UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Haucke, M., Miosga, J., Hoekstra, R., & van Ravenzwaaij, D. (2021). Bayesian Frequentists: Examining the Paradox Between What Researchers Can Conclude Versus What They Want to Conclude From Statistical Results. *Collabra: Psychology*, *7*(1). https://doi.org/10.1525/collabra.19026

Methodology and Research Practice

Bayesian Frequentists: Examining the Paradox Between What Researchers Can Conclude Versus What They Want to Conclude From Statistical Results

Matthias Haucke[1] [a], Jonas Miosga[2], Rink Hoekstra[3], Don van Ravenzwaaij[2]

[1] Department of Clinical Psychology and Psychotherapy, Freie Universität Berlin, Berlin, Germany, [2] Department of Psychometrics and Statistics, University of Groningen, Groningen, the Netherlands, [3] Department of Pedagogical and Educational Sciences, University of Groningen, Groningen, the Netherlands

Keywords: nhst, bayesian statistics, meta-science, statistical fallacy, statistical significance

https://doi.org/10.1525/collabra.19026

Collabra: Psychology

Vol. 7, Issue 1, 2021

A majority of statistically educated scientists draw incorrect conclusions based on the most commonly used statistical technique: null hypothesis significance testing (NHST). Frequentist techniques are often claimed to be incorrectly interpreted as Bayesian outcomes, which suggests that a Bayesian framework may fit better to inferences researchers frequently want to make (Briggs, 2012). The current study set out to test this proposition. Firstly, we investigated whether there is a discrepancy between what researchers *think* they can conclude and what they *want* to be able to conclude from NHST. Secondly, we investigated to what extent researchers want to incorporate prior study results and their personal beliefs in their statistical inference. Results show the expected discrepancy between what researchers think they can conclude from NHST and what they want to be able to conclude. Furthermore, researchers were interested in incorporating prior study results, but not their personal beliefs, into their statistical inference.

1. Introduction

Null hypothesis significance testing (NHST) is used in most scientific disciplines, including Psychology (Rucci & Tweney, 1980), Economics (McCloskey & Ziliak, 1996) and Medical Sciences (Chavalarias et al., 2016; Goodman, 1999). In NHST, an alternative hypothesis (for example, there is a mean difference between a treatment group and a control group) is tested against a null hypothesis (for example, there is no mean difference between a treatment group and a control group). The measured test statistics (e.g., *t*-statistics, *F*-statistics) indicate the difference between one's data and the null model prediction. The philosophical underpinning of NHST is called frequentism, and allows researchers to draw conclusions that are based on the average performance of these test statistics for a hypothetical infinite repetition of experiments. Thus, a *p*-value is the probability of obtaining the observed test statistics or more extreme ones, assuming the model assumptions (e.g., linearity, independence) are met and the null hypothesis is true (Greenland et al., 2016).

Despite the central role of NHST in the scientific process, the results from these techniques are misinterpreted by a majority of statistically educated scientists (Falk & Greenbaum, 1995; Haller & Krauss, 2002; Hoekstra et al., 2014; Lyu et al., 2020; Oaks, 1986). For instance, Oaks (1986) presented a scenario to Psychology researchers and students and asked them about their endorsement of six false statements regarding a significant *p*-value (see Table 1). These statements were: (1) You have absolutely disproved the null hypothesis; (2) You have found the probability of the null hypothesis being true; (3) You have absolutely proved your experimental hypothesis; (4) You can deduce the probability of the experimental hypothesis being true; (5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision; and (6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

None of the six presented statements are valid interpretations of a significant *p*-value.[1] On average, 2.5 incorrect

a matthias.haucke@gmx.de

1 A *p*-value quantifies the probability of obtaining the data at hand or more extreme data given that the null hypothesis is true. Statements 1 and 3 are false, because nothing can be proved absolutely. Statements 2 and 4 are false, because *p*-values do not speak to the probability of either the null or the alternative hypothesis being true. Statement 5 is false, because the probability of making a wrong decision would be the probability of the null hypothesis being true (Statement 2). Statement 6 is false, because reliably obtaining qualitatively similar results would imply the *p*-value relates to the probability of the alternative being true, which is not the case.

Table 1. The six presented statements taken from Haller & Krauss (2002).

Number	Statement
1	You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).
2	You have found the probability of the null hypothesis being true.
3	You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
4	You can deduce the probability of the experimental hypothesis being true.
5	You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
6	You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

statements were endorsed independent of the participants' statistical background. In a subsequent study, these findings were replicated when participants were given a prior explanation of the correct interpretation of NHST (Falk & Greenbaum, 1995). Misinterpretations of NHST were even found among statisticians and researchers who are teaching statistics (Haller & Krauss, 2002; Lecoutre et al., 2003). In addition, these misconceptions of NHST have been found in numerous statistical textbooks (Gigerenzer, 2004), such as *Introduction of Statistics for Psychology and Education*:

> [a significant test result] .. *is the probability that an observed difference is real....if the probability is low, the null hypothesis is improbable*.. (Nunnally et al., 1975, p. 194)

The most common misconception, which was endorsed by 73% of methodology instructors and 68% of psychology students, was "*You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision*" (Haller & Krauss, 2002). Thus, a majority of statistically educated researchers misinterpret frequentist statistics as the probability of one's hypothesis being true. That is, researchers misinterpret *p*-values as the probability of some hypothesis being true given the data (P(H0|D)), when in fact *p*-values are the probability of the observed data, or data more extreme, given the hypothesis (e.g., a right tailed test P(⩾D|H0)). This is a crucial difference, as a given set of data might lead to the rejection of the null hypothesis but could also lead to the rejection of some alternative hypothesis (Wagenmakers et al., 2011). The probability of the alternative hypothesis cannot explicitly be taken into account in a frequentist framework, but it can be in a Bayesian framework (Dienes, 2011; Etz & Vandekerckhove, 2018; Falk & Greenbaum, 1995).

