UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Wong, T. K., Kiers, H., & Tendeiro, J. (2022). On the Potential Mismatch Between the Function of the Bayes Factor and Researchers' Expectations. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.36357

Methodology and Research Practice

On the Potential Mismatch Between the Function of the Bayes Factor and Researchers' Expectations

Tsz Keung Wong[1, a], Henk Kiers[1], Jorge Tendeiro[2]

[1] Department of Psychology, University of Groningen, Groningen, The Netherlands, [2] Hiroshima University, Hiroshima, Japan

Keywords: Bayes Factor, Questionable Reporting and Interpreting Practice, Null Hypothesis Bayesian Testing, Statistical test

https://doi.org/10.1525/collabra.36357

Collabra: Psychology

Vol. 8, Issue 1, 2022

The aim of this study is to investigate whether there is a potential mismatch between the usability of a statistical tool and psychology researchers' expectation of it. Bayesian statistics is often promoted as an ideal substitute for frequentists statistics since it coincides better with researchers' expectations and needs. A particular incidence of this is the proposal of replacing Null Hypothesis Significance Testing (NHST) by Null Hypothesis Bayesian Testing (NHBT) using the Bayes factor. In this paper, it is studied to what extent the usability and expectations of NHBT match well. First, a study of the reporting practices in 73 psychological publications was carried out. It was found that eight Questionable Reporting and Interpreting Practices (QRIPs) occur more than once among the practitioners when doing NHBT. Specifically, our analysis provides insight into possible mismatches and their occurrence frequencies. A follow-up survey study has been conducted to assess such mismatches. The sample (N = 108) consisted of psychology researchers, experts in methodology (and/or statistics), and applied researchers in fields other than psychology. The data show that discrepancies exist among the participants. Interpreting the Bayes Factor as posterior odds and not acknowledging the notion of relative evidence in the Bayes Factor are arguably the most concerning ones. The results of the paper suggest that a shift of statistical paradigm cannot solve the problem of misinterpretation altogether if the users are not well acquainted with the tools.

Several studies which focus on the researchers' understanding of p-values and confidence intervals demonstrate the prevalence of misinterpretation of these common statistical tools (Haller & Krauss, 2002; Hoekstra et al., 2014; Morey, Hoekstra, Rouder, & Wagenmakers, 2016; Oakes, 1986). Moreover, two recent large-scale survey studies showed that statistical misinterpretation is pervasive across different fields (X.-K. Lyu et al., 2020) and still prevails in the field of psychology (Z. Lyu et al., 2018). Haucke et al. (2020), Hoekstra et al. (2012), and Fidler and Loftus (2009) also demonstrated that there was a clear difference between what researchers can conclude versus what they want to conclude from frequentist statistical results in their experiment. This current prevalence of misunderstanding of frequentist statistical tools might reveal an underlying and potential mismatch between what psychology researchers can achieve and what they would like to do with these tools.

Meanwhile, some researchers point out that Bayesian statistics and thinking are more in line with researchers' interests than frequentist statistics (Assaf & Tsionas, 2018; Cohen, 1994; Dienes, 2011; Haller & Krauss, 2002; Haucke et al., 2020; Hoekstra et al., 2014; Iverson et al., 2009; Morey, Hoekstra, Rouder, Lee, et al., 2016; Vandekerckhove et al., 2018; Wagenmakers et al., 2018). The reason is that Bayes' rule allows computing the so-called 'inverse probability', that is, the probability of a hypothesis being true given the observed data. The inverse probability thus clearly contrasts with the p-value, which gives the probability of observing the data (or more extreme data) given a null hypothesis is true. Various researchers have argued that the inverse probability is more aligned with researchers' interests. For this reason, it is often concluded that Bayesian analysis is better fitted to the needs of researchers than frequentist analysis, and therefore Bayesian inference has been advocated as a replacement of frequentist inference (Wagenmakers, 2007). Furthermore, Null Hypothesis Bayesian Testing (NHBT), which employs Bayes factors (BFs) comparing a null hypothesis to an alternative hypothesis, is seen as an improvement of NHST as it allows the user to quantify relative evidence for either of the two competing hypotheses (Wagenmakers et al., 2018). In particular, the direct possibility of quantifying evidence for the null hypothesis relative to the alternative hypothesis is often viewed as a clear practical example. Also, the fact that

a Correspondence concerning this article should be addressed to Tsz Keung Wong, Department of Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS, Groningen, the Netherlands. Email: t.k.wong3004@gmail.com

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

the BF in NHBT appears not to be affected by optional stopping of data collection, is often seen as an advantage (e.g. see Rouder, 2014; Tendeiro et al., 2021).

NHBT tests a precise point-null hypothesis against a general composite alternative hypothesis (Ly et al., 2016; Robert et al., 2009). The alternative hypothesis is the aggregate of all the other possible values of the parameter, combined with a specific prior probability function (or "within-model prior probability distribution"). The BF test statistic, here denoted BF_{01}, is the marginal likelihood ratio of the null hypothesis (H_0) over the alternative hypothesis (H_1). It measures the data evidence for H_0 relative to H_1 or specifically, how many times more likely it is to find the observed data if H_0 is true than if H_1 is true.

The interpretation of the BF is twofold (Tendeiro & Kiers, 2019). Firstly, the BF is the likelihood ratio of the observed data given two competing hypotheses. For instance, a BF_{01} of 2 means that the data are twice more likely to be observed under the null than under the alternative. Secondly, the BF is the change factor in going from prior odds (i.e., the ratio of prior probabilities of the two hypotheses) to posterior odds (i.e., the ratio of posterior probabilities of the two hypotheses given the data). Specifically, the posterior odds equal the prior odds multiplied by the BF. In formula:

$$\mathrm{PostOdds}_{01} = \mathrm{BF}_{01} \times \mathrm{PriorOdds}_{01} \quad (1)$$

where the indices "01" indicate that the odds and the BF are given for H_0 versus H_1.

Hence the BF indicates the shift of our prior belief on the relative plausibility of two competing hypotheses to what our belief should be now we have observed the data. A BF_{01} of 2 suggests that the relative degree of belief between the two hypotheses should shift (by a factor 2) towards the null relative to the alternative, after observing the data. Only in the special case where the prior odds equal 1, the BF equals the posterior odds.

Currently, the use of Bayesian analysis and in particular NHBT is getting more frequent and common (Tendeiro & Kiers, 2019, p. 774). Although much existing literature describes why Bayesian testing is better than frequentist testing in terms of the usability of these statistical tools, quite a bit of criticism has appeared on the approach as well (e.g. Gelman & Shalizi, 2012; Robert, 2016) and possible problems with it should not be overlooked. As indicated for instance by Tendeiro and Kiers (2019), NHBT might be prone to misinterpretations, and like NHST, need not match perfectly with the researchers' questions either. Because of such potential misinterpretations and misunderstandings, practitioners are advised to have a sufficient and basic understanding of these tools before using them. Otherwise, the poor use of Bayesian methods could lead to a set of problems of its own.

In the light of the often-heard advice to use Bayesian analysis by shifting from the use of p-values to BFs, in-depth scrutiny of the latter is needed to investigate whether it is indeed well matched with researchers' interests. This leads us to the main research question: "Is there a mismatch between what psychology researchers can conclude and would like to achieve with NHBT?"

The present paper consists of two parts. First, a literature review of published applied studies in which researchers employed NHBT is conducted for exploring the potential existence of misinterpretations. We highlight what researchers seemingly would like to achieve through NHBT and what may be common misinterpretations. Secondly, a survey study on the interpretation of the BF associated with NHBT among actual researchers is conducted to assess the potential mismatch mentioned above. The format and the style of the survey are analogous to previous survey studies on the use of p-values and CIs (Haller & Krauss, 2002; Hoekstra et al., 2014).

Study 1

Method

Selecting the set of papers to study

The targeted publications are empirical psychological research papers that employed Jeffreys' BF-based null hypothesis testing in their empirical research. Google Scholar offers a straightforward way to broadly search for scholarly literature. Therefore, a search on Google Scholar using the search terms ["psychology", "Bayes factor" and "Bayesian test"] was conducted. More specifically, the terms were placed under the search bar *with all of the words* in the advanced search menu. Besides, it was chosen that the publications should be dated before and including 2020. We did this on 2020/11/16 and obtained 408 references. The use of keywords "psychology" and "Bayes Factor" corresponds to our research interest. The phrase "Bayesian test" was added for two reasons. First, it helped at purifying the returned references to match our research interest (i.e., psychological publications using NHBT). Second, it reduced the number of returned references from around 13,000 to 400, which made the investigation accomplishable. Despite the arguably small sample size, our search suffices to answer our research question on exploring the potential mismatch and the different possible misinterpretation.

