Chapter 12

Bayesian Hypothesis Testing: An Alternative to Null Hypothesis Significance Testing (NHST) in Psychology and Social Sciences

Alonso Ortega and Gorka Navarrete

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.70230

Abstract

Since the mid-1950s, there has been a clear predominance of the Frequentist approach to hypothesis testing, both in psychology and in social sciences. Despite its popularity in the field of statistics, Bayesian inference is barely known and used in psychology. Frequentist inference, and its null hypothesis significance testing (NHST), has been hegemonic through most of the history of scientific psychology. However, the NHST has not been exempt of criticisms. Therefore, the aim of this chapter is to introduce a Bayesian approach to hypothesis testing that may represent a useful complement, or even an alternative, to the current NHST. The advantages of this Bayesian approach over Frequentist NHST will be presented, providing examples that support its use in psychology and social sciences. Conclusions are outlined.

Keywords: Bayesian inference, Bayes factor, NHST, quantitative research

1. Introduction

> *"Scientific honesty then requires less than had been thought: it consists in uttering only highly probable theories: or even in merely specifying, for each scientific theory, the evidence, and the probability of the theory in the light of this evidence".* Lakatos[1, p. 208].

The nature and role of experimentation in science found its origins in the rise of natural sciences during the sixteenth and seventeenth centuries [2]. Since then, knowledge meant that theories have to be corroborated either by the power of the intellect or by the evidence of the senses [1]. However, until the mid-late 1800s, "psychological experiments had been performed, but the science was not yet experimental" [3, p. 158]. It was not until 1875 that—either at

Wundt laboratory in Leipzig or at James' laboratory in Harvard—experimental procedures were introduced and contributed to the development of psychology as an independent science [3]. From almost one and a half centuries, scientific research mostly relies on empirical findings to provide support to their hypotheses, models, or theories. From this point of view, psychology and social sciences must take distance from rhetorical speculations, desist from unproven statements and build its knowledge on the basis of empirical evidence [1, 4]. Almost a decade ago, Curran reemphasized that the aim of any empirical science is to pursue the construction of a cumulative base of knowledge [5]. However, it has also been emphasized that such a cumulative knowledge—for a true psychological science—is not possible through the current and widespread paradigm of hypothesis testing [5–9]. Since approximately two decades ago, some explicit claims have appeared in peer review articles, such as "*Psychology will be a much better science when we change the way we analyze data*"[7], *"We need statistical thinking, not statistical rituals"* [10], "*Why most research findings are false*" [11] or "*Yes, psychologists must change the way they analyze their data…"* [12]. Most critiques have been directed toward the current—and still predominant—approach to hypothesis testing (i.e., NHST) and its overreliance on *p-values* and *significance levels* [6, 11, 13], emphasizing its pervasive consequences against the construction of a cumulative base of knowledge in psychological science [8]. Despite all warnings, they seem not to have generated a noteworthy echo in the scientific community, even though "it is evident that the current practice of focusing exclusively on a … decision strategy of null hypothesis testing can actually impede scientific progress" [14, p. 100]. Therefore, it seems reasonable to suggest that there is a need to make considerable changes to how we usually carry out research, especially if the goal is to ensure research integrity [6]. Regarding this matter, a frequently proposed alternative has been moving from the exclusive focus on *p-values* to incorporate other existing techniques such as "power analysis" [15] and "meta-analysis" [16], or to report and interpret "effect sizes" and "confidence intervals" [7]. However, in our view, a sounder alternative would be to move from a Frequentist paradigm to a Bayesian approach, which allows us not only to provide evidence against the null hypothesis but also in favor of it [17]. Furthermore, Bayesian analysis allows us to compare two (or more) competing models in light of the existent data and not only based in "theoretical probability distributions," as in the Frequentist approach to hypothesis testing [18].

A Bayesian approach would offer some interesting possibilities for both individual psychology researchers and the research endeavor in general. First, Bayesian analysis allows us to move from a dichotomous way of reasoning about results (e.g., either an effect exists of it does not) to a less artificial view that interprets results in terms of magnitude of evidence (e.g., the data are more likely under H_0 than H_a), and therefore, allows us to better depict to which extent a phenomenon may occur. Second, a Bayesian approach naturally allows us to directly test the plausibility of both the null and the alternative hypothesis, but the current NHST paradigm does not. In fact, when a researcher does not reach a desired *p-value* oftentimes it is—falsely—assumed that the effect "does not exist." As a consequence, the researcher's chances of getting his or her results published decrease dramatically, which moves us to our third argument. As broadly known, the most scientific peer-reviewed journals do not show much interest in results, which are "non-statistically significant." This common practice—or scientific standard—sadly reinforces the idea of thinking in terms of relevant or irrelevant findings. In our view,

such standards do not promote scientific advance and quickly lead us to ignore some promising but "non-significant" findings that may be further explored, fed into meta-analysis, of just be considered by other researchers in the field. Of course, systematically ignoring a portion of the research undermines the primary goal of scientific inquiry that is to collect evidence and not only to reject hypothesis. The facts and ideas exposed in this introductory section set forth the necessity to reanalyze the way in which scientific evidence has been conceived during the NHST era.

The following sections will: (a) concisely address the NHST procedure, (b) introduce a Bayesian framework to hypothesis testing, (c) provide an example that highlights the advantages of a Bayesian approach over the current NHST in terms of the way in which scientific evidence is quantified, and (d) briefly summarize and discuss the benefits of a Bayesian approach to hypothesis testing.