The relevance of directly testing the alternative hypothesis can be illustrated by the court case of Sally Clark (see e.g., Rouder et al., 2016; Wagenmakers et al., 2018). Sally Clark was accused of murdering her two children. The judge and jury had to weigh two competing explanations for the death of the two infants: either Sally Clark murdered her own children, or both children died from sudden infant death syndrome (SIDS). Based on the estimate that the probability that SIDS would happen twice in a row was is roughly 1 in 73 million, Sally Clark was convicted for double murder. This exemplifies NHST reasoning, in which only a single hypothesis taken into account (H0: the children died because of SIDS), and an alternative is accepted by default if this single hypothesis is found to be unlikely. As H0 was very improbable, the jury did not accept the explanation of SIDS occurring twice. However, a crucial comparison was overlooked: the chances of multiple SIDS against the chances of multiple homicide. Only looking at the probability of two SIDS in a row is not informative, as the two deaths being the result of a mother murdering her two infant children, is not taken into account (and, as it happens, this explanation is even less likely). From a Bayesian perspective, what matters is the *relative* likelihood of both hypotheses. According to this reasoning, the relative plausibility of SIDS versus murder given the two infant children dying actually favors SIDS as an explanation by a factor 9 (see Hill, 2005, for details and assumptions for this calculation).

Despite considerable statistical training, many scientists revert to the habit of misinterpreting the conditional probabilities of the data given a certain hypothesis as the probability that the hypothesis is true. In order to be able to make such claims, statistical inference following a Bayesian framework is necessary (Gigerenzer, 2004). For an additional list of p-value misinterpretations, we refer the interested reader to Badenes-Ribera and colleagues (2015). The goal of the present study is to delve into the reasons why researchers incorrectly endorse (some of) these statements. Could it be that researchers *want* to be able to make these statements once they obtain a positive result? Are researchers perhaps Bayesians at heart?

In Bayesian statistics, the prior p(H) is combined with the likelihood p(D|H) to arrive at a posterior belief p(H|D). By dividing likelihoods of two rival hypotheses, it is possible to calculate the relative probability of the data under each hypothesis. The ratio of the probability of the data given the alternative hypothesis and the probability of the data given the null hypothesis is called a Bayes factor. The Bayes factor is the Bayesian way of quantifying statistical evidence and it is quite different from *p*-values. A Bayes factor can quantify how much our observed data shifts the balance of evidence from one hypothesis (e.g., the null hypothesis H0) to another (e.g., the alternative hypothesis H1; for more details see Dienes, 2011):

$$\frac{P(H1|D)}{P(H0|D)} = \frac{P(D|H1)}{P(D|H0)} \times \frac{P(H1)}{P(H0)}$$

where the quantity on the left is called the posterior odds, the quantity on the right is called the prior odds, and the quantity in the middle is the *Bayes factor* (BF_{10}). The subscript indicates that the Bayes factor quantifies the relative evidence provided by the data for the alternative hypothesis (H1) and the null hypothesis (H0). The Bayes factor can be

thought of as an updating factor, it reflects the change in belief about the relative likelihood of two hypotheses after the data has been observed (see e.g., van Ravenzwaaij & Etz, 2020).

The BF is directly interpretable in relation to one's hypothesis. For example, a BF_{10} = 19 means that the alternative hypothesis is 19 times more probable than the null hypothesis given the data (Rouder et al., 2009). In addition to being able to directly test one's hypotheses against each other, there are other advantages of using Bayes factors, such as being able to quantify evidence in favor of the null hypothesis, being able to employ sequential testing without the need to correct for multiple testing, and being able to compare strength of evidence across different studies (van Ravenzwaaij & Wagenmakers, 2019; Wagenmakers et al., 2018).

There are two methods to define one's prior probability: a subjective and an objective approach (Wagenmakers, 2007). In a subjective Bayesian approach, one incorporates their own beliefs about a possible parameter or hypothesis into the prior probability distribution, possibly informed by prior study results (for more details see Kruschke, 2014). In an objective Bayesian approach, one uses a predefined prior instead. One possibility is to use a default prior distribution that is comparatively uninformative in the sense that it allocates probability density to a wide range of possible parameter values (Rouder et al., 2009).

It has been proposed that current frequentist statistical practice is beset by a difference between what it can provide and what researchers desire from them (Morey et al., 2016, but see Lakens, 2019). Combined with the abundance of misinterpretations of frequentist statistics, perhaps a Bayesian framework might provide researchers with a more appropriate tool for conducting statistical inference (Gigerenzer, 2004, 2018). In this study, we attempt to find empirical evidence for these claims.

