vec{1} + X_\gamma \beta_\gamma, \sigma^2 I), \qquad (9)$$

where Y is the criterion variable, $\vec{1}$ is a vector of ones, I is the $n \times n$ identity matrix, and α and σ^2 are the intercept and variance parameters that are present in all models. Importantly, X_γ and β_γ are the included columns of the design matrix and regression coefficients, respectively. The baseline model $\mathcal{M}_0$ is specified by γ of all zeroes, which all 2^p models can be compared to. For $\text{BF}_{\gamma 0}(d)$ to not depend on the units of measurements Y and X_γ, we set $\pi_0(\alpha, \sigma) \propto \sigma^{-1}$, and the coefficients are reparametrised in

terms of $\xi_\gamma = \left(\frac{1}{n} X_\gamma^T X_\gamma\right)^{1/2} \beta_\gamma / \sigma$, which implies that the regression is effectively executed on the standardised X_γ. Predictive matching and information consistency follow from choosing a zero-centered multivariate Cauchy prior on ξ_γ within $\mathcal{M}_\gamma$ (Zellner and Siow 1980). To facilitate computation, the multivariate Cauchy prior can be rewritten as a mixture of normals (Liang et al. 2008), which amounts to $\beta_\gamma \sim \mathcal{N}\left(0, g\left(X_\gamma^T X_\gamma\right)^{-1} \sigma^2\right)$ and $g \sim \text{InvGam}\left(\frac{1}{2}, \frac{n\kappa^2}{2}\right)$.

With 2^p models, the importance of an individual predictor may be assessed by averaging over the model space. In JASP, default prior model probabilities are assigned with a beta-binomial distribution. In particular, with beta(1, 1), this yields $P(\mathcal{M}_\gamma) = \frac{1}{p+1}\binom{p}{|\gamma|}^{-1}$, which implies that the prior mass is first divided evenly over the classes of models with $0, 1, \ldots, p$ number of active predictors, and then, within each such class, the $1/(p + 1)$ mass is divided evenly over the total number of available models. This beta-binomial prior assignment acts as an automatic correction for multiplicity (Scott and Berger 2010).

The JASP implementation of this linear regression analysis uses the R package BAS (Clyde 2010), which has been designed for computational efficiency and offers a range of informative output summaries.

Bayesian Analysis of the US Crime Rate Data

We analyse a well-known data set concerning US crime rates (Ehrlich 1973). This data set contains aggregate information about the crime rates of 47 US states, as well as a variety of demographic variables such as labour force participation and income inequality. Without the indicator variable for a Southern state, there are 14 predictors in total, and hence $2^{14} = 16,384$ models to consider. After identifying the criterion variable and the potential predictors, ticking the box 'Posterior Summaries of Coefficients' makes JASP return the information shown in Table 2. For each regression coefficient, averaged across all 16,384 models, the table provides the posterior mean, the posterior standard deviation, the prior and posterior model probabilities summed across the models that include the predictor (i.e., $P(\text{incl})$ and $P(\text{incl} \mid \text{data})$, respectively), and a posterior central 95% credible interval.

A more detailed inspection is provided by ticking the option 'Marginal posterior distributions'. For instance, for the predictor 'police expenditure', this results in Fig. 3, which shows a model-averaged posterior density for the

Table 2 Posterior summaries of regression coefficients for the US crime rate data