2. Null hypothesis significance testing (NHST)

> "*Never use the unfortunate expression: accept the null hypothesis.*" Wilkinson and the Task Force on Statistical Inference APA Board of Scientific Affairs[19, p. 602].

The most influential methods to modern null hypothesis significance testing (NHST) were developed by Fisher, and by Neyman and Pearson in the early and mid-1900s [20]. Since then, the NHST has been broadly used to provide an association between empirical evidence and models or theories [21]. In the traditional NHST procedure, two hypotheses are postulated: a null hypothesis (i.e., H_0) and a research hypothesis, also called alternative (i.e., H_a), which describe two contrasting conceptions about some phenomenon [22]. When conducting a NHST, researchers usually pursue to reject the null hypothesis (H_0) on the basis of a *p-value*. When the observed *p-value* is lower than a predetermined significance level (i.e., alpha, usually corresponding to $\alpha = 0.05$), the conclusion is that such *p-value* constitutes supporting evidence that favors the plausibility of the alternative hypothesis [23]. However, a more important feature of this procedure that remains unknown for most scientists, including psychology researchers, is that the NHST constitutes an amalgamation of two irreconcilable schools of thought in modern statistics: the Fisher test of significance, and the Neyman and Pearson hypothesis test [24, 25]. To this respect, Goodman stated that "it is not generally appreciated that the *p-value*, as conceived by Fisher, is not compatible with the Neyman and Pearson hypothesis in which it has become embedded" [25, p. 485]. In this synthesized NHST, the Fisherian approach includes a test of significance of *p-values* obtained from the data, whereas the Neyman and Pearson method incorporates the notion of error probabilities from the test (i.e., Type I and Type II).

2.1. Origins and rationale of NHST

First, in the early 1900s, Fisher [26, 27] developed a method that tested a single hypothesis (i.e., null or H_0), which has been mainly referred to as a hypothesis of "no effect" between variables (e.g., relationship, difference). The null hypothesis, as conceived by Fisher, has a known

distribution of the test statistic t. Thus, as the test statistic moves away from its expected value, then the null hypothesis becomes progressively less plausible. In other words, it appears less likely to occur by chance. Then, if H_0 achieves a probability of occurrence sufficiently lower than the significance level (i.e., a small *p-value*) then it should be rejected. Otherwise, no conclusion can be reached. Subsequently, the question that logically arises is: what *p-value* is sufficiently small to reject H_0? The answer to this question was clearly addressed by Fisher when he stated that this threshold should be determined by the context of the problem, and it was not until the 1950s that Fisher presented the first significance tables to establishing rejection thresholds [22]. However, Fisher [28] refused the idea of establishing a conventional significance level and, in its place, recommended reporting the exact *p-value* instead of a significance level (e.g., $p = 0.019$, but not $p < 0.05$; see [10]). Similarly, May et al. indicated that the choice of a significance level should depend on the consequences of rejecting or failing to reject the null hypothesis [29]. Despite these recommendations about threshold determination, most scientists from different research fields adopted standard significance levels (i.e., $\alpha = 0.05$ or $\alpha = 0.01$), which have been used—or misused—regardless of the hypotheses being tested.

Later, in 1933, Neyman and Pearson proposed a procedure in which two explicitly stated rival hypotheses were contrasted, being one of them still considered as the "null" hypothesis, as in the Fisher test [30]. Neyman and Pearson rejected Fisher's idea of only testing the null hypothesis. In this scenario, there are now two hypotheses (i.e., the null and the alternative), and based on the observed *p-value*, the researcher has to decide whether to reject or not to reject the null hypothesis. This decision rule faces the researcher with the probability of committing two kinds of errors: Type I and Type II. As defined by Neyman and Pearson, the Type I error is the probability of falsely rejecting H_0 (i.e., null) when H_0 is true [30]. Conversely, the probability of failing to reject H_0 when H_0 is false is the Type II error. For the sake of simplicity, an analogy of both kinds of errors can be found in the classic fairy tale "The boy who cried wolf!" When the young shepherd, called Peter, shouted out: "Help! the wolf is coming!" The village's people believed the young boy warning and quickly came to help him. However, when they found out that all was a joke, they got angry. To believe in the boy's false, alarm can be considered as a Type I error. Peter repeated the same joke a couple of times and, when the wolf actually appeared, the villagers did not believe the young shepherd's desperate calls. This situation is analogous to be engaged in a Type II error [31].

Within this NHST framework, the Fisher's *p-value* is then used to dichotomize effects into two categories: significant and non-significant results [21]. Consequently, on one hand, obtaining significant results led us to assume that the phenomenon under investigation can be considered as "existing" and, therefore, can be used as supporting evidence for a particular model or theory. On the other hand, non-significant results are usually (and erroneously) considered as "noise," implicating the nonexistence of an effect [21]. In this last case, there are no findings that could be reported. From this view, the evidence in favor of a research finding is then solely judged on the ability to reject H_0 when a sufficiently low *p-value* is observed. This simple and appealing decision rule may constitute a very seductive way of thinking about results, that is: A phenomenon either exists or it does not. However, thinking in this fashion is fallacious, led to misinterpretations of results and findings, and more importantly "it can distract us from a higher goal of scientific inquiry. That is, to determine if the results of a test have any practical value or not" [32, p. 7].