In a first task, we examine whether there is a difference between what researchers *think* they can conclude from statistical results and what they *want* to conclude from them. First, we tested whether researchers endorse false statements regarding NHST. Second, we investigated how much researchers would like to be able to make such statements after conducting statistical inference. Demonstrating a discrepancy between what traditional tests do and what researchers want them to do shows that the standard tools for statistical inference do not (completely) match the researchers' needs. Our study goes one step further and attempts to demonstrate not only the mismatch between what traditional tests do and what researchers want them to do, but also researchers' awareness of this mismatch. We expect that researchers will score higher on the items indicating they *want* to be able to draw the conclusions in the six statements than on the items indicating they *can* draw the conclusions in the six statements.

In a second task, we examined to what extent researchers think they typically incorporate two types of subjective priors, their own beliefs and prior study results, into their statistical analyses, and to what extent they *want* to incorporate these two types of subjective priors into their statistical analyses. We hypothesize that researchers feel uncomfortable with using their own beliefs because it may lead to different results depending on the person, thus losing the appearance of objectivity.

2. Methods

Participants

The study was emailed to the corresponding authors of all articles published in 2015, 2016, and 2017 in the following journals:

1. *Journal of Experimental Psychology: General*
2. *Psychological Science*
3. *Journal of Abnormal Psychology*
4. *Journal of Consulting and Clinical Psychology*
5. *Journal of Experimental Social Psychology*
6. *Journal of Personality and Social Psychology*

These journals were chosen to represent a sample of researchers in diverse fields of psychology (experimental, social, neuro-, and clinical), a sampling strategy previously used by Cramer et al. (2016). Participants who did not respond after two weeks received a reminder. After checking for duplicates and invalid email addresses, 1282 unique addresses were left. In total we obtained 117 participants for a response rate of 9%. Based on our preregistered power calculations, this is well enough to reliably obtain a Bayes factor higher than 10 for an underlying effect size of 0.5, but is a bit short for reliably obtaining a Bayes factor higher than 10 for an underlying effect of 0.2 (see https://osf.io/r75qd/). Respondents indicated to be faculty member (55%), graduate student (18%), post-doctorate (12%), external researcher (12%), and other (3%).

Materials and Procedure

This study made use of a within-subjects design, consisting of two tasks that required answering a multiple-choice questionnaire. After indicating their academic position, participants proceeded with the tasks. The verbatim email text and the questionnaire can be found at https://osf.io/r75qd/.

In the first task participants were asked to read a hypothetical research scenario and statistical results ("Please carefully read the text. Afterwards, indicate for each of the statements your confidence that it is true or false. "False" means that the statement does not follow logically from the information above. Also note that several or none of the statements may be correct.

Suppose you have a treatment that you suspect may reduce symptoms of migraine. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means t-test and your result is (t(18) = 2.7, p = 0.01).").[2]

2 Note that due to experimenter error, the df reads 18 instead of 38. The discrepancy was commented on by a single participant. As both

Afterwards, participants were asked to indicate their confidence in the truthfulness of six statements, taken from Haller & Krauss (2002; see Table 1), about the observed statistical results from a scale from 1 (confident it is false) to 9 (confident it is true). Moreover, participants were asked to indicate how much they *would like* to make each statement, on a scale from 1 (not at all interested) to 9 (very interested) ("For all following six statements shown, we would like to ask you to indicate to what extent you would want to be able to draw this kind of conclusion, provided the statistical techniques you use would be suitable to draw this kind of inference. In other words: is this a useful kind of statement for a researcher to make, provided they are able to? If this statement is true or false, does not matter here. Keep in mind that there are no right or wrong answers.").

In a second task, we asked researchers how they would act in a hypothetical scenario as a proxy for what researchers typically do. Participants were presented with the following text: *"Suppose you are about to study the effect of a new drug against depression. A former study with a drug similarly synthesized showed moderate to strong effects in the treatment of depression. However, due to your knowledge about the development of the drug you are suspicious and have strong beliefs against the drug's efficiency."* Afterwards participants were asked to indicate to what extent they typically take into account (1) results from a previous study ("To what extent do you typically take into account the previous study in your statistical analysis?") and (2) their own beliefs in a statistical analysis ("To what extent do you typically take into account your belief in a statistical analysis?"), on a scale from 1 (never) to 9 (always). Finally, participants were asked to indicate to what extent they *would like* to take into account (1) results from previous studies and (2) their own beliefs in a statistical analysis, on a scale from 1 (not all at interested) to 9 (very interested) ("These questions are related to the scenario before. Suppose you do know a way how to incorporate existing information in your analysis. Please indicate below to whether you want to take into account results from a previous study and your own beliefs into any statistical analysis. Note that there are no right or wrong answers"). Prior to data collection, we preregistered this study, the preregistration document may be found at https://osf.io/r75qd/.

3. Results

There was no missing data and we did not exclude any participant. A visual inspection of the frequency distribution of the collected scores for Task 1 indicated that the scores did not seem to follow a normal distribution: Scores corresponding to confidence ratings about the truthfulness of the statement were heavily skewed to the right for each of the six statements (see Figure 1). This suggests that many researchers were confident that they cannot draw the proposed conclusion. In contrast, scores corresponding to participants' desire to be able to make each of the six statements were mostly skewed to the left (see Figure 2). This indicates that many researchers did wish to be able to draw these conclusions.

