Three Insights from a Bayesian Interpretation of the One-Sided P Value

Maarten Marsman and Eric-Jan Wagenmakers
University of Amsterdam

Correspondence concerning this article should be addressed to:
Maarten Marsman or Eric-Jan Wagenmakers
University of Amsterdam, Department of Psychology
Weesperplein 4
1018 XA Amsterdam, The Netherlands
E-mail may be sent to M.Marsman@uva.nl or EJ.Wagenmakers@gmail.com.

Abstract

P values have been critiqued on several grounds but remain entrenched as the dominant inferential method in the empirical sciences. Here we elaborate on the fact that in many statistical models, the one-sided P value has a direct Bayesian interpretation as the approximate posterior mass for values lower than zero. The connection between the one-sided P value and posterior probability mass reveals three insights: (1) P values can be interpreted as Bayesian tests of direction, to be used only when the null hypothesis is known from the outset to be false; (2) as a measure of evidence, P values are biased against a point null hypothesis; (3) with N fixed and effect size variable, there is an approximately linear relation between P values and Bayesian point null hypothesis tests.

Keywords: Hypothesis testing, Bayesian inference, Null hypothesis, Estimation.

Across the empirical sciences –be it in medicine, biology, neuroscience, economics, sociology, or psychology– the classical P value is arguably the single most influential concept for statistical inference. Scientific claims about the presence of hypothesized effects are judged fit for publication only when the associated statistical tests yield $P < .05$, in which case researchers feel sanctioned to "reject the null hypothesis" and consequently embrace

This work was supported by the ERC grant "Bayes or Bust!" from the European Research Council.
Correspondence concerning this article may be addressed to Maarten Marsman or Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, the Netherlands. Email address: M.Marsman@uva.nl or EJ.Wagenmakers@gmail.com.

the alternative hypothesis. Despite its stranglehold on statistical reporting, however, the P value has been subject to intense scrutiny and numerous critiques; accessible overviews are provided by Johnson (1999), Morrison and Henkel (1970), Mulaik and Steiger (1997), Nickerson (2000), and Wagenmakers (2007).[1]

The P value detractors usually do not mince words. For instance, Edwards (1965, p. 400) argued that "Classical significance tests are violently biased against the null hypothesis.", Berger and Delampady (1987, p. 330) stated that "when testing precise hypotheses, formal use of P-values should be abandoned. Almost anything will give a better indication of the evidence provided by the data against H0.", Meehl (1978, p. 817) claimed that "the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology.", and Rozeboom (1997, p. 335) echoed this statement when he called P value significance testing "surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students".

Undeterred by such verbal onslaughts, some researchers believe that the critiques against P values are overstated or misplaced. For instance, Wainer (1999, p. 212) feels "a little at a loss to understand fully the vehemence and vindictiveness" of the P value critiques, Hagen (1997, p. 22) praises the logic of P value hypothesis testing, calling it "elegant" and "extraordinarily creative", and Leek and Peng (2015, p. 612) point out that "Arguing about the P value is like focusing on a single misspelling, rather than on the faulty logic of a sentence", and recommend that statisticians "need to stop arguing about P values".

In this article we continue to argue over P values. We depart by outlining a well-known Bayesian interpretation of the one-sided P value, and then sketch three immediate consequences. By doing so we hope to increase the field's awareness of what P values are and what they are not (Schervish, 1996).

Point of Departure: A Bayesian Interpretation of the One-Sided P Value

The Bayesian interpretation of the one-sided P value has a long and ongoing history (e.g., Berger & Mortera, 1999, Table 3; Casella & Berger, 1987; Greenland & Poole, 2013; Jeffreys, 1961; Lee, 2012, pp. 143–145; Lindley, 1965; Marin & Robert, 2007, p. 33; Morey & Wagenmakers, 2014; Pratt, 1965; Pratt, Raiffa, & Schlaifer, 1995; Rouanet, 1996). The main result may be summarized as follows. Consider Bayesian parameter estimation for the location parameter μ in a statistical model from the exponential family, assume the prior on μ is uniform on the real line, and denote the observed data by y. Then the proportion of the posterior distribution with mass lower than zero equals the one-sided classical P value, that is (e.g., Lindley, 1965; Pratt et al., 1995, p. 533),

$$\int_{-\infty}^{0} p(\mu \mid y)\mathrm{d}\mu = P_1. \tag{1}$$

[1] A selective and 14-year old listing of over 400 articles arguing against the use of p-values is available at `http://warnercnr.colostate.edu/~anderson/thompson1.html`.