2.2. NHST: Common misconceptions and criticisms

As previously stated, most problems and criticisms to the current NHST paradigm appear as a result of the mismatch of these essentially incompatible statistical approaches [10, 33, 34]. In this line, Nickerson stated that "A major concern expressed by critics is that such testing is misunderstood by many of those who use it" [35, p. 241]. Some of these misconceptions are common among researchers and are interpretative in nature. As a matter of fact, Badenes-Ribera et al. recently reported the results of a survey conducted to 164 academic psychologists who were questioned about the meaning of *p-values* [36]. Results confirmed previous findings regarding the occurrence of wrongful interpretations of *p-values*. For instance, the false belief that the *p-value* indicates the conditional probability of the null hypothesis given certain data (i.e., $p\ (H_0|D)$), instead of the probability of witnessing a given result, assuming that the null hypothesis is true [37]. This wrong interpretation of a *p-value* is known as "the inverse probability" fallacy. Another common misconception regarding *p-value*s is that they provide direct information about the magnitude of an effect, that is, a *p-value* of 0.00001 represents evidence of a bigger effect than a *p-value* of 0.01. This conclusion is wrong because the only way to estimate the magnitude of an effect is to calculate the value of the effect size with the appropriate statistic and its confidence interval (e.g., Cohen's *d*; see [38]). This erroneous interpretation of a *p-value* is known as "the effect size" fallacy. A comprehensive review of these and other common misconceptions is out of the scope of this chapter, but several resources on these topics are available for the interested readers (see [14, 35, 37–40]).

Likewise, the rationale under the NHST has been largely criticized. Most criticisms against NHST are focused on the way in which data are (unsoundly) analyzed and interpreted, for example:

- **a.** NHST only provides evidence against the plausibility of H_0, but does not provide probabilistic evidence in favor of the plausibility of $H_{a.}$
- **b.** NHST uses inference procedures based on hypothetical data distributions, instead of being based on actual data.
- **c.** NHST does not provide clear rules for stopping data collection; therefore, as long as sample size increases any H_0 can be rejected (see [9, 18]).

However, an issue that is of particular interest for this chapter is related to the use of *p-values* as a way to quantify statistical evidence [13, 41]. As previously stated in this chapter, rejecting H_0 does not provide evidence in favor of the plausibility of H_a, and all that can be concluded is that H_0 is unlikely [9]. Conversely, failing to reject H_0 simply allows us to state that—given the evidence at hand—one cannot make an assertion about the existence of some effect or phenomenon [42]. Hence, rejecting H_0 is not a valid indicator of the magnitude of evidence of a result [43]. In Schmidt's words: "… reliance on statistical significance testing in psychology and the other social sciences has led to frequent serious errors in interpreting the meaning of data, errors that have systematically retarded the growth of cumulative knowledge" [16, p. 120]. Despite the existence of scientific literature that highlights the weaknesses of NHST [9, 16, 21, 22, 39, 43–46], it is still considered as the: "*sine qua non* of the scientific method" [10, p. 199]. Moreover, NHST is arguably the most widely used method of data analysis in psychology since the mid-1950s and still governs the interpretation of quantitative data in social science

research [35, 47]. In Krueger's words: "NHST is the researcher's workhorse for making inductive inferences" [45, p. 16]. An immediate matter of concern is that most of scientific discoveries, in a wide range of research fields, are based on a procedure that still generates controversy (see [12, 48–50]). Since the focus of research should be on what data tell us about the magnitude of effects, it seems necessary to shift from our reliance on NHST to more robust alternatives [14]. Some recommended practices include estimates based on effect sizes, confidence intervals, and meta-analysis [6]. However, a sounder alternative comes from the Bayesian paradigm through the use of a simple estimate of the magnitude of evidence called Bayes factor (BF) [17]. This approach to hypothesis testing has shown several benefits. First, it is not oriented to pursue the rejection of H_0; on the contrary, it provides a way to obtain evidence for and against H_0. Second, it does not use arbitrary thresholds (i.e., significance levels) to reach dichotomous decisions about the plausibility or implausibility of H_0; on the contrary, it directly contrasts the magnitude of evidence for and against both H_0 and H_a. Third, it permits the continuous update of evidence as long as new data are available, which is in line with the nature of scientific inquiry. Bayesian methods have been largely suggested as a practical alternative to NHST [9, 17, 23, 51], but—until now—they have not received enough attention from researchers in psychology and social sciences.

3. Bayesian hypothesis testing: An alternative to NHST

> *"(…) prior and posterior are relative terms, referring to the data. Today's posterior is tomorrow's prior."* Lindley[52, p. 301] .

In the field of statistics, probabilities can be interpreted under two predominant paradigms: Frequentist inference and Bayesian inference. The former makes predictions about experiments whose outcomes depend basically upon random processes [53]. The latter assigns probabilities to any statement, even when a random process is not involved [54]. In a Bayesian framework, a probability is a way to embody an individual's degree of belief in a statement. Since the mid-1950s, there has been a clear predominance of the Frequentist approach to hypothesis testing, both in psychology and social sciences. The hegemony of Frequentist inference and its null hypothesis significance testing (NHST) might be partially attributed to the massive incorporation of such approaches in psychology undergraduate programs [9] and also to the fact that the Neyman and Pearson approach had the most well-developed computational software to conduct statistical inference [18]. However, the current scenario has drastically changed, and the development of sampling techniques like Markov-Chain Monte Carlo (MCMC; see [55, 56]) along with the availability and improvement of specifically developed software (e.g., WinBUGS, see [57, 58]; JAGS, see [59, 60]; JASP, see [61]) makes exact Bayesian inferences possible even in very complex models. As a result, "Bayesian applications have found their way into most social science fields" [22, p. 665], and psychologists can now easily implement Bayesian analysis for many common experimental situations (see for example JASP Statistics: https://jasp-stats.org/).