The histograms for Task 2 show that data for this task is less strongly skewed (see Figure 3).

Because of the violation of the normality assumption, we deviated from the preregistration document in which we specified we would conduct a Bayesian *t*-test. Instead, we conducted a non-parametric alternative to the Bayesian *t*-test: the Bayesian Signed Rank Sum test (van Doorn et al., 2017). Other than this deviation, the analyses were conducted as planned.

To quantify the statistical evidence, we computed Bayes factors. For the first and second task, we conducted a set of two-sided Bayesian Signed Rank Sum tests, each with a default folded Cauchy effect size prior width of $r = \sqrt{2}/2$ (i.e., 0.707 or "medium"; for details see Rouder et al., 2009). Gibbs sampling (Geman & Geman, 1984; van Ravenzwaaij et al., 2018) was used to sample from the posterior distribution 100,000 times. According to the classification scheme suggested by Jeffreys (1998), we considered a Bayes factor of 10 (in favor or against the alternative hypothesis) as strong evidence and $1/3 \leqslant BF \leqslant 3$ as inconclusive evidence. In the present context, Bayes factors quantify the relative likelihood of the data under the two-sided alternative versus the likelihood of the data under the null hypothesis. Our analysis for Task 1 resulted in six Bayes factors, each of which quantifies whether there is a discrepancy between what respondents think can be concluded and what they would like to be able to conclude.

For the second task, the data is a range of scores from 1 to 9 on two statements which quantify whether respondents typically take into account their own prior beliefs and/or prior study results and whether they would like to take into account their own prior beliefs and/or prior study results. The two resulting Bayes factors quantify evidence for the discrepancy between what respondents typically take into account and what they would like to take into account. Bayes factors pointing towards the null hypothesis are indicative of no discrepancy, whereas Bayes factors pointing towards the alternative hypothesis are indicative of a discrepancy. All planned analyses and the associated R-code can be found in our preregistration document at https://osf.io/r75qd/.

In Task 1, most participants were confident to varying degrees that the statements were false, as indicated by the distributions of scores and the means that were well below 5 (see Figure 1). In comparison to the results of Haller & Krauss, the performance on interpreting *p*-values (Statement 5: You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision) appears to have increased. Similar to Haller & Krauss Statement 4 (with mean 3.17), 5 (mean = 3.81) and 6 (mean = 3.57) were more often rated as correct than Statement 1

values for the df lead to a p-value of 0.01, the critical quantity for purposes of this study, we do not think this error has affected the study outcomes. *Interestingly this error was also made in the original study by Haller & Krauss (2002) and has since been repeated by several authors (van Ravenzwaaij et al., 2019).*

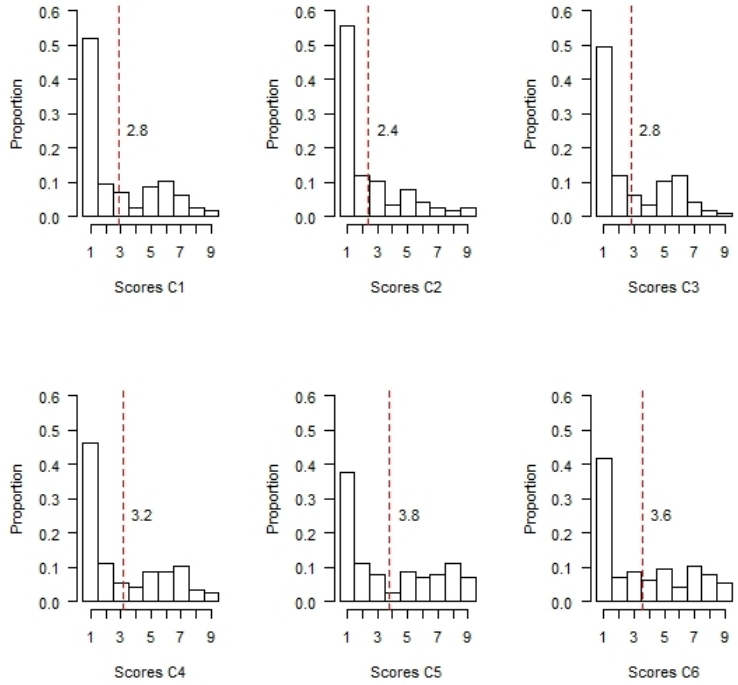

***Figure 1*. Proportion of scores for confidence ratings about truthfulness (C) for Statements 1 through 6 (Task 1). Red dashed vertical lines indicate mean score proportion.**