Despite our keywords, many results from the search did not lead to targeted papers. For this reason, an exclusion and inclusion procedure was employed. That is, the search led to a large number of methodological rather than empirical papers, and also to non-psychological papers. Of these 408 papers, 228 were considered to be methodological papers as judged from their abstracts, titles, keywords, and journals (e.g., method development and statistical tutorials), and 39 were considered to be non-psychological papers (as was concluded from the fact that their journals were not covered by APA PsycInfo). Furthermore, 23 studies mainly employed Bayesian information criteria, which do not match our interest in the BF for use in NHBT and were therefore excluded. Of the remaining references, the following were excluded for various reasons: 13 papers because they were not in English, 8 publications were PhD theses (and hence it would complicate the procedure concerning the fact that they contain multiple studies including methodological papers, non-psychological papers, and duplications of published papers), 1 master thesis, 6 duplications, 3 preregistered plans of research, 1 appendix, 2 in press, 5 preprints, 1 paper without employing any Bayesian statistical tool, 1 irrelevant result (library declaration and deposit agreement) and 2 other studies which were about

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

criminology and accounting. Preprints and in press papers were excluded since these papers are still subject to change. One paper not covered by APA PsycInfo was included nevertheless since it was published in the journal "Comprehensive Results in Social Psychology", which indicates it is a psychological study. One paper that used the replication BF instead of a BF for null hypothesis testing was excluded. One paper that mainly used BFs for finding the best out of 12 models rather than testing an alternative hypothesis against the null hypothesis was excluded as well. Also, any reported replication BF in other papers was either excluded or ignored. Lastly, one reference is lost due to a downloading error. After these exclusions and inclusions, we ended up with 73 publications fitting our in- and exclusion criteria.

Material

A recording sheet was created for marking researchers' reporting and interpreting behaviour concerning the use of the BF. This was based on a pilot study in which ten references were randomly selected from the 73 targeted papers. These papers were analysed and evaluated in detail. The results of this pilot analysis suggested to us that there were at least 8 recurrent Questionable Reporting or Interpreting Practices (QRIPs) in which the researchers gave incomplete, and (possibly) misleading statements for interpreting the BFs. These were:

1. *Explicitly describing, defining, or elaborating BFs as posterior odds ratios:* Mentioning that the BF is a posterior odds ratio of two competing hypotheses given the data, which as can be seen from formula (1), is only true if the prior odds equal 1.
2. *Null and alternative hypotheses were not specified:* In particular, it is not mentioned whether a one-sided or two-sided test is used, and/or the statistical representation of the hypothesis (such as $H_0 = 0$, $H_A > 0$) is missing.
3. *Simplifying the alternative hypothesis by ignoring the use of its associated prior distribution for computing a marginal likelihood*[1]*:* The reason or justification for the chosen within-model prior is not provided or it is not even mentioned what prior distributions were used with the alternative hypothesis.
4. *Presenting BFs as evidence for one hypothesis without mentioning it is relative to another competing hypothesis:* It is not mentioned that the reported strength of the evidence for one hypothesis depends on what the other hypothesis is.
5. *Making an absolute statement concerning the truthfulness of a hypothesis by using BFs:* It is simply mentioned that there *is* an effect or that there *is no* effect, while in fact, on the basis of statistical measures one cannot confirm or disconfirm a hypothesis; the best one could say is that the BF gives evidence more in favor of one hypothesis than another.
6. *Using BFs as if they are posterior odds ratios*[2]*:* Based on an observed BF it is concluded that either of the two hypotheses is more probable or likely. This would be true if the BF gives the posterior odds ratio, but as can be seen from formula (1) it does not. In fact, BFs can only tell under which hypothesis the *observed data* is more likely to occur (see p. 2).
7. *Considering BFs as measure of effect size:* ascribing a small BF to a small effect size, or a big BF to a big effect size.
8. *Mismatch between the research hypothesis and the statistical hypothesis:* the two used statistical models do not coincide with the researcher's hypothesis. For instance, one might have a one-sided substantive prediction (e.g., Drug A has positive effect) for the alternative hypothesis, but the associated statistical model is two-sided.

For each article, it was recorded whether each QRIP was present or absent. For each identified QRIP, the associated statement in the paper was also recorded for possible reevaluation purposes. In addition, any further Questionable Reporting or Interpreting Practice other than the 8 QRIPs above was recorded.

Researchers' reporting practice might differ from each other in other respects as well. Some might neglect to report effect sizes, and some might ignore the importance of mentioning the prior model for the BF. In any case, drawing a conclusion solely on the basis of a single BF to us seems unideal, because the same BF value may refer to quite different situations with respect to uncertainty and effect size. For example, the single sample JZS based test in the BF[3] (Morey & Rouder, 2018) R routine gives $BF_{01} = 4$ both for d = .14 and n = 34, and for d = .07 and n = 738, which clearly differ strongly in effect size and sample size. Here, the BF combines different outcomes into the same degree of evidence, but in the former case the evidence for the null is mainly due to the large uncertainty, while in the second case it is mainly due to the small effect size. People can see a bigger picture if more information is given. To see to what extent other information is actually given in practice, an additional assessment of the reporting practice was carried out by recording the main reported elements in the

1 Publications which discussed or mentioned the sensitivity of the BF to the choice of prior distribution are exempted from this QRIP, since they indicate that the authors were aware of the influence of within-model priors on the BF. Additionally, the authors raised the awareness of the readers about this notion of BF.

2 All publications with the specification of any prior odds ratio are immune to this QRIP since these authors used more than the BF to interpret their result and draw their conclusion. It has been verified that in the cases we signaled as occurrence of QRIP 6, no prior odds ratio had been specified in the paper.

3 This test compares H_0: d =0 to the alternative H_1: d ≠ 0 for which the so-called Jeffreys-Zellner-Siow (JZS) priors have been used, with scale factor $\sqrt{2}/2$; it is the default in the routine BayesFactor, and is used here only for illustrative purposes.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

Table 1. Frequencies and percentages (in parentheses) of papers in which the observed Questionable Interpreting and Reporting Practice (QRIP) was observed.

QRIP1- Describing BF as posterior odds	3 (4.1%)
QRIP2- Not specifying Null and alternative hypotheses	18 (24.7%)
QRIP3- Simplifying the use of the prior model	51 (69.9%)
QRIP4- Not mentioning comparison of models	44 (60.3%)
QRIP5- Absolute statement	41 (56.2%)
QRIP6 - Using BF as posterior odds	13 (17.8%)
QRIP7- Considering BF as Effect size	9 (12.3%)
QRIP8- Mismatch between statistical and research hypothesis	2 (2.7%)
Other QRIPs	18 (24.7 %)

N = 73

statistical analysis. These elements include the following: The term "Bayes Factor", exact value of the BF, descriptive statistics, posterior distribution (including credible interval and highest density interval), posterior probability of a hypothesis, effect size, within-model prior distribution,[4] one or more other illustrative within-model prior distributions[5] and, lastly, mentioning the link between the BF and the posterior odds. All these statistical elements are crucial to evaluate the results of a study.

Results

The observed frequencies and percentages for each QRIP are shown in Table 1. The references of the citations in this section are anonymized and referred to their data ID in the dataset since these examples are for illustrative purposes only.

The confusion between BF and posterior odds (QRIP 1 & 6)

The actual description of a BF as posterior odds is not prevalent in the sample since there are only 3 papers describing the BF as posterior odds. For instance, article [13] mentioned "The test yields a statistic—the Bayes factor, B_{01}—that is a posterior odds ratio relating the probability that the null hypothesis is true to the probability that the alternative hypothesis is true, given the data". Another example would be from article [21] "...BF_{10} quantifies the posterior probability ratio of the alternative hypothesis as compared to the null hypothesis".

However, *interpreting* BF as posterior odds (QRIP 6) is four times more frequent than describing it as such (QRIP 1). This phenomenon comes with a pattern of reporting like "H_1 is (much) more likely or probable than H_0" if BF_{10} is (much) higher than 1. For instance, article [38] wrote "The Bayesian test for correlations resulted in a Bayes Factor of BF_{01} = 5.88, indicating that the null hypothesis (i.e., absence of correlation) is ~6 times more likely than the alternative hypothesis (i.e., presence of correlation)". Here we considered that in practice the words likely and probable are used interchangeably and have the same meaning.[6] Then concluding that H_1 is (much) more likely or probable than H_0 clearly is correct if the posterior odds (i.e., the probability that H_1 is true divided by the probability that H_0 is true, given the data) is (much) higher than 1. However, this is *not correct* on the basis of the BF, because a BF higher than 1 may be associated with any posterior odds ratio (i.e., higher than 1 or smaller than 1) depending on the prior odds, because the posterior odds equal the BF times the prior odds. This reporting pattern was observed in multiple publications. It was reported in different manners as well, such as "A model with an effect of soundtype was also more likely than the null model of no difference in soundtype (BF_{10} > 100,000). Finally, a model in which these factors interact was more likely than a model with no interaction (BF_{10} = 4.95)" [47] and "Bayes factors showed that the alternative hypothesis with scores lower than chance was over 1,000 times more likely given the data" [3]. The difference of the frequencies between QRIP1 and QRIP 6 could be ascribed to the fact that not every author reported BFs with an elaboration about the definition of the BF.

Research hypotheses and statistical hypotheses (QRIP 2 & 8)

In 18 publications, the alternative hypotheses were not clear or were not mentioned at all (QRIP 2), which leads to ambiguity in the reported statistics. Please note that the actual number is bigger than 18 since we did not count the papers that conducted ANOVAs. In these cases, the null and

4 Unlike QRIP3, the reported prior models were recorded only without noting whether any justification for the chosen prior was provided or not.