3.1. Bayes in a nutshell

In Bayesian inference, our degrees of belief about a set of hypotheses are quantified by probability distributions over those hypotheses [47, 62], which makes the Bayesian approach fundamentally different from the Frequentist approach, which relies on sampling distributions of data [47]. A Bayesian analysis usually implicates the updating of prior knowledge or information in light of newly available experimental data [63]. The latter clearly reflects the aim of any empirical science, which is to strive for the elaboration of a cumulative base of knowledge. Any Bayesian analysis implies the combination of three sources of information as follows:

a. a model that specifies how latent parameters (e.g., θ) generate data (e.g., D);

b. prior information about those parameters (i.e., prior distribution); and

c. the observed data (i.e., likelihood).

This prior information, represented by $p(\theta)$, represents our degree of uncertainty about the parameters included in the model. Conversely, this prior distribution may also represent our degree of knowledge about the same parameters. Then, the more informative is our prior distribution, the less will be our degree of uncertainty about the parameters. The likelihood is the conditional probability of observing the data under some latent parameter (i.e., $p(D|\theta)$). Following the Bayes theorem [64], the combination of these three elements produces an updated knowledge about the model parameters after the data have been observed, which is also known as the posterior distribution. The change from the prior to the posterior distribution reflects what has been learned from the data (see **Figure 1**). Thus, within a Bayesian framework, a researcher can invest more effort in the specification of prior distributions by translating existing knowledge about the phenomenon under study into prior distributions [65]. As suggested by Lee and Wagenmakers "such knowledge may be obtained by eliciting prior beliefs from experts, or by consulting the literature for earlier work on similar problems" [65, p. 110].

As shown in **Figure 1**, the strength of each source of information is indicated by the narrowness of its curve. A narrower curve is more informative about the value of parameters, whereas a wider one is less informative.

Bayes' rule specifies how the prior information $p(\theta)$ and the likelihood $p(D|\theta)$ are combined to arrive at the posterior distribution denoted by $p(\theta|D)$, in Eq. (1):

$$p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{p(D)} \tag{1}$$

Eq. (1) is usually paraphrased as:

$$p(\theta|D) \propto p(D|\theta)p(\theta) \tag{2}$$

which means, "the posterior is proportional (i.e.,$\propto$) to the likelihood times the prior." In other words, the observed data (i.e., likelihood) increases our previous degree of knowledge (i.e.,

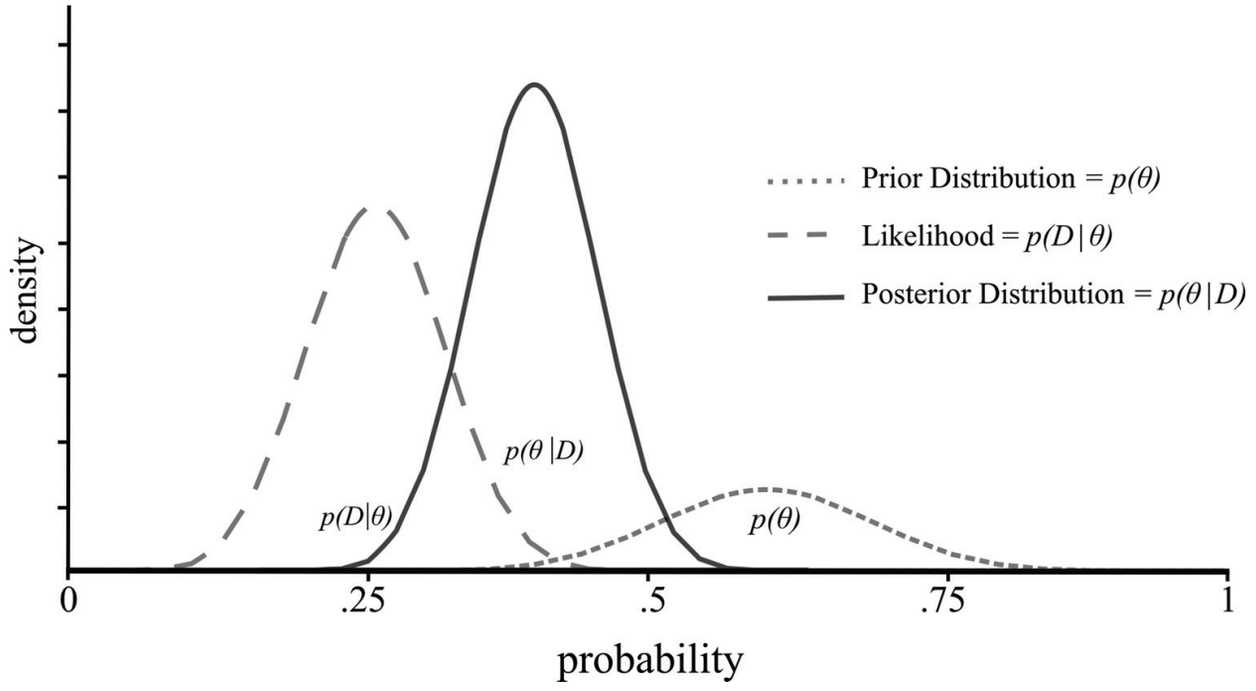

Figure 1. Prior, likelihood and posterior probability distributions.

prior) in a proportional way to its informative strength, producing a new state of knowledge about the parameters of the model (i.e., posterior). One of the benefits of the Bayesian approach is that the prior (i.e., $p(\theta)$; our present knowledge about the model parameters moderates the influence provided by the data (i.e., $p(D|\theta)$). This compromise leads to less pessimism when data are unexpectedly bad and less optimism when it is unexpectedly good [66]. Both influences are beneficial and help us to make more realistic inferences and take better decisions. For more detailed information on Bayesian inference, see, for instance, O'Hagan and Forster [54], Kruschke [59], and Jackman [67].