Coefficient	Mean	SD	P (incl)	P (incl | data)	BF$_{inclusion}$	95% Credible interval	
						Lower	Upper
Intercept	6.72	0.03	1.00	1.00	1.00	6.67	6.79
M	1.30	0.59	0.50	0.93	13.08	0.00	2.27
Ed	1.88	0.62	0.50	0.98	45.82	0.63	3.14
Po1	0.60	0.58	0.50	0.74	2.88	− 0.45	1.86
Po2	0.22	0.59	0.50	0.59	1.43	− 1.03	1.47
LF	0.14	0.44	0.50	0.45	0.81	− 0.68	1.34
M.F	− 0.72	1.50	0.50	0.49	0.98	− 5.22	1.30
Pop	− 0.04	0.05	0.50	0.62	1.61	− 0.16	0.01
NW	0.09	0.05	0.50	0.86	6.23	0.00	0.18
U1	−0.09	0.24	0.50	0.50	1.00	− 0.73	0.41
U2	0.29	0.24	0.50	0.78	3.50	− 0.02	0.76
GDP	0.38	0.43	0.50	0.63	1.72	− 0.10	1.35
Ineq	1.42	0.37	0.50	1.00	341.82	0.68	2.21
Prob	− 0.23	0.11	0.50	0.93	13.33	− 0.41	0.00
Time	− 0.14	0.18	0.50	0.61	1.58	− 0.57	0.11

M, percentage of males aged 14–24; *Ed*, mean years of schooling; *Po1*, police expenditure in 1960; *Po2*, police expenditure in 1959; *LF*, labour force participation rate; *M.F*, number of males per 1000 females; *Pop*, state population; *NW*, number of non-whites per 1000 people; *U1*, unemployment rate of urban males aged 14–24; *U2*, unemployment rate of urban males 35–39; *GDP*, gross domestic product per capita; *Ineq*, income inequality; *Prob*, probability of imprisonment; *Time*, average time served in state prisons. Table from JASP

associated regression coefficient. The vertical bar represents the point mass at zero; the posterior mass away from zero is obtained as a weighted mix of the posterior distributions for the models that include 'police expenditure' as a predictor. The mixing is apparent from the slightly irregular shape of the model-averaged posterior.
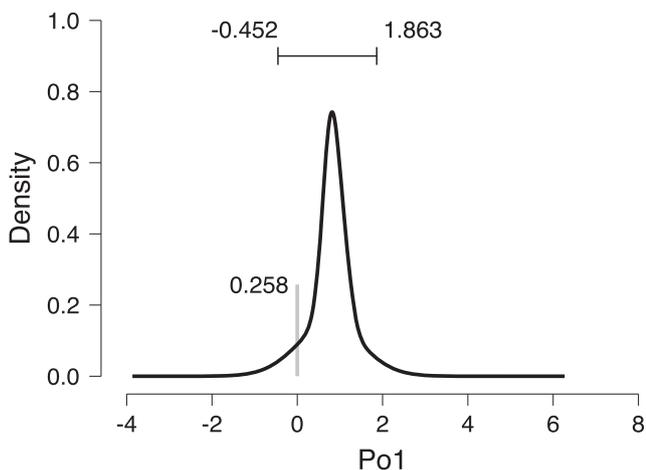
Fig. 3 Model-averaged posterior densities over police expenditure in 1960 (Po1)

Concluding Remarks

Why have theoretical statisticians been unable to move the needle on the dominance of the *p* value? What has prevented practitioners from adopting a more inclusive statistical approach? The answer, we believe, is that statistics has largely failed to offer practitioners a workable alternative to the *p* value. In this paper, we have argued that the testing methodology of Sir Harold Jeffreys—developed in the 1930s and recently implemented in the JASP program—constitutes such a workable alternative. With JASP, users can compute *p* values; but, with just a few mouse clicks, they can also conduct a relatively comprehensive Bayesian analysis. Users may adjust prior distributions, quantify and monitor evidence, conduct sensitivity analyses, engage in model averaging, and more.

We believe that Bayesian statistics deserves to be applied more often, especially in practical, run-of-the-mill research problems. A pragmatist may argue that –irrespective of one's statistical convictions—it is prudent to report the results from multiple paradigms. When the results point in the same direction, this bolsters one's confidence in the conclusions; when the results are in blatant contradiction, one's confidence is tempered—either way, something useful has been learned.