5 Considering the influence of the within-model prior on the BF, it was recorded whether the authors reported additional BFs with different priors for illustrative purposes.

6 E.g., see https://matheducators.stackexchange.com/questions/7456/probable-vs-likely-choosing-the-appropriate-word

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

the alternative hypothesis are clear because of the nature of ANOVAs (H_0: group means are equal, H_1: at least one group mean is different from another). Conclusions derived from the tests without the specification of the alternative hypothesis might lead to unsound or invalid conclusions. It can be exemplified by one of the two papers that demonstrated the mismatch between statistical hypothesis and research hypothesis (QRIP 8). The authors of article [10] conducted a study about the effect of nicotine on pupil size and hypothesized that "... pupil size would be smaller after the administration of nicotine, as compared to placebo...". They found that "Results suggested that pupil size during target assignment in the nicotine condition (M = 4.553, SD = 0.782) was significantly smaller than in the placebo condition (M = 4.681, SD = 0.827), F(1, 28) = 8.232, p = .008, η^2_p = 0.227, $BF_{incl.}$ = 1.316e+6 (Drug as the best model)", and concluded, "Results from both NHST and Bayesian analyses suggested that pupil size in the nicotine condition was smaller than pupil size in the placebo condition". Even though the authors made use of descriptive statistics to reach the final conclusion, the hypothesis tested does not correspond to their research hypothesis. A one-side prediction was formed but ANOVA was conducted (i.e the use of ANOVA implicitly implied the within-model prior was two-sided rather than one-sided.) . Although such behavior might be easier to handle in frequentist testing (since it only involves dividing p-values by 2), it is problematic under NHBT since the value of the BF changes by an unknown amount, because an entirely different prior for the alternative hypothesis is employed.

The researchers from article [16] conducted a study for investigating "... whether alcohol affects saccade countermanding; i.e., whether saccadic SSRT is lengthened during acute intoxication". However, model comparison by ANOVA was conducted, which lead to a similar problem as the example from the previous paragraph.

Not specifying the prior distribution used with the alternative model (QRIP 3 & 5)

The alternative hypothesis in NHBT is usually the aggregate of all non-null values. Since the null hypothesis is often associated with "there is no difference between the means" or "there is no association", the interpretation of the alternative could be "there is a difference between the means" or "there is an association". It may be easily overlooked that the alternative hypothesis actually incorporates a prior probability density of all specified non-null values, and hence is just one instance of an alternative hypothesis specifying that there is an effect while many other instances are possible. For this reason, researchers should provide justification for the chosen within-model prior. Furthermore, one cannot simply conclude the presence or the absence of an effect on the basis of the BF since it only quantifies relative evidence for the two competing hypotheses. However, 51 papers simplified the use of the within-model prior and 41 papers made absolute statements based on their observed BFs.

It is hard to illustrate that prior distributions have not been specified, but, as it did not happen that often, it may be illustrative to quote a paper in which the prior distribution *was* specified and justified, see [35]: "the prior distribution as a truncated gaussian with $\mu = 1$ and $\sigma = 2$ (...) with a lower bound of the distribution at 0 and an upper bound at 2..." and "These values have been chosen taking into account the typical range of effects observed in previous studies". Article [63] provided a similar reason for their chosen prior: "Effect sizes used to set priors were obtained from related results of previous studies". The authors of article [42] added "Typically, researchers entertain a distribution of possible alternatives to the null, some of which are judged more likely than others. For example, we may anticipate a high probability of a small-to-medium effect, a moderate probability of a large effect, and a very remote chance of a gargantuan effect. (...) The prior distribution, therefore, influences the BF, and the choice of a prior distribution is at the researcher's disposal; it may be based on knowledge about likely values established in previous work, or it may be chosen to be minimally informative". As can be seen from Table 1, in almost half of the cases no such explanations were given (n = 24) or the priors model were not reported (n = 27).

Two examples of QRIP 5 follow. Article [22] reported that "Of the 24 resulting comparisons, four produced a Bayes factor greater than 1 (i.e., providing evidence for an effect of initial category structure).(...). The other three comparisons *showed a difference* between the one-dimensional linear boundary and the other initial category structures for the Shepard circles" where the italics are ours, to pinpoint the absolute statement. Article [41] asserted "Both frequentist and Bayesian analyses revealed no main effects for responses of 'environment,' errors on SART, and no interaction for any of the measures studied". Needless to say, both absolute assertions are questionable and unjustified by the statistical tool used.

Presenting BFs as evidence for one hypothesis without mentioning it is relative to another competing hypothesis (QRIP 4)

Interpreting BFs as absolute rather than relative evidence prevails among practitioners in the sample in which 44 papers demonstrated this questionable practice. This QRIP also comes with a specific pattern of reporting like "The Bayes Factor provides evidence in favor of H_1 (or H_0)". These reports did not mention that the strength of evidence always depends on the two specified models. Suppose one finds BF_{01} is bigger than one. It is unideal to simply conclude that the evidence is in favour of the null hypothesis ("there is no effect" or "there is no difference between the means") since the BF is always relative to another competing hypothesis and the value of the BF depends on the prior distribution associated with the alternative hypothesis. Article [72] stressed this notion by mentioning that "The Bayes factor BF_{+0} quantifies the evidence that the data provide for H_+ (i.e. the presence of the compensatory control effect) relative to H_0 (i.e. the absence of the compensatory control effect)". The researchers of the article obtained a BF of 5.41 and interpreted the result as "This means that the data are about 5.41 times more likely under the null model including only religiosity, *compared to* the

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

alternative model that also includes the control-threat manipulate" (italic is ours).

Considering the Bayes Factor as a measure of effect size (QRIP 7)

The value of a BF does not indicate the magnitude of an effect or effect size. There are 9 papers in the sample that nevertheless did make a statement to this effect. Article [54] reported "As both the confidence interval and the Bayes factor do not point towards a true difference and the t-test is borderline significant, this can be considered a very small or non-existent effect." A similar questionable reporting practice is in article [15]: "A repeated-measures ANOVA with Cue (novel; standard) and Electrode (Fz; Cz; Pz) as factors showed that there is substantial evidence for a null effect (BF_{01} = 5.676), suggesting that the effect of expectations on the N2 was not very strong.", article [32] wrote "Bayesian statistics again supported that sexual motivation had a strong effect on attractiveness ratings in the control condition ($Bf_{1,0}$ = 25.99)", and article [16] wrote "Both a Pearson's correlation and a Bayesian correlation analysis revealed no strong relationship between the difference in relative change from pre-drink to post-drink between alcohol and placebo for the saccadic task and the manual tasks (r = 0.197, p = 0.22, BF = 0.41)". For article [16], it has to be noted that the Pearson's correlation did support the claim "there is no strong relationship in the experiment", however, such assertation cannot be supported by the BF.

Other QRIPs

The most common QRIP we found in the sample was basing conclusions on an inconclusive BF (n = 16). For this part of the study, we simply defined a BF as inconclusive if its value was between $10^{-0.5}$ and $10^{0.5}$, as Jeffreys (1948) calls these BFs *not worth more than a bare mention.*[7] Any BF close to 1 indicates the evidence is in favor of one hypothesis as much as of the other competing hypothesis. It is described as *absence of evidence* (Keysers et al., 2020). However, the researchers in the studied papers confuse *absence of evidence* with *evidence of absence.* For instance, article [10] asserted that "A chi-squared test with a contingency table revealed that responses in the placebo condition had no association with those in the nicotine condition, χ^2 (2, N = 29) = 2.193, p = .139, BF_{10} = 1.17". The BF_{10} only suggests that the likelihood of obtaining the observed data under the null model is more or less the same as under the model with interaction. This questionable reporting practice might inherit from NHST, as can be illustrated by the following example. From Article [21]: "The length-scale parameter λ did not differ significantly between the three experiments (all p > 0.5, BF_{10} = 1.1)".

Two other rare QRIPs were observed. Article [18] gave a wrong impression about the function of the BF: "The JZS-Bayes factor gives the probability of evidence that the null hypothesis is true". Article [61] gave a problematic interpretation of the alternative hypothesis: "The results of a Bayesian t test (with the hypothesis H_1: $d'_{PHA} \neq d'_{HC}$) shows a Bayes factor (BF_{01}) of 0.51, suggesting that hypothesis H1 (that the PHA shows a better detection performance than the HC) is 1.96 times more likely than hypothesis H0 (that the PHA and the HC do not differ significantly), providing only weak anecdotal support for H1". The given statistical alternative hypothesis posited that there is a difference, but a one-sided interpretation was given.

The quotations above and additional examples from the studied papers for illustrating the presence and the absence for each QRIP are summarized in Appendix A.

Table 2. The frequencies and the percentages (in parentheses) observed for each reported statistical element

BF	71 (97.3%)
Exact value of BF	66 (90.4%)
Descriptive Statistics	68 (93.2%)
Posterior distribution	13 (17.8%)
Posterior probability of a hypothesis	5 (6.8%)
Effect size	59 (80.8%)
Within-model prior distribution	46 (63%)
Other illustrative within-model prior distribution	3 (4.1%)
Linkage between BF and posterior odds	7 (9.6%)

Reported Statistical Elements in the papers

The frequencies and the percentages observed for each reported statistical element is shown in Table 2.

