Journal of Mathematical Psychology 72 (2016) 1–5

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments"

Joris Mulder [a,*], Eric-Jan Wagenmakers [b]

[a] Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands
[b] Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

HIGHLIGHTS

- Bayes factors are increasingly being used by psychologists to test statistical hypotheses and substantive theories.
- Differences between Bayes factor tests and classical significance tests are highlighted.
- The statistical software packages that are currently available for computing Bayes factors are described.
- An overview is presented of new contributions about Bayes factor tests in psychological research as part of this special issue.

ARTICLE INFO

Article history:
Available online 17 February 2016

Keywords:
Bayes factors
p values
Psychology

ABSTRACT

In order to test their hypotheses, psychologists increasingly favor the *Bayes factor*, the standard Bayesian measure of relative evidence between two competing statistical models. The Bayes factor has an intuitive interpretation and allows a comparison between any two models, even models that are complex and nonnested. In this introduction to the special issue "Bayes factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments", we first highlight the basic properties of the Bayes factor, stressing its advantages over classical significance testing. Next, we briefly discuss statistical software packages that are useful for researchers who wish to make the transition from *p* values to Bayes factors. We end by providing an overview of the contributions to this special issue. The contributions fall in three partly overlapping categories: those that present new philosophical insights, those that provide methodological innovations, and those that demonstrate practical applications.

1. Introduction

Many empirical researchers seek to evaluate and test hypotheses by comparing theoretical predictions to observed data. The dominant statistical vehicle for this activity is null hypothesis significance testing using *p* values. Despite their popularity, the literature contains an intense and ongoing debate about the usefulness of *p* values for testing scientific expectations (e.g., Berger & Sellke, 1987; Cohen, 1994; Edwards, Lindman, & Savage, 1963; Hubbard & Armstrong, 2006; Wagenmakers, 2007; Wainer, 1999, among many others).

One important critique of *p* values is that they cannot be used to quantify evidence in favor of the null hypothesis; a *p* value can only be used to falsify that null hypothesis. This is a limitation for replication research (Wagenmakers, Verhagen, & Ly, in press), or when the null hypothesis reflects a surprising prediction from a substantive theory (Gallistel, 2009). When the *p* value is larger than the chosen significance level we enter a state of suspended disbelief: there are insufficient grounds to reject the null hypothesis but we cannot claim evidence in its favor. In other words, the *p* value does not allow one to discriminate absence of evidence (i.e., uninformative data) from evidence of absence (i.e., data supporting the null hypothesis; Dienes, 2014). Another important critique is that *p* values tend to overestimate the evidence against the null hypothesis (Berger & Delampady, 1987; Johnson, 2013; Sellke, Bayarri, & Berger, 2001). This critique is particularly relevant in light of the present discussion about the lack of reproducibility of key results in psychology (Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). A third critique is that *p* values are computed as integrals over the

* Corresponding author.
E-mail address: j.mulder3@uvt.nl (J. Mulder).

http://dx.doi.org/10.1016/j.jmp.2016.01.002
0022-2496/

sample space of more extreme outcomes, and therefore depend on the sampling plan (i.e., the intention with which the data are collected, Berger & Berry, 1988a,b). This is a serious practical limitation for research fields in which there is no known sampling plan and data become available over time, as is common in ecology, geophysics, and astronomy.

A final critique we mention here is that p values are limited regarding the types of hypotheses that can be tested. For example, p values cannot be used for testing two nonnested regression models, such as a model $\mathcal{M}_1$ with "gender" and "income" as explanatory variables versus a model $\mathcal{M}_2$ with "educational level" and "age" as explanatory variables. Furthermore, p values are of limited use for testing hypotheses with order constraints on the parameters of interest (Braeken, Mulder, & Wood, 2015). This is unfortunate because psychologists often use order constraints to formulate expectations. For example, a strong treatment is expected to have more effect than a mild treatment, and a mild treatment is expected to have more effect than a placebo treatment.