3.2. Bayes factor

Bayesian approaches for hypothesis testing are comparative in nature. Different models often represent competing theories or hypotheses, and the focus of interest is on which one is more plausible and better supported by the data [65]. Therefore, the Bayesian approach allows to quantify the plausibility of a given model or hypothesis (i.e., H_0) against that of an alternative model (i.e., H_a). For any comparison of two competing models or hypotheses (e.g., H_a vs. H_0), we can rely on an estimate of evidence known as the Bayes factor [52]. One of the attractive features of the Bayes factor is that it follows the principle of parsimony: When two models fit the data equally well, the Bayes factor prefers the simple model over the more complex one [68]. Nonetheless, in contrast to the NHST approach, "Bayesian statistics assigns no special

status to the null hypothesis, which means that *Bayes factors* can be used to quantify evidence for the null hypothesis just as for any other hypothesis" [65, p. 108].

Before observing the data, the *prior odds* of H_a over, e.g., H_0, are $p(H_a)/p(H_0)$, and after having observed the data we have the *posterior odds* $p(H_a|D)/p(H_0|D)$. Therefore, the ratio of the posterior odds and the prior odds is defined as the Bayes factor:

$$BF_{H_aH_0} = \frac{(D|H_a)}{(D|H_0)} = \frac{\frac{\{p(H_a|D)\}}{\{p(H_0|D)\}}}{\frac{\{p(H_a)\}}{\{p(H_0)\}}} = \frac{posterior\,odds}{prior\,odds} \quad (3)$$

Eq. (3) shows the Bayes factor for given data D and two competing hypotheses (i.e., H_0 vs. H_a), which is a measure of the evidence for H_a against H_0 provided by the data. In other words, the Bayes factor is the probability of the data under one hypothesis relative to the other. For instance, a $BF_{H_aH_0}$ = 3 indicates that H_a is three times more plausible relative to H_0 than it was a priori. From this view, the Bayes factor may be considered as analogous to the Frequentist likelihood ratio. Nevertheless, in the Bayesian context there is no reference at all to theoretical probability distributions as it is customary in a Frequentist approach. In a Bayesian framework, all inferences are made conditional on the observed data, and therefore, the Bayes factor has to be interpreted as a summary measure of the information provided by the data about the relative plausibility of two models or hypotheses (e.g., H_a vs. H_0). Jeffreys [52] suggests the following scale for interpreting the Bayes factor (**Table 1**), although some people argue against the use of thresholds, least we fall in a different version of the old $p < 0.05$ ritual (see, for instance, [69]).

Bayes factor			**Interpretation**
	>	100	Extreme evidence for H_a
30	–	100	Very strong evidence for H_a
10	–	30	Strong evidence for H_a
3	–	10	Moderate evidence for H_a
1	–	3	Anecdotal evidence for H_a
1			No evidence
1/3	–	1	Anecdotal evidence for H_0
1/10	–	1/3	Moderate evidence for H_0
1/30	–	1/10	Strong evidence for H_0
1/100	–	1/30	Very strong evidence for H_0
	<	1/100	Extreme evidence for H_0

Adapted from Jeffreys [52, p. 433], and Lee and Wagenmakers [65, p. 105].

Table 1. Evidence categories for the Bayes factor.[1]

4. Bayesian vs. Frequentist approaches to hypothesis testing: An example

Bayes factors to evaluate the amount of evidence in favor or against H_0 and H_a are one of the big selling points of the Bayesian framework.[1] As stated in the previous section, the core idea is that the magnitude of evidence in favor of the null hypothesis compared to that of the alternative hypothesis can be estimated (or vice-versa). As we have seen, this approach has multiple advantages, such as departing from a *hit-or-miss* approach to results reporting, or being able to show evidence in favor of the null. The possibility of providing evidence in favor of both the null and the alternative hypotheses has some important advantages. One of them is that it helps to overcome one of the most common issues behind the well-known file-drawer effect, in that results do not suddenly become meaningless when the *p-value* is over certain threshold. Another advantage is that it gives us more freedom when establishing hypothesis, particularly in topics where hypothesizing the absence of differences may be necessary for theoretical advance.

In this section, an example from a field known as Bayesian reasoning will be presented, which deals with how people update their beliefs when new evidence is available (e.g., when receiving a positive result in a medical test, how likely it is that I have a disease?). There is a long standing debate in the field about why people are unable to solve medical screening problems such as the one shown in **Table 2** when the information is shown in a standard probability format (i.e., single-event probabilities; for instance, 1% have cancer), but have a comparatively better time when the same information is shown in a standard frequency format (i.e., natural frequencies; for instance, 10 in 1000 have cancer). As it is often the case, the debate about these issues is very complex (for a review, see [71]), and the present example will focus on a single unnuanced aspect with the goal of showing the usefulness of the Bayesian statistics paradigm.