Almost all papers reported BFs except two papers providing posterior probabilities of the models instead. Furthermore, five papers did not give exact values, but only reported the gradation of the BFs. An example can be found in the quotation from article [22] under section QRIP 3 & 5, where it was only mentioned for which tests BFs larger than 1 were found. Five other papers did not provide any descriptive statistics. One of those is a stimulating study, three studies are reanalyses of existing datasets, and one paper is about the models goodness of fit. A small minority of publications made use of Bayesian estimation tools (e.g., credible interval and/or highest density interval) and/or posterior probability distributions next to their reported BFs. However, most of the papers reported effect sizes. Despite the importance of the within-model prior (i.e., probability distribution used with the alternative model) for the BF, 27 publications did not specify or mention the chosen within-prior model for the tests and 3 papers provided BFs associated to priors other than their own specified prior models as illustrative examples for the readers.

Lastly, the relationship between the BF and the posterior odds were elaborated in 7 publications. For instance, article [13] stressed that "We report the raw Bayes factor for each

7 $10^{0.5} \approx 3.16$

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

single-degree-of freedom analysis, which allows readers to draw their own subjective conclusions about the degree to which the evidence favors either or neither hypothesis for a given comparison", which pinpoints the difference between evidence and degree of belief. Article [74] explicitly mentioned "The Bayes factor indicates the change from prior to posterior odds brought about by the data.". Articles [31], [42], [60], and [72] also elaborated the linkage by introducing Bayes' theorem.

Discussion

The results of Study 1 indicate that at least 8 Questionable Reporting and Interpreting Practices were signaled repeatedly in the set of studied papers. More specifically, 67 (or 92% of) sampled papers showed at least one of the 8 prespecified QRIPs. Among these QRIPs, simplifying the use of the within-model prior by using the default prior or not mentioning the chosen prior (QRIP 3), presenting the BF as absolute rather than relative evidence (QRIP 4), and making an absolute statement based on the BF (QRIP 5) are the most common ones in the *sample*. Moreover, a large majority of the publications reported BFs with descriptive statistics, prior model and effect size, but a small minority did not. However, the use of Bayesian estimation tools and posterior probability distributions is rare, which can possibly be attributed to the choice of keywords in the search on Google Scholar since the keywords pertain to testing only. Providing BFs based on other illustrative within-model priors is observed in a small number of papers despite the influence of the prior model on the BF. Additionally, only a few papers demonstrated the linkage between the BF and posterior odds.

The observed QRIPs also suggest there are some potential mismatches between the usability of the BF and researchers' expectations. Quite a few researchers seem to expect that the BF can establish the presence or the absence of an effect without acknowledging that the chosen prior is just one instance of the alternative hypothesis. Their conclusion based on the BF would be sound if the prior distribution indeed (closely) corresponds to a prior distribution of general interest. However, researchers cannot ensure that this is the case, and it might actually be very unlikely that such general consensus on prior distributions would actually exist. For this reason, providing the information about the chosen prior model and using other illustrative prior distributions is important for the readers and the researchers themselves to make a balanced conclusion. Furthermore, many papers seem to demonstrate the expectation that the BF can offer *absolute* (or direct) evidence in favor of either of the two competing hypotheses, which is a questionable interpretation. The users should realize that the BF is a relative measure of evidence for the two competing hypotheses. The BF-based evidence for one hypothesis is always dependent on another hypothesis. Failures of recognizing these two notions (the influence of within-model prior and relative notion of evidence) of the BF might lead to a dichotomous mindset and faulty interpretation.

Study 1 faces several limitations concerning its generalizability. The actual number of published papers in psychology which employed NHBT is unknown. The generalizability of this study entirely depends on the search engine on Google Scholar, and on our inclusion and exclusion criteria. Moreover, the use of the keyword "Bayesian Test" in the search led to a great reduction in the number of the returned papers, which might have led to insufficient coverage of all targeted papers. As such, the sample might not be representative of the population. If the engine indeed did return the majority of or all the targeted papers, we would expect that this study gives a general picture of the practice of Bayesian testing in the psychology community as a whole. Otherwise, its generalizability is limited. Despite all these, it has to be stressed that the primary aim of the study was to assess the mere existence of such QRIPs, and it clearly did so *in the sampled papers*. Future research with larger sample size is needed to explore how prevalent these QRIPs are among psychology researchers in general.

The observed QRIPs might not necessarily imply there is a mismatch between the usability of the tools and researchers' expectations since there can be a difference between what researchers wrote and what researchers' expectations are, but it offers valuable insight into what kind of mismatch there could be. For this reason, we next conducted a survey that mainly focuses on what researchers expect from NHBT. The study is similar to previous studies about the misinterpretation of the p-value and the confidence interval, in which the participants were given a hypothetical result of a statistical analysis with a few interpretational statements and were asked to judge if each statement is true or false (Haller & Krauss, 2002; Hoekstra et al., 2014; Oakes, 1986).

Study 2

Method

Participants

Our targeted participants are psychology researchers who employed the BF-based null hypothesis testing in their empirical research, and Bayesian methodologists. Many of the references in Study 1 were reused since the authors (including correspondence authors and co-authors) of the 228 excluded methodological papers, the 73 targeted papers, and one of the other excluded papers, corresponded to our research interest. For this reason, additional searches on Google (Scholar), Linkedin, and ResearchGate were conducted to verify the identity of the authors and collect their email addresses based on the information that is given in the publication. In total, 805 authors and their email addresses were successfully identified and collected. Out of those, there were 226 duplications, after elimination of which we ended up with 579 authors, which served as the targeted set of participants for our study.

Material and procedure

Invitation emails were sent alongside a link to the survey on Qualtrics. The participants were informed that they can leave their email addresses at the end of the survey for receiving the result of the study, including the marking scheme, their score, and the overall mismatch of all participants. They were instructed that a hypothetical result of a

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

Bayesian analysis would be given with 11 interpretational statements in turn, and they would be asked to judge if the given statements were "True" or "False". The participants could choose the option "Don't know" if they were uncertain about the truthfulness of the statements. The option "Don't Know" was introduced since knowing that the participants indeed are uncertain about the truthfulness of a statement is more informative than a randomly chosen answer when in doubt. The statements were presented one at a time and in randomized order, in order to reduce the possibility that the participants could infer the answers based on a comparison of the statements rather than their knowledge.

The presented hypothetical result of a Bayesian analysis was the following:

> *"A group of researchers conducted an experiment in which three newly developed drugs for a deadly virus have been compared to a control group. The dependent variable is the recovery rate, the null hypothesis posits that the drug does not influence the recovery rate while the alternative hypothesis posits the drug does influence the recovery rate. After the experiment, two-sided Bayesian t-tests with a default prior were conducted. For drug A, it was found that BF_{10} is 2 while for drug B BF_{10} of 20 was found. For drug C, BF_{10} of 0.1 was found."*

The 11 statements were the following:

1. *The probability of Drug A having an effect is one-tenth of that for drug B.*
2. *For Drug B, the probability of obtaining the observed data is 20 times higher under the alternative than under the null.*
3. *The alternative hypothesis (drug B has an effect) is 20 times more probable than the null hypothesis (drug B has no effect).*
4. *Drug B demonstrated a positive effect on the recovery rate.*
5. *There is evidence suggesting that drug A demonstrated an effect on the recovery rate relative to the null hypothesis (there is no effect).*
6. *For drug C, there is evidence in favor of the null hypothesis (there is no effect).*
7. *For drug B, the Bayes factor disproved the null hypothesis (there is no effect).*
8. *The effect of drug B is bigger than the effect of drug A.*
9. *The strength of drug B affecting recovery rate cannot be known with the given information.*
10. *If the same group of researchers will conduct the same experiment repeatedly, they can be expected to find the presence of the effect from drug B in 20 out of 21 times.*
11. *Based on the given Bayes Factor for drug A, drug A demonstrated no effect since the Bayes Factor is smaller than 3.*

At the end of the survey, the participants were asked to categorize themselves in one of the following groups: (1) "Psychological researcher", (2) "Methodologist (and/or Statistician)", (3) "Methodologist (and/or Statistician) and Psychological researcher", (4) "Researcher in a field other than psychology" and (5) "Other (Please specify, e.g., Methodologist and Economist)".

The participants were categorized into three groups for making comparisons in terms of the performance in the questionnaire: *psychology researchers* group (1), *experts* in methodology and/or statistics (2 and 3), and *non-psychologists* group (4). There were 6 participants that chose option (5), who were assigned to the group that seemed to match best with the specification they gave: 5 participants who specialized in methodology (and/or statistics) and a field other than psychology were assigned to the *experts* group, while one expert in psychology and neuroscience was assigned to the *psychology researchers* group.

For each of the statements, we now specify the responses that we considered to be the best fitting ones. A detailed explanation of each statement can also be found in Supplement Document S1.