These and other considerations have stimulated statisticians and scientists to explore alternative methods for testing theories (e.g., Hoijtink, Klugkist, & Boelen, 2008; Mulder, Hoijtink, & Klugkist, 2010; Rouder, Morey, Speckman, & Province, 2012; Vanpaemel, 2010). Recently, there has been an increasing interest in the use of the Bayes factor, the standard Bayesian method for model selection and hypothesis testing (Jeffreys, 1961; Kass & Raftery, 1995; Lewis & Raftery, 1997; O'Hagan & Forster, 2004). As a result, Bayes factors have been effectively used for testing hypotheses in various subdisciplines of psychology, such as cognitive psychology (Cavagnaro & Davis-Stober, 2014; Massaro, Cohen, Campbell, & Rodriguez, 2001), experimental psychology (Kammers, Mulder, de Vignemont, & Dijkerman, 2009a), clinical psychology (van den Hout et al., 2012), and developmental psychology (van de Schoot et al., 2011).

The current special issue "Bayes factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments" brings together a series of papers about Bayes factor tests for psychological research. The papers can roughly be divided into three categories. The first category consists of papers that explore the philosophical foundations of the Bayes factor, such as its interpretation as statistical evidence (Morey, Romeijn, & Rouder, 2016) and the origin of default Bayes factors as proposed by Sir Harold Jeffreys (Jeffreys, 1961; Ly, Verhagen, & Wagenmakers, 2016b). In the second category papers present new statistical developments, such as hypothesis testing based on the odds of correct rejection of the null hypothesis to incorrect rejection (Bayarri, Benjamin, Berger, & Sellke, 2016) and Bayes factors for testing order constraints on correlations (Mulder, 2016). The third category presents new applications of Bayes factor tests, such as category learning (Vanpaemel, 2016), differential item functioning in educational assessment (Verhagen, Levy, Millsap, & Fox, 2016), and sport statistics (Wetzels et al., 2016).

Before discussing the contributions in this special issue in more detail we highlight some fundamental properties of Bayes factor tests and its relation to classical tests in Section 2. In Section 3 currently available statistical software packages are discussed that can be used for computing Bayes factors without needing to know all the details of statistical modeling. Finally, an overview is given of the contributions of this special issue, followed by some closing remarks.

2. Differences between Bayes factors and null hypothesis significance tests

The Bayes factor, originally advocated by Jeffreys (1961), aims to quantify the relative evidence that the data provide for two competing hypotheses. For instance, a Bayes factor of a null hypothesis $\mathcal{H}_0$ against an alternative $\mathcal{H}_1$ of $B_{01} = 10$ implies that the data are ten times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$. Bayes factors are computed by assessing the relative predictive adequacy of the hypotheses under consideration, as provided by the so-called *marginal likelihood* (Kass & Raftery, 1995; Morey et al., 2016).

The goal of a null hypothesis significance test (NHST) on the other hand is to determine whether there is enough evidence in the data to reject the null hypothesis, while controlling the probability of incorrectly rejecting the null (i.e., the type I error probability), using a significance level α. A NHST is constructed such that the probability of not rejecting an incorrect null (i.e., the type II error probability) is minimized, resulting in a test with maximal power. This methodology dates back to Neyman and Pearson (see Lehmann, 1959, for a classic reference on this paradigm). In practice a NHST is typically performed using the p value, which was originally proposed by Fisher. A p value smaller than α indicates there is enough evidence to reject the null and a p value larger than α indicates there is not enough evidence in the data to reject the null. This mechanism automatically implies that a NHST can only be used to falsify the null hypothesis; it cannot be used to quantify evidence in favor of the null, even when the sample size N is large and the p value is close to 1.