Standard probability format
The probability of breast cancer is 1% for women at age 40 who participate in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? _____%
Standard frequency format
Ten out of every 1000 women at age 40 who participate in routine screening have breast cancer. Eight of every 10 women with breast cancer will get a positive mammography. Ninety-five out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at age 40 who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? ____out of____

Table 2. Standard probability and standard frequency format problems, as shown by Gigerenzer and Hoffrage [72].

[1]However, we recommend the interested reader to revise a recent paper by Lakens [70], which describes an approach to test for equivalence within a Frequentist framework.

Some authors [73, 74] argue that the crucial factor explaining the differences between the two versions is not the representation format (i.e., probabilities or natural frequencies), but the reference class or more specifically the computational complexity is caused by the reference class of the problems [75]. In brief, as the probability version has a relative reference class, and all the numbers refer to the group above them (e.g., 80% from the 1% who have breast cancer will get a positive mammography). To solve the problem, we need to use the base-rates (in this example, percentage of women with and without breast cancer; 1 and 99%), and the percentage of women who got a positive mammography amongst those two groups (e.g., 80 and 9.6%; see Eq. (4)). In the frequency version, as the reference class is absolute, and all numbers can be seen as referring to the 1000 women, we can ignore the base-rates and directly use the positive mammographies for women with and without cancer (8 and 95; see Eq. (5)). The above-mentioned authors hypothesized that when reference class and computational complexity are taken into account, there is no difference between probabilities and natural frequencies. In other words, they expect the null hypothesis to be true (**Figure 2**).

$$p(H|D) = \frac{1\% \times 80\%}{1\% \times 80\% + 99\% \times 9.6\%} = 0.077 \tag{4}$$

$$p(H|D) = \frac{8}{8+95} = 0.077 \tag{5}$$

Now, imagine two PhD students, a Frequentist (i.e., Student 1) and a Bayesian (i.e., Student 2). After reading a critical but often ignored Fiedler's paper [73], they had the idea that computational complexity class (and not representation format) is the key issue when trying to understand how people solve Bayesian reasoning problems. They devise a very simple experiment where two different groups of people will be asked to solve one Bayesian reasoning problem that will be shown either in single-event probabilities or in natural frequencies. In both cases, the arithmetic complexity (i.e., number of arithmetic steps required to solve the problem) will be exactly 2. That is, to solve the problems, participants would need to do two arithmetic operations, a sum and a division. They used a test with a 100% sensitivity and 0% specificity, which could not have any clinical application, but it is useful to get a few arithmetic steps out of the probability format and check if computational complexity underlies Bayesian reasoning. With this manipulation, the algorithms to solve the probability and frequency versions become

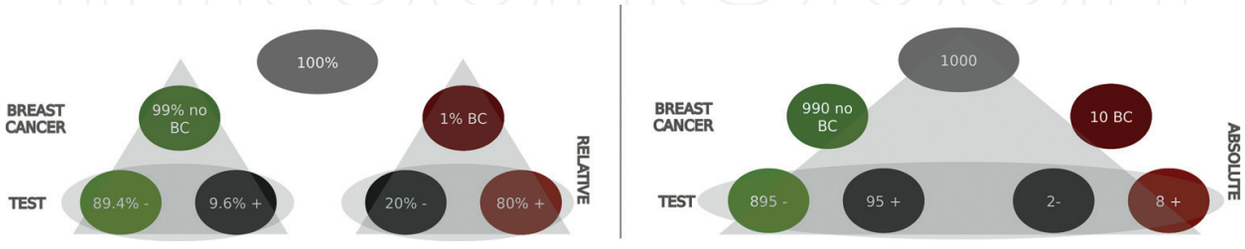

Figure 2. Relative and absolute reference classes represented by the reference of the last row (test results). In the Relative reference class, the information about the test, for example, 80% positive (+) and 20% negative results (−) refers to the 1% women with BC, but not to the 100% of the women (it is not an 80% of the 100%!). However, in the absolute reference class, the same information, 8+ and 2−, refers to the women with BC, but also to the 1000 women directly. This translates in the need to use Eq. (4) for relative probabilities and Eq. (5) for absolute frequencies.

Eqs. (6) and (7), respectively. It is easy to see how both have become roughly equivalent now in terms of arithmetic complexity.

$$p(H|D) = \frac{10\% \times 100\%}{10\% \times 100\% + 90\% \times 100\%} = \frac{10\%}{10\% + 90\%} = 0.1 \tag{6}$$

$$p(H|D) = \frac{10}{10 + 90} = 0.1 \tag{7}$$

As it can be deduced, Student 1 would have a Fisherian approach to statistics and Student 2 a Bayesian approach. Both run an experiment with a total of 62 participants (31 per group),[2] and have the following results:

Contingency tables

	Representation format		
Accuracy	**Natural frequencies**	**Probabilities**	**Total**
0	23	24	47
1	8	7	15
Total	31	31	62

4.1. PhD Student 1—Frequentist

Student 1, as the most good NHST practitioners would do, conducts a Chi-square test and reports that he did not obtain a significant effect of representation format when arithmetic steps were equal (χ^2 = 0.088, p = 0.767). He is happy, because this is congruent with his hypothesis. He then writes a brief report detailing his idea and experimental results and sends the manuscript draft to his advisor. A few days later, he receives his advisor feedback, telling him that his non-significant results could be caused by a number of reasons, and as a consequence, the non-significant results are hard to interpret.