Statements 1, 2, and 3, which mainly focus on the difference between the BF and posterior odds, correspond to QRIPs 1 and 6 in Study 1. To us, the best fitting responses for Statements 1 and 3 are "False" and "Don't Know". "False" for us reflects that the given statements do not necessarily follow from the given information since the BF is the marginal likelihood ratio of obtaining the observed data under the two competing hypotheses. The BF is not a posterior probability or the posterior odds of a drug having an effect or not. For the latter, prior odds are needed to calculate the posterior odds ratios, and this important information is missing. Therefore, the truthfulness of the statement could also be considered uncertain because it is based on insufficient information. For Statement 2, the expected response is "True" since the BF can be interpreted as a ratio of the probability of obtaining the observed data under the alternative versus the probability of obtaining the observed data under the null.

Statement 4 is related to QRIP 2 and 8 about the setting of a test (one-sided test and two-sided test). Based on the result from a two-sided Bayesian test, one can only conclude whether there is evidence in favor of the (two-sided) hypothesis "there is an effect" relative to the null. Therefore, to us the best fitting responses for Statement 4 are "False" and "Don't Know".

Statements 5, 6 and 7 are relevant to QRIP 3, 4 and 5, which are pertaining to the function of the BF as a relative measure of evidence of two competing hypotheses and the importance of the chosen prior distribution. To us, the best fitting response for Statement 5 is either "True" since it shows that the BF is a measure of evidence in favor of a hypothesis relative to another, or "Don't Know" since some researchers consider BFs close to 1 as meaning inconclusive evidence (compare the explanation for Statement 11). To us the best fitting responses to Statements 6 and 7 are "False" as they fail to interpret the BF as relative evidence, and, in fact, one cannot prove or disprove a hypothesis anyway.

Statements 8 and 9 are about the linkage between the BF and effect size (QRIP 7). The BF is not a measure of the magnitude of an effect, but it indicates the relative plausibility of getting the data under the two competing hypotheses. Claiming the effect of one drug is bigger than another is impossible with the given information. So, we considered options "False" and "Don't Know" the best fitting ones for Statement 8 while for Statement 9 this was the option "True".

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

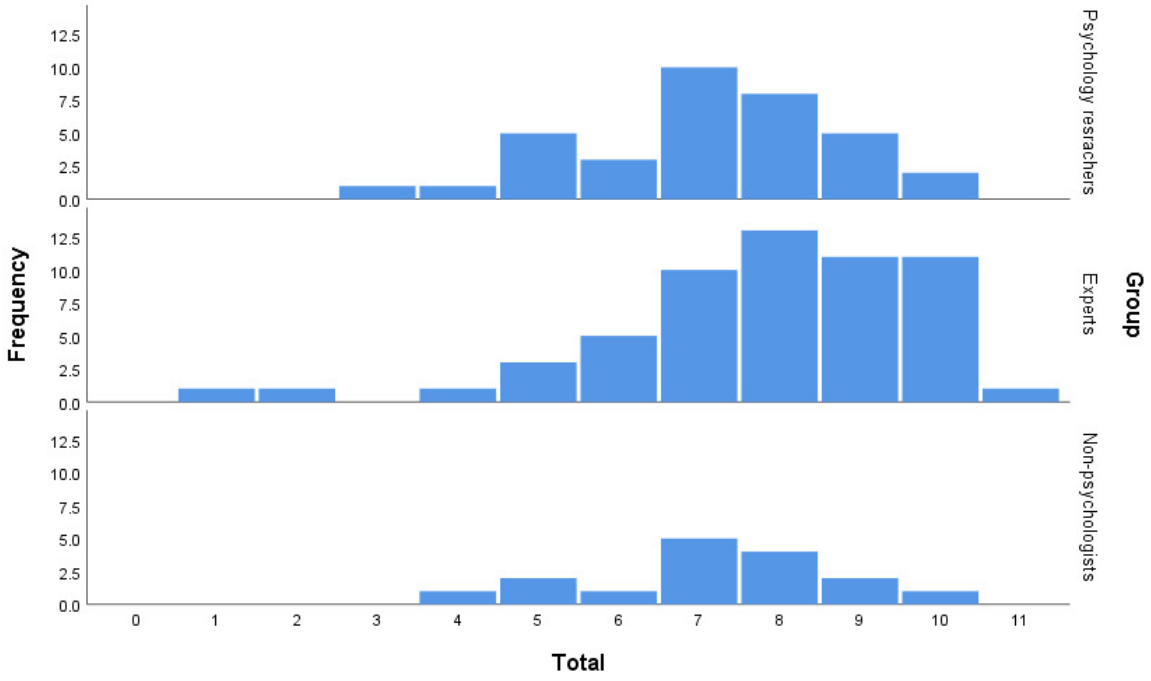

Figure 1. The distribution of the total score (number of best fitting responses given) in the different groups

For statement 10, we considered the option "False" as the best fitting one, because the BF is a measure of evidence based on the observed data only. It does not tell anything about future data.

Statement 11 is about an 'inconclusive' BF value. The BF suggests the absence of evidence, not evidence of no effect, so to us the best fitting response is "False".

It has to be stressed that these best fitting responses are from the authors' perspectives only. The ultimate "correct" answers to the statements are up to discussion and debate, and surely, there will be different opinions as to what can and what cannot be concluded.

Results

In total, 146 researchers responded to our invitation to participate in the study. Of these, 29 only gave consent but they did not proceed with the study, while 9 answered some of the statements but the question for the grouping was not answered. Therefore, these replies with missing data were excluded from the dataset. After the exclusion, the sample size is 108 and the completion rate is 18.7%. The dataset consists of responses from 35 *psychology researchers*, 57 *experts* in methodology and/or Statistics, and 16 *non-psychologists*.

The mean total score of answering the best fitting responses is 7.11 for the *psychology researchers*, 7.82 for the *experts* in methodology and statistics and 7.18 for the *non-psychologists*. The distributions of the total score can be found in Figure 1.

Figure 2 provides the visualization of the percentages of endorsing the best fitting answers for each statement and group. The first observation one could make is that by far most proportions lie between .5 and .8. This implies that within all categories, for most items, there was noticeably disagreement on the answers. Experts performed better than the other groups for most of the items. The difference in performance between the groups is not noticeable in most of the statements except for Statement 3. The proportions for Statement 6 are by far the lowest for all groups; obviously, this finding hints at a special situation with this statement, as we will discuss later on. Overall, the deviation from what we considered the best fitting responses is slightly larger among the applied researchers than the experts, considering the small difference of the means scores between the groups. Surprisingly, such a deviation is even present among the experts in methodology and statistics, since the percentages of endorsing the expected response in Study 2 often clearly fall short of 100%. A final noteworthy finding is the low proportion of best fitting answers for Statement 3 in the psychology researchers group.

Table 3 show the frequencies of endorsing different options for each statement in various groups. The corresponding proportions of endorsing each option can be found in Supplement Document S1.

As for Statement 1 and 2, most of the participants answered expectedly, with small frequencies of picking the option "Don't know". Despite Statement 3 being based on the same QRIPs, there are drops in the proportions endorsing the best fitting answers. Moreover, the "Don't know" was chosen more often than for Statement 2.

As to Statement 4, there are 11 (31.4%) psychology researchers and 16 (28.1%) experts who believed it to be true even though there is a mismatch between the conclusion and the research hypothesis.

Regarding Statement 5, the majority of the participants gave the best fitting responses and the proportions for psychology researchers and experts are around 70%. Again,

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

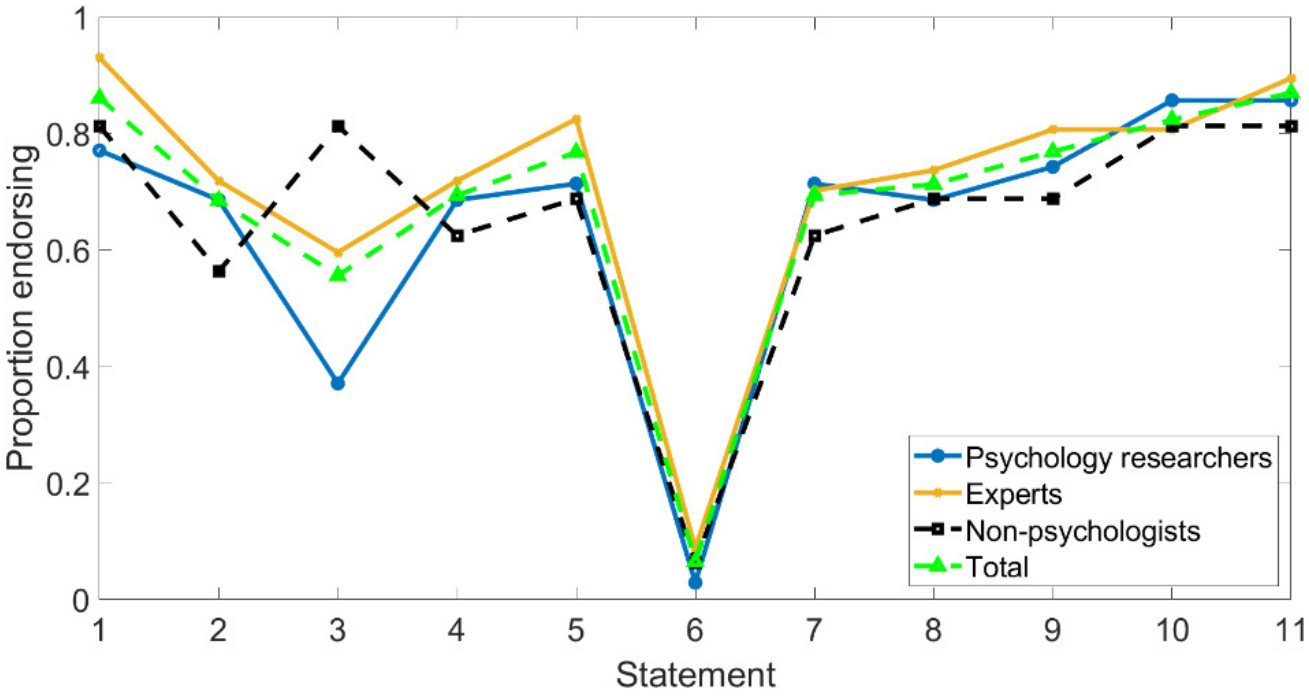

Figure 2. The proportions of participants answering the best fitting answer for each statement in each group.