Another fundamental difference is the scale of the outcome of both tests. In a Bayes factor test the outcome is the relative evidence in the data for $\mathcal{H}_0$ against $\mathcal{H}_1$, which lies on a *continuous* scale from 0 (which implies infinitely more evidence for $\mathcal{H}_1$ against $\mathcal{H}_0$) to infinity ∞ (which implies infinitely more evidence for $\mathcal{H}_0$ against $\mathcal{H}_1$). Based on the outcome of the Bayes factor, researchers can judge for themselves whether the evidence is sufficiently compelling in the context of the research question at hand. It can also happen that both hypotheses predict the observed data about equally well, in which case the Bayes factor is approximately 1. On the other hand, the outcome of a NHST, as advocated by Neyman and Pearson, is a *dichotomous decision*: There is either enough evidence in the data to reject the null, i.e., the evidence against the null is "significant", or there is not enough evidence in the data to reject the null, i.e., the evidence against the null is "not significant".

For researchers who perform a NHST it may not be satisfactory that the outcome of the test is dichotomous because the decision is based on the significance level which is arbitrarily chosen. An undesirable consequence is that the paradigm of Neyman and Pearson, who advocated making a dichotomous decision about rejecting the null while controlling the type I and type II error probabilities, is sometimes mixed up with the paradigm of Fisher, who advocated interpreting the p value in a NHST as a continuous measure of evidence against the null while avoiding any clear formulation about the effect under the alternative. For instance researchers tend to interpret a p value in the range $0.05 < p < 0.10$, $0.01 < p < 0.05$, and $p < 0.01$ as "mildly significant", "significant", or "highly significant", respectively, while having the idea that the type I error probability is controlled. This practice however results in an inflation of the type I error probability because the significance level α is chosen after observing the data where α is specified as small as possible but still larger than the observed p value.

The cause of this mixup may be that on the one hand researchers want to include the alternative hypothesis in the testing procedure, for example via the type II error probability as advocated by Neyman and Pearson (but not by Fisher). On the other hand researchers want to interpret the evidence in the data on a continuous scale as advocated by Fisher (but not by Neyman and Pearson). In that sense one could argue that the Bayes factor test has the best of both worlds. First the Bayes factor quantifies the evidence in the data on a continuous scale and no dichotomous decision has to be made about which hypothesis to select based on an arbitrarily chosen cut-off value. Second this measure

of evidence is a *relative* measure which balances between the plausibility of the null hypothesis and the alternative hypothesis where the prior under the alternative formalizes the anticipated effect if the null is not true. A Bayes factor of, say, $B_{01} = 10$, simply implies that researchers need to adjust their prior beliefs about the relative plausibility of the competing hypotheses by a factor of 10. It is up to the scientific community to determine whether this evidence is sufficiently compelling to warrant publication, something that needs to be assessed in the context of the prior plausibility of the competing hypotheses (Dreber et al., in press; Jeffreys, 1935).

Finally we note another important difference between both approaches regarding consistency. Roughly speaking a statistical test is called consistent if the true hypothesis is always selected if the sample size is large enough. The Bayes factor test is generally consistent (e.g., O'Hagan, 1995). Thus as the sample size goes to infinity, the Bayes factor of $\mathcal{H}_0$ against $\mathcal{H}_1$ either goes to 0 (if $\mathcal{H}_1$ is true) or to ∞ (if $\mathcal{H}_0$ is true). In contrast, NHST is not consistent. When the null hypothesis is true we still have a probability of incorrectly rejecting the null hypothesis that is equal to the chosen α-level, even in the case of extremely large samples. The Bayes factor on the other hand always points toward the true hypothesis as long as the sample is large enough.

3. Statistical software for computing Bayes factors

Easy-to-use statistical software is crucial in order for researchers and practitioners to use Bayes factors in their own field. Currently several statistical packages are available that can be used for computing Bayes factors without needing to know all the intricate details of statistical modeling: JASP (Love et al., 0000), the BayesFactor package in R (Morey & Rouder, 2015), BIEMS (Mulder, Hoijtink, & de Leeuw, 2012), BIG (Gu, Mulder, Decović, & Hoijtink, 2014), and BOCOR (Mulder, 2016). All programs are freely downloadable and easy to use.