References

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

Clyde, M.A. (2010). BAS: Bayesian adaptive sampling for Bayesian model averaging 0.90. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/BAS/index.html.

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

Dickey, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *41*, 204–223.

Edwards, W., Lindman, H., Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Ehrlich, I. (1973). Participation in illegitimate activities: a theoretical and empirical investigation. *Journal of political Economy*, *81*, 521–565.

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

Gönen, M., Johnson, W.O., Lu, Y., Westfall, P.H. (2005). The Bayesian two-sample t test. *The American Statistician*, *59*, 252–257.

Gronau, Q.F., Ly, A., Wagenmakers, E.-J. (2017). Informed Bayesian *t*-tests. arXiv:1704.02479.

Grünwald, P., de Heide, R., Koolen, W. (2019). Safe testing. arXiv:1906.07801.

Haaf, J.M., Ly, A., Wagenmakers, E.-J. (2019). Retire significance, but still test hypotheses. *Nature*, *567*, 461.

Hendriksen, A., de Heide, R., Grünwald, P. (2018). Optional stopping with Bayes factors: a categorization and extension of folklore results, with an application to invariant situations. arXiv:1807.09077.

Howie, D. (2002). *Interpreting probability: controversies and developments in the early twentieth century*. Cambridge: Cambridge University Press.

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

Jeffreys, H. (1980). Some general points in probability theory. In *Bayesian analysis in econometrics and statistics: essays in honor of Harold Jeffreys (eds. A. Zellner and B. Kadane, Joseph), 451–453. Amsterdam: North-Holland*.

Johnson, V.E., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*, 143–170.

Li, Y., & Clyde, M.A. (2018). Mixtures of g-priors in generalized linear models. arXiv:1503.06913.

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

Ly, A., Etz, A., Marsman, M., Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior Research Methods*, *51*, 2498–2508.

Ly, A., Komarlu Narendra Gupta, A.R., Etz, A., Marsman, M., Gronau, Q.F., Wagenmakers, E.-J. (2018a). Bayesian reanalyses from summary statistics and the strength of statistical evidence. *Advances in Methods and Practices in Psychological Science*, *1*, 367–374.

Ly, A., Marsman, M., Wagenmakers, E.-J. (2018b). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*, *72*, 4–13.

Ly, A., Verhagen, A.J., Wagenmakers, E.-J. (2016a). Harold Jeffreys's default Bayes factor hypothesis tests: explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

Ly, A., Verhagen, A.J., Wagenmakers, E.-J. (2016b). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55.

Pisa, D., Alonso, R., Rábano, A., Rodal, I., Carrasco, L. (2015). Different brain regions are infected with fungi in Alzheimer's disease. *Scientific reports*, *5*, 15015.

Polya, G. (1954). *Mathematics and plausible reasoning: Vol. I. Induction and analogy in mathematics*. Princeton: Princeton University Press.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org. ISBN 3–900051–00–3.

Raftery, A.E., Madigan, D., Hoeting, J.A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*, 179–191.

Robert, C.P., Chopin, N., Rousseau, J. (2009). Harold Jeffreys's theory of probability revisited. *Statistical Science*, *24*, 141–172.

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

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

Scott, J.G., & Berger, J.O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*, 2587–2619.

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

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., Zwaan, R.A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on psychological science*, *11*, 917–928.

Wasserstein, R.L., & Lazar, N.A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*, 129–133.

Wasserstein, R.L., Schirm, A.L., Lazar, N.A. (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician*, *73*, 1–19.

Wrinch, D., & Jeffreys, H. (1919). On some aspects of the theory of probability. *Philosophical Magazine*, *38*, 715–731.

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

Wrinch, D., & Jeffreys, H. (1923). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *45*, 368–375.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F. (Eds.) *Bayesian Statistics: Proceedings of the 1st international meeting held in Valencia*, (Vol. 1 pp. 585–603). Berlin: Springer.