Table 3. The frequencies of endorsing different options for each statement in various groups (with the expected responses for the statements in the first column).

		Group								
	Best fitting	Psychologist researchers			Experts			Non-psychologist		
		True	False	Don't Know	True	False	Don't Know	True	False	Don't Know
S1	F or D	8	20	7	4	45	8	3	9	4
S2	T	24	9	2	41	15	1	9	4	3
S3	F or D	22	8	5	23	25	9	3	9	4
S4	F or D	11	13	11	16	21	20	6	8	2
S5	T or D	21	10	4	42	10	5	10	5	1
S6	F	32	1	2	49	5	3	14	1	1
S7	F	6	25	4	11	40	6	6	10	0
S8	F or D	11	12	12	15	21	21	5	6	5
S9	T	26	6	3	46	10	1	11	3	2
S10	F	2	30	3	3	46	8	0	13	3
S11	F	3	30	2	4	51	2	2	13	1

Note. T and F denote options "True" and "False" respectively, while D denotes option "Don't know"

only a minority of participants selected "Don't know" for the statement. Statement 7 has similar statistical figures as Statement 5. As for Statement 6, an overwhelming majority of the participants chose the option "True" regardless of the groups.

For Statement 8, concerned with the association between the BF and effect sizes (QRIP 7), the proportions of endorsing the expected option are around 70% in each group. For each group, roughly the same numbers of participants answered "Don't know" and "False". For Statement 9, most of the participants (n = 83) picked "True" and only 6 participants chose "Don't know". Lastly, the proportions of answering the best fitting option ("False") for Statements 10 and 11 are noticeably higher than the other statements except Statement 1.

Discussion

The results of Study 2 suggest: (1) a potential mismatch between the participants' interpretation and what can actually be expected from the BF, and (2) a potential disagreement on what can be and what cannot be concluded based

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

on the BFs among the methodologists. Before discussing further the implications of Study 2, a comparison of the results between Study 1 and Study 2 is made. In Study 1, the percentages of the studied papers describing BF as posterior odds (QRIP 1) and using it as posterior odds (QRIP 6) are 4.1% and 17.8% respectively while in Study 2 only 24% of psychology researchers (n = 9) gave the expected response for all three Statements 1, 2 and 3 that are associated with these QRIPs. Most researchers did not interpret the BF as posterior odds in the studied papers, but more than half of the participants in the survey seem to expect that the BF can actually be used as posterior odds (Statement 3). This might suggest that posterior probability and posterior odds are the tools researchers are actually looking for rather than the BF.

In 51 out of 73 publications, simplified use of the prior probability distribution for the alternative hypothesis (QRIP 3) was observed. Using the BF without noting it is relative (QRIP 4) is also common among the studied papers. In Study 2, this phenomenon manifested itself with Statement 6. As for Statements 8 and 9, about the linkage between effect size and BF (QRIP 7), the results of both studies agree as the data show that a fairly small percentage (roughly 20-30%) in each group made such a link. More specifically, only 4 psychology researchers (11.4%) gave unexpected responses to both statements. As for the potential misinterpretation of inconclusive BFs (Statement 11), both studies pointed out that most of the participants by far correctly distinguished the concepts evidence of absence and absence of evidence. The prevailing misinterpretation of frequentist statistical tools is considered as evidence suggesting that there is a discrepancy between the function of the tools and the researchers' expectations towards these tools (Gigerenzer, 2004; Haller & Krauss, 2002). Meanwhile, Bayesian statistics is proposed as an ideal substitute for frequentist tools since it coincides better with researchers' expectations and needs (Gigerenzer, 2004; Haller & Krauss, 2002; Haucke et al., 2020). This study aimed at investigating the potential mismatch of the NHBT with psychology researchers. The results of Study 2 suggest that a mismatch indeed exists for a considerable number of *psychology* researchers.

The results suggest there are two clear mismatches among the psychology researchers. A number of participants apparently expected one can draw direct probabilistic statements about the hypotheses based on the BF (Statement 3) and interpret the BF as evidence for one hypothesis without noting that it is a relative measure of evidence on two specified hypotheses (Statement 6). As far as Statement 3 is concerned, the participants and the authors who committed QRIP 6 in Study 1 might draw a probabilistic conclusion by assuming the prior odds equal to 1 as a tacit default. Therefore, Statement 3 could be "True" and these publications commit no QRIPs that is associated with the confusion over BF and posterior odds. However, the authors see such supposition as a potential QRIP considering that the chosen or the assumed prior odds need to be justified as well (e.g., based on previous evidence or simply subjective intuition). Moreover, equal prior odds might only be suitable in certain specific situations since it means "equal probabilities for the single null value 0 for the effects size versus all non-null values for the effect size". In the studied papers, only article [9] provided justification for their chosen prior odds ". Because previous studies have found mixed results with regards to the effect of the number of tasks on task switching performance, the prior probabilities assumed for H_1 and H_0 were 0.5 ("uninformative priors")."

As for Statement 6, such a mismatch prevails among all participants in the various groups since less than one-tenth of the participants endorsed the answer that to us is best fitting. A possible explanation of this phenomenon could be that the participants might consider that interpretation from Statements 5 and 6 is similar, and it is not necessary to stress the notion of "relative evidence" as the given BF for drug C indicates there is clear evidence for the null relative to the alternative. Moreover, the participants might well be aware of the notion of relative evidence in the BF and assumed it is well known by other readers and experts as well. Therefore, one might report the BF as in Statement 5 at their convenience as a shortcut. For these reasons, it might also be argued that our interpretation of Statement 6 is too strict and demanding.

However, the authors disagree with these views. NHBT using the BF only quantifies evidence for the point-null hypothesis "relative to" a specific within-model prior distribution that is associated to the alternative hypothesis. The value of the BF is sensitive to that chosen prior distribution. Therefore, simply concluding there is evidence in favor of the null hypothesis without noting that it is relative evidence can be misleading, as it implies that this is evidence in favor of all conceivable alternative hypotheses, while an indefinite number of within-model priors is available. The evidence might shift from favoring the null to the alternative and vice versa when different within-model priors are chosen for the alternative hypothesis. For this reason, pointing out the notion of "relative evidence" is crucial. Second, the notion of relative evidence might not be well known by non-methodologists (and/or non-statisticians), hence in particular applied researchers might neglect the influence of the within-model prior on the BFs when interpreting results. It could possibly lead to mindless use (e.g., making use of a within-model prior without giving any justification) and hence questionable interpretation of the BF in research.

As for the confusion between BF and posterior odds, the result for Statement 3 is noteworthy since the non-psychologists did not only outperform psychology researchers, but also the experts in methodology and statistics. Initially, it was planned to study only the potential mismatch between psychologists and their expectation, and possibly make comparisons with the experts. However, a reasonable amount of data from non-psychologists were collected, hence, their data were also presented. It has to be noted that the sample size of the non-psychologists group (n = 16) is small relative to the group size (n = 57) of the experts, and hence the role of chance could be strong. Since we do not have random samples here it is difficult to assess the uncertainty, but we give a tentative answer by means of Bayesian estimation, as follows. Suppose the assumption of randomization would hold and consider the number of participants endorsing the expected response as the number of successes. The outcome would then follow a binomial dis-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

Table 4. 95% Highest density intervals of the proportion endorsing the expected response for Statement 3 in each group with different priors and 95% Confidence Intervals.