The first program, JASP, has a point-and-click graphical user interface (jasp-stats.org). JASP is a spreadsheet program that features both classical and Bayesian data analysis methods. Many of the statistical models used by social scientists (e.g., ANOVA, regression, repeated measures) are implemented in JASP, often by using the functionality of the BayesFactor package.

The second program is the BayesFactor package in R. This package provides much of the same functionality as JASP, and users who are comfortable with R may prefer to work with the BayesFactor package.

The third program, BIEMS, also produces Bayes factors for commonly used statistical models via a graphical user interface (jorismulder.com). BIEMS is particularly useful for testing hypotheses with order constraints (possibly in addition to equality constraints) between the parameters of interest, say, $\mu_1 > \mu_2 > \mu_3$, i.e., group mean 1 is expected to be larger than group mean 2, and group mean 2 is expected to be larger than group mean 3. The program has a point-and-click tool for formulating hypotheses with equality and/or order constraints in an easy manner.

Finally, BIG (informative-hypotheses.sites.uu.nl) and BOCOR (jorismulder.com) can be used for computing Bayes factors between hypotheses with only order constraints. BIG can be used to test constrained hypotheses in general statistical models, such as structure equation models, and BOCOR can be used to test constrained hypotheses on correlations coefficients.

The Bayes factors in the above software packages can be computed without needing to formulate prior distributions for the model parameters. The motivation was that users who are new to Bayesian statistics may find it difficult to translate one's prior beliefs into distributions. It may also be the case that prior information is simply unavailable. For this reason default priors can be used for computing Bayes factors in JASP, the BayesFactor package, BIEMS, BIG, and BOCOR. The default prior in JASP and the BayesFactor package builds on the work of Jeffreys (1961), Zellner and Siow (1980), Rouder, Morey, Speckman, and Province (2012b) and Rouder, Speckman, Sun, and Iverson (2009). The default prior in BIEMS builds on earlier work of Berger and Pericchi (1996) and Mulder et al. (2010, 2009), and it contains the information of a minimal experiment to ensure that the prior distribution is not unrealistically vague but also not too informative. In BIG and BOCOR very vague proper priors are specified, which is allowed when computing Bayes factors between hypotheses with only order constraints on the parameters of interest (Klugkist & Hoijtink, 2007; Mulder, 2014). Note that arbitrarily vague proper priors should not be used when testing hypotheses with strict equality constraints due to the Jeffreys–Bartlett–Lindley paradox (Bartlett, 1957; Jeffreys, 1961; Lindley, 1957; Ly, Verhagen, & Wagenmakers, 2016a).

4. Contributions of the special issue to Bayesian hypothesis testing in psychological research

As indicated above, Bayes factors avoid many of the limitations inherent to p value testing. The development of Bayes factors for testing statistical models and its application to evaluate scientific theories therefore remains an active area of research. This is witnessed by the many applications of Bayes factor tests in psychology (e.g. Cavagnaro & Davis-Stober, 2014; Kammers et al., 2009a; Massaro et al., 2001) and other fields of research such as genetics (Sawcer, 2010), ecology (King, Morgan, Gimenez, & Brooks, 2010), and management research (Andraszewicz et al., 2015; Braeken et al., 2015). The increasing interest in Bayes factor hypothesis testing motivated the current special issue "Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments" for the *Journal of Mathematical Psychology*. The contributions in this special issue aim to (i) provide new insights about the philosophical underpinnings of Bayes factors for testing statistical hypotheses, (ii) present methodological advancements of Bayes factor tests for yet unexplored testing problems, and (iii) show how Bayes factors can address research questions in various applications which could not be properly addressed using alternative approaches. The technical level of many contributions is relatively low so that most readers are able to understand the new insights and key results. The contributions can be divided into three partly overlapping categories outlined below.