Thus, for the classical statistician the one-sided P value represents the outcome of a significance test that assumes the null hypothesis is true, whereas for the Bayesian statistician the one-sided P value can be obtained from an estimation procedure (i.e., posterior updating of μ) that assumes the null hypothesis is false.

Furthermore, in this specific case the Bayesian estimation outcome is directly related to a Bayesian test for direction, one in which we contrast $H_+ : \mu > 0$ (i.e., the effect is positive) against $H_- : \mu < 0$ (i.e., the effect is negative). When the prior is symmetric around $\mu = 0$, the Bayes factor hypothesis test (Jeffreys, 1961; Kass & Raftery, 1995; Ly, Verhagen, & Wagenmakers, in press) simplifies to

$$\begin{aligned} \text{BF}_{+-} &= \frac{p(y \mid H_+)}{p(y \mid H_-)} \\ &= \frac{\int_0^\infty p(\mu \mid y)\text{d}\mu}{\int_{-\infty}^0 p(\mu \mid y)\text{d}\mu} \\ &= \frac{1 - P_1}{P_1}, \end{aligned} \tag{2}$$

where P_1 denotes the classical one-sided P value. Hence, there is a direct and exact relation between the Bayes factor for a test of direction and the one-sided P value such that $\log(\text{BF}_{-+}) = \text{logit}(P_1)$.

As mentioned above, the relationship is exact for location parameters in models from the exponential family when these parameters are assigned uniform priors; for other parameters and prior distributions the relationship is approximate (e.g., Casella & Berger, 1987; Greenland & Poole, 2013; for a critique see Gelman, 2013). In what follows we explore three consequences and insights afforded by the Bayesian interpretation of the one-sided P value.

First Consequence: P Values Are Meaningful Only When the Null Hypothesis is False

The Bayesian interpretation of the one-sided P value is that it is a test for direction, as the logit of the one-sided P value equals the log of the Bayes factor that contrasts $H_+ : \mu > 0$ (i.e., the effect is positive) against $H_- : \mu < 0$ (i.e., the effect is negative). Consequently, from this Bayesian perspective, the one-sided P value is appropriate only when $H_0 : \mu = 0$ is known from the outset to be false or uninteresting (Jeffreys, 1961, p. 387; but see DeGroot, 1973).

The interpretation of a one-sided P value as a test for direction –not as a test for the null hypothesis– is relevant because a common critique against the use of P values is that null hypothesis is nearly always false. For instance, Johnson (1999, p. 764) complains "P is calculated under the assumption that the null hypothesis is true. Most null hypotheses tested, however, state that some parameter equals zero, or that some set of parameters are all equal. These hypotheses, called point null hypotheses, are almost invariably known to be false before any data are collected". The same sentiment was expressed by Cohen (1990, p. 1308): "A little thought reveals a fact widely understood among statisticians: The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world. It can only be true in the bowels of a computer

processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null hypothesis is always false, what's the big deal about rejecting it?"

From a Bayesian perspective, however, the one-sided P value is not a test that involves the null hypothesis at all – instead, it is a test for the direction of an effect, suitable exactly for those scenarios where Johnson (1999) and Cohen (1990) argued it is meaningless. Note that in the Bayesian interpretation, collecting a large enough sample does not confirm the obvious; instead, what will be confirmed is the true direction of the effect. Paradoxically, the threat to the validity of the Bayesian interpretation of the one-sided P value is not that the null hypothesis is false, but that the null hypothesis is true. For when the null is exactly true, the test is between two directional models that are both equally wrong: the truth is literally in the middle (see also Sanborn & Hills, 2014; but see Rouder, 2014).

In sum, from a Bayesian perspective the one-sided P value represents a test for direction, a test that is valid only when the null hypothesis is false. For readers familiar with the popular argument against P values (i.e., "the null is never true") this line of argumentation may come as a surprise.