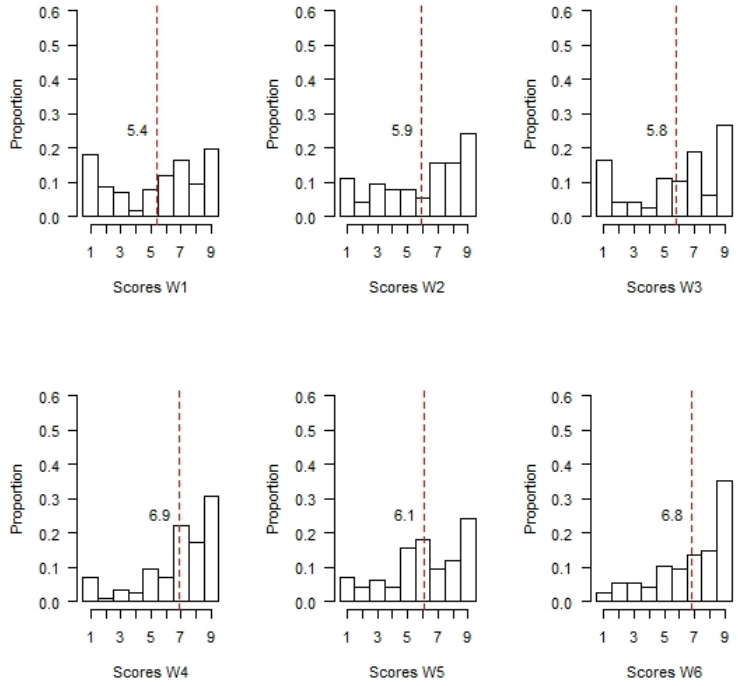

***Figure 2*. Proportions of scores for wanting (W) to be able to make Statements 1 through 6 (Task 1). Red dashed vertical lines indicate mean scores.**

(mean = 2.84), 2 (mean = 2.43), or 3 (mean = 2.79). This is in line with Haller & Krauss, who found Statements 1-3 to be

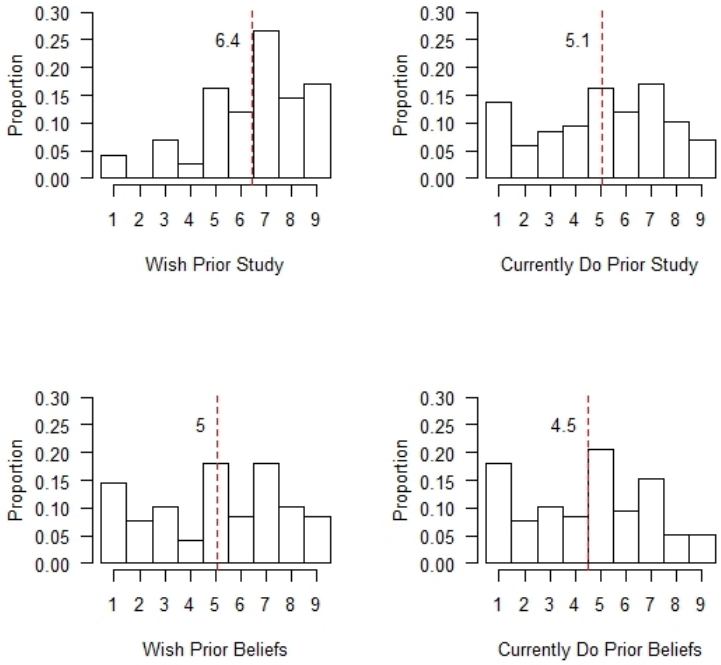

***Figure 3*. Proportion of scores on "Currently Do" and "Wish" questions about incorporating previous study results and personal beliefs in the analysis (Task 2). Red dashed vertical lines indicate mean scores.**

answered mostly correct (10-34% incorrect among Scientific Psychologists and Psychology students in comparison to Statement 4 (33%-59% incorrect) and Statement 5 (67-78% incorrect). In our data we see this trend as well, which suggests that Statements 1-3 might be inherently easier to flag as incorrect compared to the later statements about the probability of H1 (Statement 4) or probability of Type I errors (Statement 5). In line with our hypothesis that there is a difference between what researchers *think* they can conclude from statistical results and what they *want* to conclude from them, it can be seen that the means for *Wish* statements are generally higher than for *Can* statements. Note that the comparison between our study and Haller & Krauss's study can only be made indirectly: Whereas they looked at proportions of incorrect answers, we looked at the averages of the proportion of the entire scale.

Bayes factors for the first task can be found in Table 2. All six Bayes factors indicate overwhelming evidence in favor of there being a difference between what researchers think they can conclude and what researchers want to be able to conclude. Inspection of the means indicates that scores on what researchers want to be able to conclude are higher than what they think they can conclude for all six statements. In addition to the individual BFs for both tasks, we also calculate a meta-analytic Bayes Factor (Rouder & Morey, 2011), resulting in a high BF_{10} = 3.84e+89.

Bayes factors for the second task can be found in Table 3. The first Bayes factor indicates overwhelming evidence in favor of there being a difference in the extent to which people take into account prior study results and the extent they wish to do so. The second Bayes factor indicates almost no evidence for a difference in the extent to which people take into account a-priori beliefs and the extent to which they want to be able to. Inspection of the means indicate that scores on what researchers want to be able to do are higher than what they typically do in both cases.

Table 2. BF_{10} for Task 1.

Test (X vs. Y)	BF_{10}
W1vsC1	$2.84*10^6$
W2vsC2	$3.28*10^6$
W3vsC3	$2.37*10^6$
W4vsC4	$1.20*10^5$
W5vsC5	$6.00*10^5$
W6vsC6	$1.18*10^7$

Table 3. BF_{10} for Task 2.