Philosophical foundations

- In *The philosophy of Bayes factors and the quantification of statistical evidence*, Morey, Romeijn, and Rouder show how the Bayes factor formalizes the important scientific concept of statistical evidence. Furthermore the authors show how Bayes factors provide a natural means for updating one's prior beliefs about scientific claims, hypotheses, or theories in light of the newly observed data.
- In *Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology*, Ly, Verhagen, and Wagenmakers discuss in an accessible manner how Sir Harold Jeffreys, one of the most influential Bayesian statisticians, initiated the development of default Bayes factor for testing statistical hypotheses which can be used without having subjective prior information. Furthermore, these authors presented useful extensions of Jeffreys' methodology such as a one-sided hypothesis test for a bivariate correlation. Interesting response papers were provided about this discussion of Jeffreys' work by Robert, and Chandramouli and Shiffrin, which was followed by a rejoinder by Ly, Verhagen, and Wagenmakers.

- In *Bayes factors, relations to minimum description length, and overlapping model classes*, Shiffrin, Chandramouli, and Grünwald investigate the theoretical and practical differences between two prominent methods, the Bayes factor and minimum description length. Although both methods have different philosophical backgrounds, the authors show in a nontechnical manner that both methods behave similarly when testing one-sided hypotheses of the success probability in a binomial experiment.
- In *How Bayes factors change scientific practice*, Dienes shows how Bayes factors can help in solving several important problems underlying the credibility crisis which currently plagues psychology. These problems are partly caused by the misuse of classical p values in null hypothesis significance testing. Bayes factors potentially resolve these issues due to the fact that the Bayes factor is a symmetric measure of evidence between two hypotheses and the fact that Bayes factors are not sensitive to the stopping rule that is used by researchers when they collect data.

Methodological advancements

- In *Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses*, Bayarri, Benjamin, Berger, and Sellke present Bayesian as well as classical methods that avoid four common problems with standard statistical testing, such as the failure to incorporate power when quantifying statistical evidence. Furthermore the authors show that the Bayes factor satisfies important frequentist principles. This holds out hope for a possible synthesis of frequentist and Bayesian hypothesis testing, a subject which statisticians have been struggling with for decades.
- In *Bayes factors for testing order-constrained hypotheses on correlations*, Mulder presents a methodology for computing Bayes factors for testing order constraints on correlation coefficients. Applications include multitrait–multimethod analyses, repeated measures studies, and tests for ordered moderator effects. The methodology is implemented in the freely downloadable software package "BOCOR". This package allows researchers to test complex order constraints on correlations using the Bayes factor in a relatively easy manner.
- In *Bayes factors for state-trace analysis*, Davis-Stober, Morey, Gretton, and Heathcote generalize existing Bayes factor methodology for state-trace analysis. The authors' methodology efficiently assesses the evidence for a monotonic relation; in addition, the authors propose a group-level Bayes factor to evaluate whether or not all individuals satisfy monotonicity. Particular attention is paid to the specification of prior distributions in order to ensure that the statistical models under test are veridical reflections of the underlying theory.
- In *Error probabilities in default Bayesian hypothesis testing*, Gu, Hoijtink, and Mulder investigate classical error probabilities of commonly used default Bayes factors for a simple Student t test. This work was motivated by the fact that Bayes factors minimize the sum of the type I and type II error probabilities when generating data via the proper priors that are specified under the hypotheses. The authors show how the prior implicitly controls for which effects one obtains good frequentist performance.
- In *Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs*, Nathoo and Masson show how to compute the BIC (i.e., a large sample approximation of the Bayes factor) for testing hypotheses in repeated measures designs. An R-package is provided to compute this BIC which only needs standard output of a classical ANOVA analysis.
- In *Automatic Bayes factors for testing variances of two independent normal distributions*, Böing-Messing and Mulder propose different default Bayes factors for a multiple testing problem of two population variances. The proposed methods can be used to quantify the evidence in the data in favor of the null hypothesis that two population variances are equal, something that is not possible using classical p value tests. The authors show how to compute these Bayes factors in a simple manner.