Second Consequence: P Values Are Biased Against $\mathcal{H}_0$

As alluded to earlier, several statisticians have remarked that P values overestimate the evidence against a point null hypothesis (e.g., Berger & Delampady, 1987; Dickey, 1977; Edwards, Lindman, & Savage, 1963; Johnson, 2013; Sellke, Bayarri, & Berger, 2001). The relation expressed in Equation 2 allows us to bypass mathematical details and present an intuitive argument: the one-sided P value corresponds to a Bayesian test for direction, in which $\mathcal{H}_+$ is pitted against $\mathcal{H}_-$; for the same data, such a test for direction generally yields a more diagnostic outcome than a test for existence, for instance, one that compares $\mathcal{H}_1$ (i.e., "there is an effect") against $\mathcal{H}_0$ (i.e., "there is no effect"). The reason why tests for direction are relatively diagnostic is because the models involved make opposite predictions: under one model the effect is predicted to be negative, whereas under the other model the effect is predicted to be positive. In contrast, for a test of existence, $\mathcal{H}_0$ is often a reduced case of $\mathcal{H}_1$, which means that the models can make similar predictions.

For example, consider a match between two avid Rummikub players. After six games, player A is leading player B by 4-2. If the choice is between $\mathcal{H}_+$: "player A is better than player B" versus $\mathcal{H}_-$: "player B is better than player A", one might have a strong intuitive preference in favor of $\mathcal{H}_+$: after all, player B is unlikely to be losing by 4-2 when she is in reality the better player. However, if the choice is between $\mathcal{H}_1$: "player A and player B are not equally good" versus $\mathcal{H}_0$: "player A and player B are equally good", one's preference is certainly less pronounced: a score of 4-2 is not that unlikely to occur when the players are equally skilled.

In sum, tests for direction are easier than tests for existence: when applied to the same data, tests for direction are more diagnostic than tests for existence. From a Bayesian perspective, the one-sided P value is a test for direction; when this test is misinterpreted as a test for existence –as classical statisticians are wont to do– this overstates the true evidence that the data provide against a point null hypothesis.

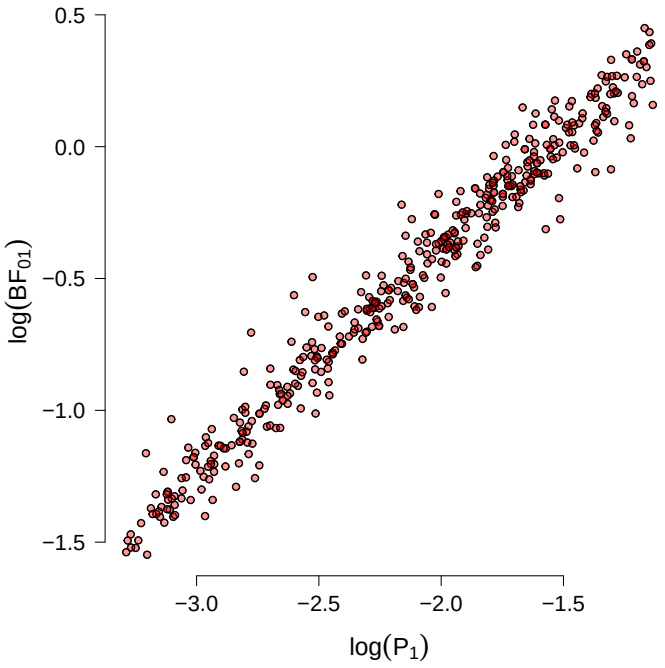

Figure 1. The highly regular relationship between one-sided P values and point null Bayes factor hypothesis tests for 440 t-test results reported by Wetzels et al. (2011) and reanalyzed by Rouder et al. (2012).

Third Consequence: With N Fixed, the Relation between P Values and Bayesian Point Null Hypothesis Tests is Approximately Linear

Several authors have explored the lawlike relationship between the classical P value and the Bayes factor against a point null hypothesis (e.g., Rouder, Morey, Speckman, & Province, 2012; Wetzels et al., 2011). Specifically, when sample size N is relatively stable and only effect size varies, lower P values will be accompanied by higher Bayes factors against the point null hypothesis. Figure 1 shows the empirical relation for 440 t-tests reported by Wetzels et al. (2011) and reanalyzed by Rouder et al. (2012).