Chi-square tests

	Value	***df***	***p***
χ^2	0.088	1	0.767
N	62		

[2]Of course, the sample size and manipulation for this experiment is more congruent with a pilot experiment than a real one that could be sent to a journal on its own. As a side note, take into account that one of the advantages of the Bayesian framework some authors propose is a sequential sampling rule, where sampling stops when the evidence (BF) is over a predetermined threshold (e.g., BF_{10} >10 | <0.1), see Lindley [76].

His advisor suggests carrying out a few more experiments using variations of the task and decent sample-sizes, to be able to perform a meta-analysis that could convince the editorial board of a journal that their endeavor is noteworthy, as they would probably have a hard time publishing those non-significant results by themselves.

4.2. PhD Student 2—Bayesian

Student 2, instead of performing a Chi-square test, prefers to use a well-known analysis among Bayesian statisticians called Bayes factor (BF; see [17, 65]). He uses a very simple to use software called JASP [61], that incorporates Bayesian contingency tables, and outputs BF results in ready to use APA formatted tables. He finds that when arithmetic steps are equal, there is a BF_{01} of 4.656, that is, there is 4.6 times more evidence in favor of the null-hypothesis than the alternative-hypothesis. Along his advisor, they send the manuscript to a journal, pushing for the relative importance of arithmetic complexity over representation format. In practical terms, it is more likely that the editor will be willing to publish this interesting result, although the amount of evidence in favor of the null would be considered moderate by some standards (see [53]).

Bayesian contingency tables tests

	Value
BF_{0+} independent multinomial	4.656
N	62

Note: For all tests, the alternative hypothesis specifies that group *Natural-Frequencies* is greater than group *Probabilities.*

As the evidence for the null effect is not very strong, they would need to run a few more studies with variations to replicate the finding and show, using BF, how much more evidence there is for the null hypothesis compared to the alternative hypothesis. Alternatively, they could increase the sample size in their experiment until the stopping rule threshold (e.g., $BF_{10} < 0.1$) is reached.

This example was aimed to describe (in a very simplified manner) one of the practical advantages of the Bayesian framework, that is, being able to present the amount of evidence for and against both the null and alternative-hypotheses. This, combined with the incremental nature of the Bayesian inference process, allows us to move further from the *hit-or-miss* approach generally reinforced by the NHST framework, in which significant results are seen as more valuable than non-significant ones.

5. Conclusion

During the past 70 years, the NHST has dominated the way in which knowledge is produced and interpreted and still governs the way in which researchers analyze their data, reach

conclusions, and report results [10, 45]. This approach has been largely criticized [9, 16, 21, 22, 39, 43–46], and "a major concern expressed by critics is that such testing is misunderstood by many of those who use it" [35, p. 241]. Some authors [9, 13] emphasized that one of the most pervasive influences of the NHST approach has been its over reliance on *p-values*, and in particular, in the way that *p-values* have been interpreted (see, for instance [35, 36, 77]). One of the most common misinterpretations of *p-values* it has been to consider a *p-value* as a valid indicator of the magnitude of evidence of a result (i.e., effect size fallacy). Regarding this point, Cohen emphasized that the only way to estimate the magnitude of an effect is to calculate the value of the effect size with the appropriate statistic and its confidence interval [38]. The correct way to interpret *p-values* is two-fold. On one hand, to reject H_0 only allows us to conclude that H_0 is unlikely. On the other hand, failing to reject H_0 simply allows us to state that—given the evidence at hand—one cannot make an assertion about the existence of some effect or phenomenon [42]. An immediate consequence of the wrong way in which a big number of researchers interpret *p-values* is that null results have been usually considered as the absence of evidence of the existence of an effect. This perspective regarding the decisions made when a given *p-value* threshold is not reached (i.e., $p < 0.05$) do not promote scientific advance and quickly leads us to a systematic bias toward ignoring promising but "non-significant" findings that may be further explored, fed into meta-analysis, of just be considered by other researchers in the field. This fact is against the pursue of any empirical science and may be harmful to the construction of a cumulative base of knowledge [5].

As a way to provide a complementary (or alternative) method to deal with the current NHST practice, we described here a Bayesian approach to hypothesis testing. A Bayesian approach allows us to think about phenomena in terms of the magnitude of evidence that supports the existence of an effect, instead of a dichotomous and artificial way of thinking in which an effect either exists or does not exist [21]. As described in previous sections, a Bayesian approach provides us a measure of evidence for and against both the null and the alternative hypotheses (i.e., Bayes factor, BF; see [17]). The use of Bayes factors helps to overcome one of the most common issues behind the well-known file-drawer effect, reducing the existent bias through which results suddenly become meaningless when the *p-value* is over certain threshold (e.g., $p > 0.05$). A straightforward feature of this approach is that "Bayesian statistics assigns no special status to the null hypothesis, which means that *Bayes factors* can be used to quantify evidence for the null hypothesis just as for any other hypothesis" [65, p. 108]. Therefore, a Bayesian approach gives us more freedom when establishing hypothesis, for example in topics where hypothesizing the absence of differences may be necessary for theoretical advance.