New applications of Bayes factor tests

- In *Evaluating evidence for invariant items: A Bayes factor approach to testing measurement invariance*, Verhagen, Levy, Millsap, and Fox present a Bayes factor test for detecting differential item functioning in educational testing. The test is applied to a mathematical test to investigate whether women answer geometry items differently than men. An attractive feature of the proposed test is that it can be used to quantify evidence in favor of the hypothesis that an item is gender invariant.
- In *Prototypes, exemplars and the response scaling parameter: A Bayes factor perspective*, Vanpaemel shows that the Bayes factor can be used to break the stalemate between prototype and exemplar theorists in category learning. As described by the author this is due to the fact that Bayes factors behave like an Occam's razor where model fit and model complexity are naturally balanced when quantifying the relative evidence in the data between the two models.
- In *Bayesian analysis of simulation-based models*, Turner, Sederberg, and McClelland explore likelihood-free posterior estimation for the Leaky, Competing Accumulator (LCA) model and the Feed-Forward Inhibition (FFI) model. The authors then compare these two models on the basis of the several performance measures including Bayes factors based on the BIC. The results reveal considerable participant heterogeneity, where the LCA does better than the FFI for some participants, but worse for others.
- In *A Bayesian test for the hot hand phenomenon*, Wetzels, Dolan, Tutschkow, van der Sluis, Dutilh, and Wagenmakers present a Bayes factor test to determine whether the performance of sports players is punctuated by streaks of exceptional performance. The results indicate that very long data sequences are needed to determine whether the hot hand really exists or not. The new method is applied to empirical basketball data and time-series data of a visual perception task.
- In *Using Bayes factors to test the predictions of models: A case study in visual working memory*, Kary, Taylor, and Donkin use Bayes factors to quantify the relative predictive adequacy of two models for visual working memory, that is, hierarchical versions of standard slots and resource models. Data from previous experiments are used to update the prior distributions, resulting in more focused predictions. The authors rightly stress the difference between fitting a model to data and evaluating that model's predictions.

5. Final remarks

The Bayes factor is increasingly used across many fields of empirical research. Reasons for its increased popularity include its intuitive interpretation as the relative evidence provided by the data between the hypotheses of interest, its flexibility to test nonnested hypotheses (possibly in the presence of order constraints), and its straightforward availability through user-friendly software packages. Together, the contributions to this special issue form another step toward a better understanding of the Bayes factor's potential to address substantive research questions in an appropriate and coherent fashion. We also would

like to alert readers to a special issue on Bayesian methods in psychology which is scheduled to appear in *Psychonomic Bulletin and Review* later this year. These and other contributions demonstrate how psychology and other empirical disciplines can benefit from using Bayes factors to test statistical hypotheses and evaluate scientific theories.

Acknowledgment

This work was partly funded by a Veni Grant awarded to the first author by the Netherlands Organization for Scientific Research (VENI Grant Number 451-13-011).

References

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R. P. P. P., Verhagen, A. J., & Wagenmakers, E. J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, *41*, 521–543.

Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*, 533–534.

Bayarri, M., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, *72*, 90–103.

Berger, J. O., & Berry, D. A. (1988a). The relevance of stopping rules in statistical inference. In S. S. Gupta, & J. O. Berger (Eds.), *Statistical decision theory and related topics. Vol. 4* (pp. 29–72). New York: Springer Verlag.

Berger, J. O., & Berry, D. A. (1988b). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*, 159–165.

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

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*, 109–122.