We now formalize the relation between P values and Bayes factors for point null hypotheses by exploiting two facts. The first fact is that the one-sided P value is the posterior mass to the left of zero (i.e., Equation 1). The second fact is that the Bayes factor hypothesis test for a point null hypothesis $\mathcal{H}_0$ versus an unrestricted alternative $\mathcal{H}_1$ is given by the Savage-Dickey density ratio (e.g., Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Wetzels, Grasman, & Wagenmakers, 2010):

$$\mathrm{BF}_{01} = \frac{p(y \mid \mathcal{H}_0)}{p(y \mid \mathcal{H}_1)} = \frac{p(\mu = 0 \mid y, \mathcal{H}_1)}{p(\mu = 0 \mid \mathcal{H}_1)}. \tag{3}$$

In words, the Bayes factor in favor of the null hypothesis $\mathcal{H}_0$ equals the ratio of the posterior ordinate to the prior ordinate, evaluated under the alternative hypothesis $\mathcal{H}_1$ and for the

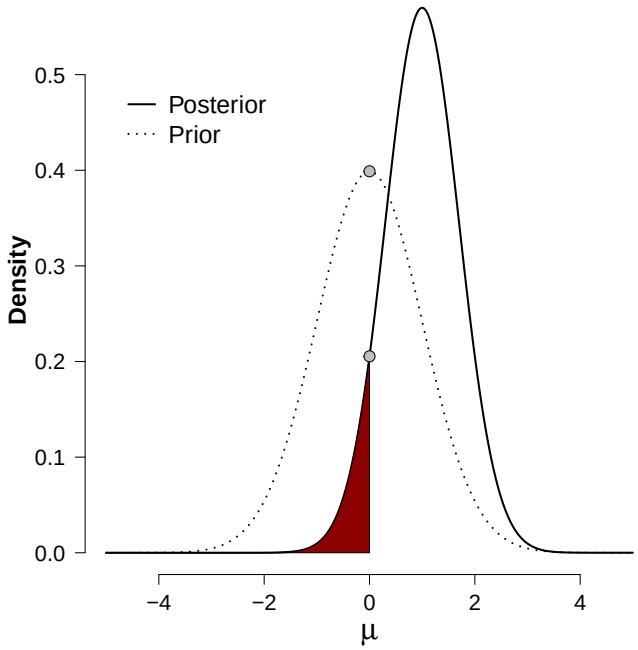

Figure 2. Prior and posterior distribution for a hypothetical data set. The shaded area of the posterior distribution indicates the mass that is lower than zero, whereas the two dots visualize the Savage-Dickey density ratio. As the posterior distribution shifts to the right, the shaded area and the posterior ordinate at $\mu = 0$ decrease simultaneously.

point of interest (here $\mu = 0$; for a short proof see O'Hagan & Forster, 2004, pp. 174–177).

We examine the following simplified scenario. The prior for the location parameter μ is a standard normal under $\mathcal{H}_1$: $p(\mu \mid \mathcal{H}_1) = \mathcal{N}(0, 1)$. Data points y_i, $i = 1, \ldots, N$, arrive and yield a posterior for μ that is a normal distribution with mean $m_\mu = \frac{N}{N+1}\bar{y}$ and variance $s^2_\mu = \frac{1}{N+1}$: $p(\mu \mid y, \mathcal{H}_1) = \mathcal{N}(m_\mu, s^2_\mu)$. We investigate the case where sample size N is fixed but $\bar{y}$ varies, that is, we keep s^2_μ constant but vary m_μ such that the posterior distribution is shifted to the right. Figure 2 shows the prior distribution and one example of a posterior distribution. The shaded area indicates $p(\mu < 0 \mid y, \mathcal{H}_1)$, the posterior mass lower than zero, and it is approximately equal to the one-sided P value; the ratio between the posterior and prior ordinate at $\mu = 0$ equals BF_{01}, the Bayes factor for the point null hypothesis (i.e., Equation 3). When the posterior distribution is shifted to the right this will simultaneously decrease both $p(\mu < 0 \mid y, \mathcal{H}_1)$ and BF_{01}.

The nature of these simultaneous changes is shown in Figure 3 for values of $P_1 \leq 0.05$ and $N = 10$. The left panel of Figure 3 shows the relation between the Bayes factor for the point null hypothesis and the posterior mass lower than zero on the untransformed scale, and the right panel shows the same relation on the log-scale. Comparison against the straight grey line segments confirms that the relation on the log-scale is approximately linear.