Test	BF_{10}
Wish Study vs Presently Do Study	$4.75* 10^4$
Wish Belief vs Presently Do Belief	2.58

4. Discussion

NHST is without a doubt the most common variant of

hypothesis testing used in a majority of quantitative scientific disciplines (Goodman, 1999; Hoekstra et al., 2006; McCloskey & Ziliak, 1996; Rucci & Tweney, 1980). Despite considerable statistical training, many scientists misinterpret the outcomes of NHST (Falk & Greenbaum, 1995; Haller & Krauss, 2002; Lyu et al., 2020; Oaks, 1986). It has been proposed that these common misinterpretations might emerge because there is a discrepancy between what researchers can conclude and what they wish to conclude from their statistical analyses (Gigerenzer, 2004; Morey et al., 2016). The results of the current study suggest that this proposed discrepancy exists. However, the results show contradicting evidence for the notion that a Bayesian framework better describes how researchers want to use inferential statistics.

In Task 1, participants were presented with a research scenario investigating a treatment that may reduce symptoms of migraine including statistically significant results from a simple independent means *t*-test. On average, participants indicated that they wish to make each of these six statements, even though they realize they are not able to under NHST. Specifically, participants wish to "absolutely disprove the null hypothesis" (Statements 1) and wish to find the probability of the null hypothesis being true (Statement 2); they wish to absolutely prove the experimental hypothesis (Statement 3) and find the probability of the experimental hypothesis being true (Statement 4). Finally, participants wish to know, if they decide to reject the null hypothesis, the probability of making a wrong decision (Statement 5) and they wish to have a reliable experimental finding in the sense that if the experiment were repeated a great number of times, the obtained statistical quantity would inform the number of times a significant result would be obtained (Statement 6).

So would Bayesian statistics offer what researchers seem to want, based on the results of our study? Bayes factors quantify the relative likelihood of the data under one hypothesis (e.g. the null hypothesis) to another (e.g. the alternative hypothesis). As such, one can quantify evidence in favor of the alternative or in favor of the null hypothesis. This cannot be done within the frequentist NHST as only the null hypothesis is explicitly taken into consideration. Therefore, it is impossible to assign a probability to any hypothesis. Bayesian inference allows for some, but not all, of the previously presented conclusions to be drawn, given a prior belief. Specifically, Bayes factors enable drawing the conclusions from Statement 2 (We can find the probability of the null hypothesis being true, given a prior belief), Statement 4 (We can find the probability of the alternative hypothesis, given a prior belief). When used for making decisions (see e.g., Aczel et al., 2020), Bayes factors can be used to make Statement 5 (We can find the probability of making a wrong decision, when rejecting the null hypothesis, given a prior belief).

Although the conclusions from Statement 1 and Statement 3 cannot be drawn under either a frequentist or a Bayesian framework (we can never disprove the null or alternative hypothesis *absolutely*), we are able to assign a concrete number to the relative evidence in favor of one hypothesis over another provided by the data. As a result, a researcher is at liberty to decide that a relative likelihood of 1000 to 1 constitutes compelling evidence for one hypothesis over another (possibly influenced by their prior belief in the plausibility of both hypotheses) and collect data until the Bayes factor is either 1000 or 1/1000. Thus, contrary to NHST, Bayesian inference does allow one to draw conclusions that *approximate* those drawn in Statements 1 and 3.

Statement 6 is not applicable to a Bayesian framework, as significant results pertain to NHST specifically. However, if a researcher were to conduct a great number of experiments, and in each of these continuously samples evidence until a BF_{10} of 10 or 1/10 was reached, then they would hit BF_{10} =10 ten times as often as BF_{10} =1/10 if the alternative hypothesis were true and they would hit BF_{10} =1/10 ten times as often as BF_{10} =10 if the null hypothesis were true (Rouder, 2014; Tendeiro et al., 2019). And if they would do so until hitting a BF_{10} of 100 or 1/100 instead, they would hit BF_{10} =100 hundred times as often as BF_{10} =1/100 if the alternative hypothesis were true and they would hit BF_{10} =1/100 hundred times as often as BF_{10} =100 if the null hypothesis were true. As such, Bayes factors allow one to draw conclusions about the probability of drawing the wrong conclusion in the long run based on the strength of evidence. Put simply – given a great number of experiments – there is a link between the size of the BF and the expected proportion of BFs that provide evidence in favor of the correct model. However, a single Bayes factor does not allow for predicting, say, the size of the Bayes factor in a replication attempt.

The misfit between what researchers can conclude versus what they want to conclude from a frequentist test might explain the common misinterpretation of frequentist NHST in line with a more intuitive Bayesian interpretation (Gigerenzer, 2004; Haller & Krauss, 2002; Lyu et al., 2020; Oaks, 1986). The results of Task 1 indicate that, indeed, researchers wish to make inferences in line with a Bayesian framework, such as finding the probability of the null and alternative hypothesis given the data and prior belief.