Berger, J. O., & Sellke, T. (1987). Testing a point-null hypothesis: the irreconcilability of significance levels and evidence (with discussion). *Journal of the American Statistical Association*, *82*, 112–122.

Braeken, J., Mulder, J., & Wood, S. (2015). Relative effects at work: Bayes factors for order hypotheses. *Journal of Management*, *41*, 544–573.

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

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, *49*, 997–1003.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psycholology*, *5*, 781.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., et al. (2016). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences of the United States of America*, in press.

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

Gallistel, C. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.

Gu, X., Mulder, J., Decović, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*, 511–527.

Hoijtink, H., Klugkist, I., & Boelen, P. A. (2008). *Bayesian evaluation of informative hypotheses*. New York: Springer.

Hubbard, R., & Armstrong, J. S. (2006). Why we don't really know what statistical significance means: Implications for educators. *Journal of Marketing Education*, *28*, 114–120.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, *31*, 203–222.

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

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.

Kammers, M. P. M., Mulder, J., de Vignemont, F., & Dijkerman, H. C. (2009a). The weight of representing the body: Addressing the potentially indefinite number of body representations in healthy individuals. *Experimental Brain Research*, *204*, 333–342.

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

King, R., Morgan, B. J. T., Gimenez, O., & Brooks, S. P. (2010). *Bayesian analysis for population ecology*. Boca Raton, FL: Chapman & Hall/CRC.

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

Lehmann, E. L. (1959). *Testing statistical hypotheses* (1st ed.). New York: Wiley.

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, *92*, 648–655.

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

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., & Verhagen, A.J. et al. (0000). Jasp (version 0.7)[computer software].

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

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

Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review*, *8*, 1–17.

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18.

Morey, R.D., & Rouder, J.N. (2015). BayesFactor 0.9.11-1. Comprehensive R Archive Network. Retrieved from http://cran.r-project.org/web/packages/BayesFactor/index.html.

Mulder, J. (2014). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, *67*, 153–171.

Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104–115.

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). Biems: A fortran 90 program for calculating Bayes factors for inequality and equality constrained model. *Journal of Statistical Software*, *46*.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*, 887–906.

Mulder, J., Klugkist, I., van de Schoot, A., Meeus, W., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*, 530–546.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society: Series B*, *57*, 99–138.

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

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012a). Default Bayes factors for anova designs. *Journal of Mathematical Psychology*,.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012b). 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.

Sawcer, S. (2010). Bayes factors in complex genetics. *European Journal of Human Genetics*, *18*, 746–750.

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

van de Schoot, R., Hoijtink, H., Mulder, J., van Aken, M. A. G., Orobio de Castro, B., Romeijn, J. W., et al. (2011). Evaluating expectations about negative emotional states of aggressive boys using Bayesian model selection. *Developmental Psychology*, *47*, 203–212.

van den Hout, M. A., Rijkeboer, M. M., Engelhard, I. M., Klugkist, I., Hornsveld, H., Toffolo, M., et al. (2012). Tones inferior to eye movements in the EMDR treatment of PTSD. *Behaviour Research and Therapy*, *50*, 275–279.

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

Vanpaemel, W. (2016). Prototypes, exemplars and the response scaling parameter: A Bayes factor perspective. *Journal of Mathematical Psychology*, *72*, 183–190.

Verhagen, J., Levy, R., Millsap, R. E., & Fox, J. P. (2016). Evaluating evidence for invariant items: A Bayes factor approach to testing measurement invariance. *Journal of Mathematical Psychology*, *72*, 171–182.

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

Wagenmakers, E. J., Verhagen, A. J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, in press.

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

Wetzels, R., Dolan, C., Tutschkow, D., van der Sluis, S., Dutilh, G., & Wagenmakers, E. J. (2016). A Bayesian test for the hot hand phenomenon. *Journal of Mathematical Psychology*, *72*, 200–209.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: proceedings of the first international meeting* (pp. 585–603). Valencia: University of Valencia Press.