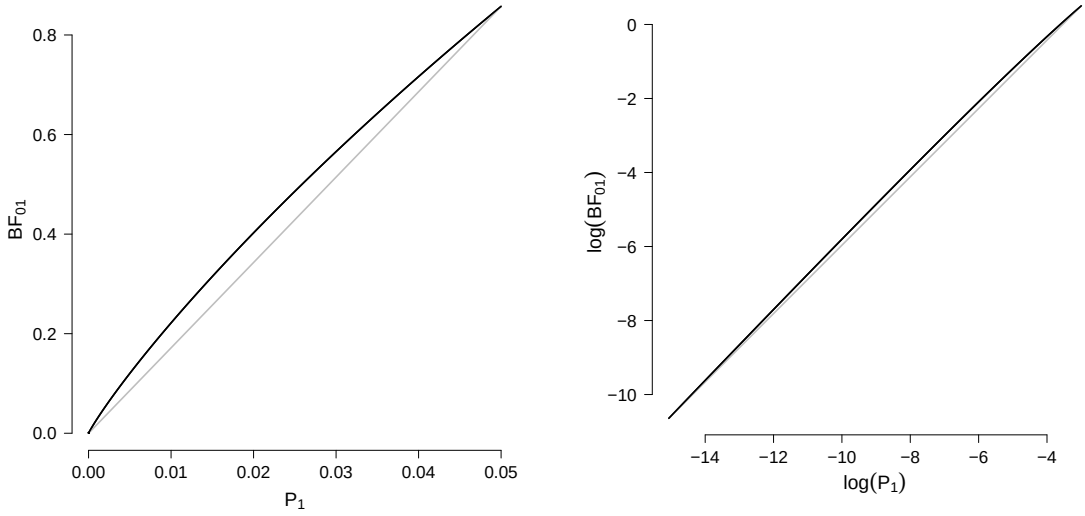

Figure 3. Lawlike relation between the one-sided P value and the point null Bayes factor BF_{01} for values of $P_1 < .05$ and $N = 10$. The left panel shows the relation on the untransformed scale, and the right panel shows the relation after a logaritmic transformation. In gray are straight line segments that connect the endpoints of the scale.

In this demonstration, the lower end-point corresponds to a value of m_μ for which the point of interest ($\mu = 0$) is five standard deviations away from the mean (i.e., the five σ rule commonly used in physics), whereas the upper end-point corresponds to a value of m_μ for which $P_1 = .05$ (i.e., the threshold level of significance used in most scientific disciplines).

An interesting observation about the relations shown in Figure 3 is that they are invariant across different choices of N and the choice of prior variance for the location parameter μ. That is, we can define a prior $p(\mu \mid \mathcal{H}_1) = \mathcal{N}(0, \tau^2)$ for the location parameter μ with $\tau^2 \neq 1$, or we can use a different value for N, and –except for a change of values on the y-axis– the same two panels would result. This means that the lawlike relation between the approximate one-sided P value and the Bayes factor is relatively general.

In sum, for a fixed value of N there exists a lawlike relation between the (approximate) one-sided P value and the Bayes factor for a point-null hypothesis. This relation implies that one can traverse from the one-sided P value to the Bayes factor and vice versa. Assuming that the relation between $\log(P_1)$ and $\log(\text{BF}_{01})$ is $\log(P_1) \approx \alpha + \beta \log(\text{BF}_{01})$, then we can compute P_1 as $\exp(\alpha)\text{BF}_{01}^{\beta}$, approximately. This also implies that when two equal-N studies have been done yielding one-sided P values P_a and $P_b = 2P_a$, we have that $P_b = 2P_a \approx \exp(\alpha)2\text{BF}_{01}(a)^\beta \approx \exp(\alpha)\text{BF}_{01}(b)^\beta$, such that $\text{BF}_{01}(b) \approx \text{BF}_{01}(a)\sqrt[\beta]{2}$.

Concluding Comments

We have demonstrated that one-sided P values can be given a Bayesian interpretation as an approximate test of direction, that is, a test of whether a latent effect is negative or positive. From a Bayesian perspective, this means that P values may be used when the