Group / Prior	Beta(6,3)	Beta(3,3)	Beta(3,6)	95% CI
psychology researchers	(.29, .58)	(.25, .54)	(.23, .50)	(.23, .54)
Experts	(.49, .72)	(.47, .71)	(.44, .68)	(.47, .71)
Non-psychologists	(.59, .91)	(.54, .90)	(.45, .82)	(.56, .94)

tribution. The 95% highest density intervals (HDI) for estimating the proportion of the groups endorsing the expected response using three quite different but (in our opinion still realistic) priors and 95% confidence intervals (CI) are given in Table 4. Specifically, all prior distributions were chosen to refer to situations with little information. We chose three different priors in order to reflect what happens if we take for granted that actually higher endorsement proportions can be expected for one group (notably) experts more than for another group, see below. The HDIs and CIs suggest that there is not a huge difference between the groups when using the same priors. Interestingly, using a set of priors suggesting realistically that experts would perform quite a bit better than the non-psychologists, that is using Beta(6,3) for experts and Beta(3, 6) for non-psychologists, then the posteriors still reveal very little difference. Furthermore, the performance of the psychologists probably is inferior to these two groups based on the given HDIs and CIs regardless of the chosen prior. However, as mentioned, we only suppose that we have random samples. It is well possible that the non-psychologists are the co-authors in some of the methodological papers since it is the most probable reason why they are targeted and invited to the study in the first place. Based on these reasons, one should be cautious in interpreting the difference between the experts and non-psychologists.

Unlike previous studies (Haller & Krauss, 2002; Oakes, 1986), the option "Don't know" was introduced in the survey. Initially, the rationale for using it was that the results would be more informative, since the answers would not be randomly chosen when the participants actually simply did not know what to answer. However, the meaning of "Don't know" can be twofold. The participants may have chosen this option because they did not know the correct answer for the question based on their knowledge about the BF, as intended. Meanwhile, it could also mean that the truthfulness of the given statements is indeed uncertain with the given hypothetical statistical results. This ambiguity makes the interpretation of the results of this study somewhat complicated. For instance, one participant scored 5 out of 11 by picking "Don't know" for all the statements. It might create uncertainty on whether the participants literally "don't know" the answer because of insufficient knowledge or because they judge the correct answer for the statement to be indeed "Don't know". The option "Don't know" comes with another unexpected side effect, since its definition overlaps with the meaning of "False". A few participants did consult the author and expressed their concern on the definition of the options "Don't know" and "False" before proceeding with the study. Therefore, our findings might not be comparable to previous research on frequentist statistics.

The education level of the authors and the participants on Bayesian Statistics could be another factor making the comparison between our findings and previous studies unideal. These psychology researchers are likely to be self-taught Bayesian who receive informal and insufficient education on using the BF, whereas the participants in the research on the misuse of p-value probably have been given substantial amounts of formal as well as informal training on the (use of the) p-value. Such a difference would make the comparison inappropriate. Indeed, the authors received an email from one recipient of our invitation to the survey study indicating that their Bayesian analysis was done by an expert rather than themselves, and another one indicating that the researcher knew nothing about BFs and suggested us a methodologist to invite for our study.

One might question whether the results of Study 2 are generalizable since the completion rate is only 18.7%. There may be a greater or smaller discrepancy among those who did not participate in the study. Moreover, 38 participants withdrew from the study after reading the instruction or during the study. It could very well make a difference in the results if these responses were completed. Also, the procedure of searching literature might not be ideal for making inference, as was already addressed in the discussion of Study 1. Lastly, the generalizability of our Study 2 as far as results for methodologists are concerned, may be susceptible to self-selection bias in a way that the methodologists who are more knowledgeable about BFs might be more likely and willing to participate in the survey study, which would result in underestimated numbers of mistakes. However, we conclude that at least for a non-negligible number of researchers discrepancies appear to exist between their expectations and the function of the BF, or else the formulations they used in their papers are not sufficiently precise to make sure that they actually understand what can and cannot be concluded on the basis of a BF. In the latter case, one can expect that this lack of precision will easily lead to misconceptions by readers of their papers. Interpreting the BF as posterior odds and not acknowledging the notion of relative evidence are arguably the most striking of these.

For statistical practice, we think our studies show that researchers should become more aware of what can and cannot be concluded with the methods by properly understanding the foundation and the underlying reasoning of these tests. Thus, the researchers can apply these tools knowledgeably and be well informed on what can be achieved and cannot be achieved with these tools in their research. As for future generations of researchers, changing the way statistical inference is taught, and modifying the educational material could help in solving the ubiquitous misunderstanding of conventional testing methods as well as Bayesian testing methods in the long run. This would en-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

tail thorough teaching of what *can* and especially also what *cannot* be concluded from a *p* value or a Bayes Factor. For instance, neither from a very small Bayes factor (of H_1 versus H_0) nor from a high *p* value, can one conclude that there probably is no effect. For this type of conclusion further information is needed. The least would be a HDI or a CI to get an idea of a plausible range of values for the effect size parameter. If this plausible range covers only values that for all practical purposes are small, one could conclude that the effect size is probably small for all practical purposes. Conversely, from a (very) high Bayes factor (of H_1 versus H_0) or a (very) small *p* value, one cannot conclude yet that the effect is sizeable. Indeed, high BFs and small *p* values can also be found for effect sizes that are negligible for all practical purposes. Strictly speaking, all one can conclude is that the effect size is probably not exactly 0. A researcher's experience with the kind of research it pertains to, may help to gauge what such outcomes actually pertain to, but for (less experienced) readers, it will always be essential to, in addition come up with, for instance, an HDI or CI to give an indication of what are plausible ranges of values for the effect size at hand. Alternatives to this could be giving standard errors or posterior distributions, in addition to observed effect size values, but in any case, what should be taught is that such single value summarizers are incomplete in order to draw useful conclusions. This actually is in line with the two decades old guideline for reporting classical statistical results by Wilkinson et al. (1999), and is here analogously applied to Bayesian statistical analysis, e.g. also see Tendeiro and Kiers (2022).

Author Note

The work described in this article was carried out as a bachelor's thesis project by Tsz Keung Wong at Department of Psychology, University of Groningen under the supervision of Henk Kiers, h.a.l.kiers@rug.nl. At the time of the publication, Tsz Keung Wong is a research master student at Tilburg University.

Contributions

Contributed to conception and design: TKW, HK, JT
Contributed to acquisition of data: TKW
Contributed to analysis and interpretation of data: TKW, HK, JT
Drafted and/or revised the article: TKW, HK, JT
Approved the submitted version for publication: TKW, HK, JT

Funding

This research was supported by a Japanese JSPS KAKENHI grant awarded to Jorge N. Tendeiro (21K20211).

Competing Interests

The authors declare no competing interests.

Data Accessibility Statement

Research material, data and R scripts are uploaded as online supporting information through the Open Science Framework (OSF), and can be accessed using this link: https://osf.io/wv8qk/

Submitted: August 10, 2021 PDT, Accepted: June 08, 2022 PDT

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at http://creativecommons.org/licenses/by/4.0 and legal code at http://creativecommons.org/licenses/by/4.0/legalcode for more information.

References

Assaf, A. G., & Tsionas, M. (2018). Bayes factors vs. P-values. *Tourism Management*, *67*, 17–31. https://doi.org/10.1016/j.tourman.2017.11.011

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066x.49.12.997

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Fidler, F., & Loftus, G. R. (2009). Why Figures with Error Bars Should Replace *p* Values. *Zeitschrift Für Psychologie / Journal of Psychology*, *217*(1), 27–37. https://doi.org/10.1027/0044-3409.217.1.27

Gelman, A., & Shalizi, C. R. (2012). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*(5), 587–560. https://doi.org/10.1016/j.socec.2004.09.033

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research*, *7*(1), 1–20.

Haucke, M., Miosga, J., Hoekstra, R., & van Ravenzwaaij, D. (2020). *Bayesian frequentists: Examining the paradox between what researchers can conclude versus what they want to conclude from statistical results*. https://doi.org/10.31234/osf.io/escvy

Hoekstra, R., Johnson, A., & Kiers, H. A. L. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement*, *72*(6), 1039–1052. https://doi.org/10.1177/0013164412450297

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Iverson, G. J., Lee, M. D., Zhang, S., & Wagenmakers, E.-J. (2009). p(rep): An agony in five fits. *Journal of Mathematical Psychology*, *53*(4), 195–202. https://doi.org/10.1016/j.jmp.2008.09.004

Jeffreys, H. (1948). *Theory of probability* (2nd ed.). Clarendon Press.

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*(7), 788–799. https://doi.org/10.1038/s41593-020-0660-4

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). An evaluation of alternative methods for testing hypotheses, from the perspective of harold jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55. https://doi.org/10.1016/j.jmp.2016.01.003

Lyu, X.-K., Xu, Y., Zhao, X.-F., Zuo, X.-N., & Hu, C.-P. (2020). Beyond psychology: Prevalence of *p* value and confidence interval misinterpretation across different fields. *Journal of Pacific Rim Psychology*, *14*, e6. https://doi.org/10.1017/prp.2019.28

Lyu, Z., Peng, K., & Hu, C.-P. (2018). P-Value, Confidence Intervals, and Statistical Inference: A New Dataset of Misinterpretation. *Frontiers in Psychology*, *9*(868). https://doi.org/10.3389/fpsyg.2018.00868

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Morey, R. D., Hoekstra, R., Rouder, J. N., & Wagenmakers, E.-J. (2016). Continued misinterpretation of confidence intervals: Response to Miller and Ulrich. *Psychonomic Bulletin & Review*, *23*(1), 131–140. https://doi.org/10.3758/s13423-015-0955-8

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs* (R package version 0.9.12-4.2). https://CRAN.R-project.org/package=BayesFactor

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Wiley.

Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, *72*, 33–37. https://doi.org/10.1016/j.jmp.2015.08.002

Robert, C. P., Chopin, N., & Rousseau, J. (2009). Harold Jeffreys's Theory of Probability Revisited. *Statistical Science*, *24*(2). https://doi.org/10.1214/09-sts284

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/10.1037/met0000221

Tendeiro, J. N., & Kiers, H. A. L. (2022). *With Bayesian Estimation One Can Get All That Bayes Factors Offer, and More*. https://doi.org/10.31234/osf.io/zbpmy

Tendeiro, J. N., Kiers, H. A. L., & van Ravenzwaaij, D. (2021). Worked-out examples of the adequacy of Bayesian optional stopping. *Psychonomic Bulletin & Review*, 1–18.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*(1), 1–4. https://doi.org/10.3758/s13423-018-1443-8

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/bf03194105

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066x.54.8.594

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

Appendix A

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

Table A. Quotations from the sample for illustrating the presence and absence of each QRIP.

	Presence of QRIP	Absence of QRIP
QRIP1 - Describing BF as posterior odds	[13] "The test yields a statistic—the Bayes factor, B_{01}—that is a posterior odds ratio relating the probability that the null hypothesis is true to the probability that the alternative hypothesis is true, given the data" [21] "…BF_{10} quantifies the posterior probability ratio of the alternative hypothesis as compared to the null hypothesis". [24]" Bayesian tests were used because these afford evaluating the relative probabilities of null and alternative hypothesis given the data"	[72] "the Bayes factor BF_{+0} then represents the ratio of the marginal likelihoods of the observed data under M_{exp} and M_{cov}(*authors' note: M refer to model*)" [31]" Thus, the change from prior odds $p(H_0)/p(H_1)$ to posterior odds $p(H_0\|D)/p(H_1\|D)$ brought about by the data is given by the ratio of $p(D\|H_0)/p(D\|H_1)$, a quantity known as the *Bayes factor*"
QRIP2 - Not specifying Null and alternative hypotheses	N/A (these papers simply did not specify the null and the alternative hypotheses)	[1] "H_0 (that there is no peak-end rule for small positive experiences) and H_1 (that there is a peak-end rule for small positive experiences) We chose a one-sided prior (reflecting our prediction of a positive peak-end difference in one direction) with a width of $r = .707$" [2] "For testing the first three hypotheses, the model of H_1 used in the Bayesian analyses was a half-normal, with a mode of zero and the standard deviation equal to an expected raw effect size of 0.66"
QRIP3 - Simplifying the use of the prior model	N/A(these papers simply did not mention the prior model nor its justification.)	[35]: "the prior distribution as a truncated gaussian with $\mu = 1$ and $\sigma = 2$ (…) with a lower bound of the distribution at 0 and an upper bound at 2…" and "These values have been chosen taking into account the typical range of effects observed in previous studies" [63] "Effect sizes used to set priors were obtained from related results of previous studies" [42] "Typically, researchers entertain a distribution of possible alternatives to the null, some of which are judged more likely than others. For example, we may anticipate a high probability of a small-to-medium effect, a moderate probability of a large effect, and a very remote chance of a gargantuan effect. (…) The prior distribution, therefore, influences the BF, and the choice of a prior distribution is at the researcher's disposal; it may be based on knowledge about likely values established in previous work, or it may be chosen to be minimally informative".
QRIP4 - Not mentioning comparison of models	[2] "A Bayesian test of whether CRT scores were positively correlated with the resit effect yielded a correlation of r = .34 and a Bayes factor of 6.1, indicating that the data produced strong evidence in favor of the existence of a positive relation between CRT scores and the effect of resit exams on study-time investment" [10] "For example, a BF of 12 means that the data provides strong evidence for the alternative hypothesis." [11] "No effect was found for use actions (a 0.5% difference, $F < 1$; Bayes factor of 3.4 favoring the null model)." [21] "Finally, the proportion of participants choosing the best option for the current context increased during the task ($r = 0.72$, $p < 0.01$, $BF_{10} = 263.2$, see Figure 2a)"	[28] "..in a Bayesian test of mean differences we obtained a Bayes factor of 2.338 in favor of the null hypothesis over the two-sided difference hypothesis" [72] "The Bayes factor BF_{+0} quantifies the evidence that the data provide for H_+ (i.e. the presence of the compensatory control effect) relative to H_0 (i.e. the absence of the compensatory control effect)". [72] "This means that the data are about 5.41 times more likely under the null model including only religiosity, *compared to* the alternative model that also includes the control-threat manipulate" [1] "we obtained a BF_{01} of 3.21, indicating moderate evidence for H_0 relative to H_1"

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

	Presence of QRIP	Absence of QRIP
QRIP5 - Absolute statement	[22] "Of the 24 resulting comparisons, four produced a Bayes factor greater than 1 (i.e., providing evidence for an effect of initial category structure).(…). The other three comparisons *showed a difference* between the one-dimensional linear boundary and the other initial category structures for the Shepard circles" [41] "Both frequentist and Bayesian analyses revealed no main effects for responses of 'environment,' errors on SART, and no interaction for any of the measures studied". [3] "As in the t-test, the prior was specified as a Cauchy distribution. There was a significant difference; BF10 > 1,000" [11] "No effect was found for use actions (a 0.5% difference, F < 1; Bayes factor of 3.4 favoring the null model)." [25] "…but there was no interaction, F(4.41,127.85) = 1.860, MSE = 0.011, η^2_p = 0.060, p = 0.115, BF_{01} = 20.654"	(see the examples of absence of QRIP4)
QRIP6 - Using BF as posterior odds	[38] "The Bayesian test for correlations resulted in a Bayes Factor of BF_{01} = 5.88, indicating that the null hypothesis (i.e., absence of correlation) is ~6 times more likely than the alternative hypothesis (i.e., presence of correlation)" [47] "A model with an effect of soundtype was also more likely than the null model of no difference in soundtype (BF_{10} > 100,000). Finally, a model in which these factors interact was more likely than a model with no interaction (BF_{10} = 4.95)" [3] "Bayes factors showed that the alternative hypothesis with scores lower than chance was over 1,000 times more likely given the data" [21] "A hierarchical Bayesian t-test revealed that the alternative hypothesis of performing better than chance was BF_{10} = 53.88 more likely than the null hypothesis of chance performance"	(see the examples of absence of QRIP4)
QRIP7 - Considering BF as Effect size	[54] "As both the confidence interval and the Bayes factor do not point towards a true difference and the t-test is borderline significant, this can be considered a very small or non-existent effect." [15]: "A repeated-measures ANOVA with Cue (novel; standard) and Electrode (Fz; Cz; Pz) as factors showed that there is substantial evidence for a null effect (BF_{01} = 5.676), suggesting that the effect of expectations on the N2 was not very strong." [16] "Both a Pearson's correlation and a Bayesian correlation analysis revealed no strong relationship between the difference in relative change from pre-drink to post-drink between alcohol and placebo for the saccadic task and the manual tasks (r = 0.197, p = 0.22, BF = 0.41)" [32] "Bayesian statistics again supported that sexual motivation had a strong effect on attractiveness ratings in the control condition ($Bf_{1,0}$ = 25.99)" [59] "The target-absent ANOVA results suggest our manipulations did not strongly affect accuracy in the dual feature difference trials (BF = 3.59 for the most likely model, which included a main effect of color dissimilarity, against the next most likely model, which did not include any factor other than subject; BF = 15.29 against the third	N/A

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

	Presence of QRIP	Absence of QRIP
	most likely model, which included main effects of color dissimilarity and shape." dissimilarity; BF = 54.63 against the main effect of shape dissimilarity; BF = 227.42 against the interaction of color dissimilarity and shape dissimilarity)"	
QRIP8 - Mismatch between the research hypothesis and the statistical hypothesis	[10] conducted a study about the effect of nicotine on pupil size and hypothesized that "… pupil size would be smaller after the administration of nicotine, as compared to placebo…". They found that "Results suggested that pupil size during target assignment in the nicotine condition (M = 4.553, SD = 0.782) was significantly smaller than in the placebo condition (M = 4.681, SD = 0.827), F(1, 28) = 8.232, p = .008, η^2_p = 0.227, $BF_{incl.}$ = 1.316e+6 (Drug as the best model)", and concluded, "Results from both NHST and Bayesian analyses suggested that pupil size in the nicotine condition was smaller than pupil size in the placebo condition" [16] conducted a study for investigating "… whether alcohol affects saccade countermanding; i.e., whether saccadic SSRT is lengthened during acute intoxication". However, model comparison by ANOVA was conducted.	N/A

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/36357-on-the-potential-mismatch-between-the-function-of-the-bayes-factor-and-researchers-expectations/attachment/92134.docx?auth_token=PyCeVDyvse0pWhvu8b0E

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36357/718907/collabra_2022_8_1_36357.pdf by guest on 22 February 2024