In a second task, participants were presented with a hypothetical scenario about the effectiveness of a new drug against depression. We asked participants whether they would like to be able to incorporate previous findings, and their own personal beliefs into their statistical analyses. Our results indicate that many researchers feel partial to being able to specify their own prior distribution. On the one hand, participants seem to want the ability to incorporate prior study results into their analysis more than they are presently able to. On the other hand, our results are ambiguous with respect to a mismatch between researchers wanting to incorporate their own personal beliefs into their statistical inference versus their ability to do so, which suggests that they might be more comfortable with a subjective prior based on previous research or an objective prior.

Whether this is reflective of an actual preference, or because this is closer related to the techniques they are currently using is an open empirical question. It might well be that incorporating one's own beliefs into study outcomes might be seen as problematic given that objectivity is often propagated as a scientific virtue. We would like to point out though that subjective beliefs may well be informed by knowledge about previous study results, so the distinction between the two may not be as clear-cut.

Researchers' potential hesitance to take their subjective

beliefs into account in the analysis might be one of the reasons why Bayes is not used instead of NHST as often as one might expect given the amount of attention for such techniques. The most common Bayesian way of thinking about probability is to define the probability of an event as the degree of belief that we assign to the truth of an event. Thus, probabilities do not exist in the world, but rather in thought and assumptions of the researcher (Navarro, 2015), and are hence necessarily subjective. Our study indicates that this Bayesian definition of probability might be less appealing than the frequentist definition of probability, which in the context of test statistics is the average performance for an infinite repetition of hypothetical experiments. However, we should be cautious about concluding too much from these findings alone. Importantly, this result suggests that the exact way we arrive at a prior distribution is a crucial factor on whether researchers accept the Bayesian notion of subjectivity (i.e., if the prior distribution is well founded by previous studies it is accepted).

In our study, we could not replicate the finding that a majority of psychological researchers endorses false statements about *p*-values (Falk & Greenbaum, 1995; Haller & Krauss, 2002; Hoekstra et al., 2014; Oaks, 1986). One reason for this might be an increased awareness about the false use of statistics since the term p-hacking was coined (Simonsohn et al., 2014). As we obtained a low response rate (9%), we should be careful with the conclusions we can draw based on our study. Possibly, those who did participate were on average more knowledgeable in statistical inference than those who did not. If true, our participants might be more likely to know about the limitations of NHST and might feel more comfortable with Bayesian statistics, although nothing in the invitation email nor in the survey mentioned Bayesian statistics.

We asked participants about their willingness to include prior beliefs or studies in the context of a study analyzing the effect of new drugs against depression. This result might be limited to a clinical context, in which participants might be more careful in including their prior beliefs (as people's well-being is at stake), therefore we must be careful to generalize the outcome to other fields. Our recruited participants had a diverse background (experimental, social, neuro and clinical psychologists), thus non-clinical researchers might be hesitant to include prior beliefs in a clinical research context, as they simply did not feel qualified to do so. In our study, we did not ask participants to indicate their field of expertise, thus future studies would need to study whether being comfortable with including prior beliefs changes with the expertise a researcher has in the context of the research question. We also did not ask participants to report what kind of statistical inference they typically employ in their own work. As such, the results of the present study do not allow linking participants' responses to their own practices, but we believe this to be a fruitful avenue to explore in future studies.

Finally, future studies should include options for open answer responses. It might be interesting to learn from researchers under which circumstances they would like to incorporate a prior and what information they would like to base such a prior on.

General conclusion

Our study suggests that there is a gap between what researchers would like to conclude and what they can conclude from their statistical analyses. Researchers seem to be interested in making inferences in line with a Bayesian framework, such as finding the probability of the null and alternative hypothesis given the data and prior. The difference between what researchers can and want to conclude might be one of the explanations for the previously found misinterpretations of statistical results (Haller & Krauss, 2002; Lyu et al., 2020). However, researchers seem to be ambiguous about what information to include in their prior distribution. On the one hand, we did not find compelling evidence that researchers wish to take their own prior beliefs into account when analyzing their data. On the other hand, we did find overwhelming evidence to suggest that researchers are interested in including information from previous studies into their analyses. This suggests that the source from which the prior distribution is derived (i.e., outcomes of previous studies) is a crucial factor on whether researchers accept the Bayesian notion of subjective probability.

Our results should be interpreted with a great deal of caution. First of all, the response rate was low (9%), and there is no way to tell whether the remaining sample is still representative of researchers. The question of what researchers want to conclude from their statistical analysis is quite complex to answer, as it depends amongst other things on the field of study, their statistical education and on the specific research question. The used questionnaires can only give some indications about what people really want, so more detailed questionnaires or qualitative studies are needed.

Allowing for these caveats, our study allows for two conclusions: (1) there was a clear difference between what researchers think can be concluded versus what they would like to be able to conclude from the fictional results paragraph presented in our experiment, and (2) researchers seems to like being able to incorporate prior results but not personal beliefs into statistical inference. Thus, our study provides modest support for the notion that researchers sympathize with some elements native to the philosophy of objective Bayesianism. Moreover, a subjective Bayesian approach is deemed more acceptable if a prior distribution is built upon previous research.

Contributions

MH drafted the manuscript, prepared the material, designed the study and analyzed the data; JM prepared the material, designed the study, gathered the data and analyzed the data; RH drafted the manuscript and designed the study; DR drafted the manuscript and designed the study.