null hypothesis is false or when its veracity is not at issue (and when a diffuse, symmetric prior on the location parameter is acceptable). When misinterpreted as tests of existence, P values overestimate the evidence against the null hypothesis, as a test for direction is generally easier than a test for existence. Finally, with N fixed and effect size variable, P values and point null Bayesian hypothesis tests are approximately linearly related on the log-scale. This latter finding may falsely suggest that tests for direction and tests for existence are closely related. Although we have demonstrated this to be the case for N fixed, the situation changes if N is variable (e.g., Cano, Carazo, & Salmerón, 2013; Girón, Martínez, Moreno, & Torres, 2006). With N variable, sharp conflicts between test of direction and tests of existence are unavoidable, a phenomenon known as Lindley's paradox (Lindley, 1957). Consider the scenario shown in Figure 2 and imagine that more data are collected, causing the posterior distribution to become more peaked. At the same time, imagine that the posterior mean moves toward zero such that the posterior area lower than zero remains constant; when this happens the posterior ordinate will increase and this strengthens the evidence in favor of the point null hypothesis. Thus, as N increases and the posterior area lower than zero remains constant, the evidence in favor of the point null hypothesis increases indefinitely. This means that in a test for direction, one may be relatively certain that the effect is positive rather than negative; for the same data, a test for existence may reveal that the null hypothesis is much stronger supported than the alternative hypothesis. Of course, the paradox seizes to feel like a paradox as soon as it is properly understood. In the foreword to his monograph *Theory of Probability*, Jeffreys already underscores the main point: "The most beneficial result that I can hope for as a consequence of this work is that more attention will be paid to the precise statement of the alternatives involved in the questions asked. It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude.".

The Bayesian interpretation of the one-sided P value presents a double-edged sword. On the one hand, researchers can feel more confident in their use of the one-sided P value; after all, it has a Bayesian interpretation and it is valid when the null hypothesis is false (and when a diffuse, symmetric prior on the location parameter is acceptable). On the other hand, it is clear that the Bayesian interpretation of the one-sided P value presents a test of direction, not a test of existence. Despite the fact that many statisticians and methodologists have argued that tests of direction are more meaningful than tests of existence, we are not convinced that their arguments resonate with medical researchers, geneticists, experimental psychologists, and researchers in similar fields where general laws and invariances are regularly tested by means of empirical investigations.

References

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Berger, J. O., & Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, *94*, 542–554.

Cano, J. A., Carazo, C., & Salmerón, C. (2013). Bayesian model selection approach to the one way analysis of variance under homoscedasticity. *Computational Statistics*, *28*, 919–931.

Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one–sided testing problem. *Journal of the American Statistical Association*, *82*, 106–111.

Cohen, J. (1990). Things I have learned (thus far). *American Psychologist*, *45*, 1304–1312.

DeGroot, M. H. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association*, *68*, 966–969.

Dickey, J. M. (1977). Is the tail area useful as an approximate Bayes factor? *Journal of the American Statistical Association*, *72*, 138–142.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, *63*, 400–402.

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

Gelman, A. (2013). p values and statistical practice. *Epidemiology*, *24*, 69–72.

Girón, F. J., Martínez, M. L., Moreno, E., & Torres, F. (2006). Objective testing procedures in linear models: Calibration of the p-values. *Scandinavian Journal of Statistics*, *33*, 765–784.

Greenland, S., & Poole, C. (2013). Living with P values: Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*, *24*, 62–68.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15–24.

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

Johnson, D. H. (1999). The insignificance of statistical significance testing. *The Journal of Wildlife Management*, *63*, 763–772.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 19313–19317.

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

Lee, P. M. (2012). *Bayesian statistics: An introduction* (4 ed.). Chichester, UK: Wiley.

Leek, J. T., & Peng, R. D. (2015). P values are just the tip of the iceberg. *Nature*, *520*, 612.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192.

Lindley, D. V. (1965). *Introduction to probability & statistics from a Bayesian viewpoint. Part 2. Inference.* Cambridge: Cambridge University Press.

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (in press). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology.*

Marin, J.-M., & Robert, C. P. (2007). *Bayesian core: A practical approach to computational Bayesian statistics.* New York: Springer.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Morey, R. D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics and Probability Letters*, *92*, 121–124.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy.* New Brunswick (N.J.): Transaction Publishers.

Mulaik, S., & Steiger, J. (1997). *What if there were no significance tests.* Mahwah, New Jersey: Erlbaum.

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.).* London: Arnold.

Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society B*, *27*, 169–203.

Pratt, J. W., Raiffa, H., & Schlaifer, R. (1995). *Introduction to statistical decision theory.* Cambridge, MA: MIT Press.

Rouanet, H. (1996). Bayesian methods for assessing importance of effects. *Psychological Bulletin*, *119*, 149-158.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.

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.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In S. Mulaik & J. Steiger (Eds.), *What if there were no significance tests* (pp. 335–392). Mahwah, New Jersey: Erlbaum.

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*, 283-300.

Schervish, M. J. (1996). P values: What they are and what they are not. *The American Statistician*, *50*, 203–206.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

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*, 158–189.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*, 212–213.

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

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298.