Competing Interests

The authors declare no competing interests. Don van Ravenzwaaij is a Section Editor at Collabra: Psychology. He was not involved in the peer review of this article.

Data Accessibility Statement

All the stimuli, participant data, preregistration and analysis script can be found on the paper's project page on the open science framework: https://osf.io/r75qd/.

Submitted: April 10, 2020 PST, Accepted: January 20, 2021 PST

REFERENCES

Aczel, B., Hoekstra, R., Gelman, A., Wagenmakers, E.-J., Klugkist, I. G., Rouder, J. N., Vandekerckhove, J., Lee, M. D., Morey, R. D., Vanpaemel, W., Dienes, Z., & van Ravenzwaaij, D. (2020). Discussion Points for Bayesian Inference. *Nature Human Behaviour*, *4*(6), 561–563. https://doi.org/10.1038/s41562-019-0807-z

Badenes-Ribera, L., Frías-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2015). Interpretation of the p value: A national survey study in academic psychologists from Spain. *Psicothema*, *27*(3), 290–295.

Briggs, W. M. (2012). *It is time to stop teaching frequentism to non-statisticians* (arXiv:1201.2590). arXiv preprint.

Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of reporting P values in the biomedical literature, 1990-2015. *Jama*, *315*(11), 1141–1148. https://doi.org/10.1001/jama.2016.1952

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., Waldorp, L. J., & Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*(2), 640–647. https://doi.org/10.3758/s13423-015-0913-5

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

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*(1), 5–34. https://doi.org/10.3758/s13423-017-1262-3

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*(1), 75–98. https://doi.org/10.1177/0959354395051004

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721–741. https://doi.org/10.1109/tpami.1984.4767596

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine*, *130*(12), 995–1004.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

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

Hill, R. (2005). Reflections on the cot death cases. *Significance*, *2*(1), 13–16. https://doi.org/10.1111/j.1740-9713.2005.00077.x

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse ofp values. *Psychonomic Bulletin & Review*, *13*(6), 1033–1037. https://doi.org/10.3758/bf03213921

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

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press. https://doi.org/10.1016/b978-0-12-405888-0.00008-8

Lakens, D. (2019). *The practical alternative to the p-value is the correctly used p-value*. https://doi.org/10.31234/osf.io/shm8v

Lecoutre, M.-P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology*, *38*(1), 37–45. https://doi.org/10.1080/00207590244000250

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*. https://doi.org/10.1017/prp.2019.28

McCloskey, D. N., & Ziliak, S. T. (1996). The standard error of regressions. *Journal of Economic Literature*, *34*(1), 97–114.

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. https://doi.org/10.1016/j.jmp.2015.11.001

Navarro, D. (2015). *Learning statistics with R*. Lulu.com.

Nunnally, J. C., Durham, R. L., Lemond, L. C., & Wilson, W. H. (1975). *Introduction to statistics for psychology and education*. McGraw-Hill Book.

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

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

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*(4), 682–689. https://doi.org/10.3758/s13423-011-0088-7

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*(3), 520–547. https://doi.org/10.1111/tops.12214

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*(2), 225–237. https://doi.org/10.3758/pbr.16.2.225

Rucci, A. J., & Tweney, R. D. (1980). Analysis of variance and the "second discipline" of scientific psychology: A historical account. *Psychological Bulletin*, *87*(1), 166–184. https://doi.org/10.1037/0033-2909.87.1.166

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Tendeiro, J., Kiers, H., & van Ravenzwaaij, D. (2019). *Mathematical Evidence for the Adequacy of Bayesian Optional Stopping*. https://doi.org/10.31234/osf.io/9t2e7

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2017). *Bayesian Latent-Normal Inference for the Rank Sum Test, the Signed Rank Test, and Spearman's ρ*. arXiv.

van Ravenzwaaij, D., Albers, C., Derksen, M., & Hoekstra, R. (2019). Citing is easy, reading is hard. *Mindwise*. https://mindwise-groningen.nl/citing-is-easy-reading-is-hard/

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte-Carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143–154. https://doi.org/10.3758/s13423-016-1015-8

van Ravenzwaaij, D., & Etz, A. (2020). *Simulation Studies as a Tool to Understand Bayes Factors*. https://doi.org/10.31234/osf.io/27ndb

van Ravenzwaaij, D., & Wagenmakers, E.-J. (2019). *Advantages Masquerading as 'Issues' in Bayesian Hypothesis Testing: A Commentary on Tendeiro and Kiers (2019)*. https://doi.org/10.31234/osf.io/nf7rp

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

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., & … Epskamp, S. (2018). Bayesian inference for psychology. *Part I: Theoretical Advantages and Practical Ramifications. Psychonomic Bulletin & Review*, *25*(1), 35–57.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, *100*(3), 426–432. https://doi.org/10.1037/a0022790

SUPPLEMENTARY MATERIALS

Peer review history

Download: https://collabra.scholasticahq.com/article/19026-bayesian-frequentists-examining-the-paradox-between-what-researchers-can-conclude-versus-what-they-want-to-conclude-from-statistical-results/attachment/50725.docx?auth_token=OJroRmrN9U1jxlO-zUbJ