However, a major problem with Bayesian statistics has historically been that they require complex and intricate mathematical calculations that were analytically intractable, at least without the required techniques and specialized software. However, this scenario changed dramatically during the 1990s with the development of sampling techniques like Markov-Chain Monte Carlo (MCMC; see [55]) along with the availability and improvement of specifically developed software (e.g., WinBUGS, see [57, 58]; JAGS, see [59, 60]) that makes exact Bayesian inferences possible even in very complex models. Nowadays, the relatively recent implementation and availability of Bayesian analysis in "easy-to-use" and open software such as JASP [61], R toolboxes such as Bayes factor [78], or more specialized ones like WinBUGS,

JAGS, or Stan (http://mc-stan.org/) makes Bayesian statistics more accessible to all researchers, academics and students. This widespread availability, paired with the advantages of the Bayesian approach described in this chapter, and several times elsewhere [79–82], should help establish the Bayesian paradigm as a viable and popular alternative to NHST.

Despite all the important Bayesian paradigm advantages, as always, there is potential for misuse. As pointed out by Morey, Bayes factor interpretation is very natural (i.e., as the amount of evidence in favor of one hypothesis in comparison to another), and does not need specific decision thresholds, as it is the case of *p-values* [83]. However, some standards that could help to communicate BF results have been proposed (see [53]) and may be helpful to people that are not familiar with them. Nonetheless, the introduction of these labels also creates an opportunity for misuse, as they could be misinterpreted as decision boundaries. It is very important to be aware of this fact, and be careful when using them, to avoid making "BF > 3" the new "$p < 0.05$."

To sum up, the main goal of this chapter has been to increase the degree of awareness regarding the limitations of the NHST approach and highlight the advantages of the Bayesian approach. We expect that the inclusion of an easy-to-understand example of a specific case where a Bayesian paradigm shows its practical utility may offer the newborn readers on this matter a glimpse to the usefulness of this alternative to the way in which they can analyze and interpret their data. As a final remark, we would like to point an often-heard recommendation for people interested in starting to use BF, which is to introduce them alongside *p-values* and effect size measures, to ease the transition to the new paradigm, and make them comprehensible to people not yet familiarized with them.

Author details

Alonso Ortega[1]* and Gorka Navarrete[2]

*Address all correspondence to: alonso.ortega@uai.cl

1 School of Psychology, Universidad Adolfo Ibáñez, Chile

2 Center for Social and Cognitive Neuroscience (CSCN), School of Psychology, Universidad Adolfo Ibáñez, Chile

References

[1] Lakatos I. Falsification and the methodology of scientific research programmes. In: Harding S, editor. Can Theories be Refuted? Dordrecht: Holland: D. Reidel Publishing Company; 1976. pp. 205-259

[2] Radder H. Toward a more developed philosophy of scientific experimentation. In: Radder H, editor. The Philosophy of Scientific Experimentation. Pittsburgh: University of Pittsburgh Press; 2003. pp. 1-18

[3] Harper RS. The first psychological laboratory. Isis. 1950;**41**(2):158-161

[4] Popper KR. Degree of confirmation. The British Journal for the Philosophy of Science. 1954;**5**(18):143-149

[5] Curran PJ. The seemingly quixotic pursuit of a cumulative psychological science: Introduction to the special issue. Psychological Methods. 2009;**14**(2):77-80

[6] Cumming G. The new statistics why and how. Psychological Science. 2013;**25**(1):7-29

[7] Loftus GR. Psychology will be a much better science when we change the way we analyze data. Current Directions in Psychological Science. 1996;**5**(6):161-171

[8] Rossi JS. A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In: Harlow L, Mulaik S, Steiger J, editors. What If There Were No Significance Tests. Mahwah, NJ: Erlbaum Associates Publishers; 1997. pp. 175-197

[9] Wagenmakers E-J. A practical solution to the pervasive problems of *p* values. Psychonomic Bulletin & Review. 2007;**14**(5):779-804

[10] Gigerenzer G. We need statistical thinking, not statistical rituals. Behavioral and Brain Sciences. 1998;**21**(2):199-200

[11] Ioannidis JP. Why most published research findings are false. PLOS Medicine. 2005;**2**(8): e124

[12] Wagenmakers EJ, Wetzels R, Borsboom D, Van Der Maas HL. Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). Journal of Personality and Social Psychology. 2011;**100**(3):426-432

[13] Llobell JP, Dolores M, Navarro F, et al. Usos y abusos de la significación estadística: propuestas de futuro ("Necesidad de nuevas normativas editoriales"). Metodologia de las Ciencias del Comportamiento, 2004; Volumen Especial: 465-469

[14] Kirk RE. The importance of effect magnitude. In: Davis SF, editor. Handbook of Research Methods in Experimental Psychology. Malden, MA: Blackwell Publishing; 2003. pp. 83-105

[15] Cohen J. A power primer. Psychological Bulletin. 1992;**112**(1):155-159

[16] Schmidt FL. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. American Psychological Association. 1996;**1**(2): 115-129

[17] Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association. 1995;**90**(430):773-795

[18] Dienes Z. Bayesian versus Orthodox statistics: Which side are you on? Perspectives on Psychological Science. 2011;**6**(3):274-290

[19] Wilkinson L, Task Force on Statistical Inference APA Board of Scientific Affairs. Statistical methods in psychology journals: Guidelines and explanations. American Psychologist. 1999;**54**:594-604

[20] Levine TR, Weber R, Hullett C, Park HS, Lindsey LLM. A critical assessment of null hypothesis significance testing in quantitative communication research. Human Communication Research. 2008;**34**(2):71-187

[21] Dixon P. The p-value fallacy and how to avoid it. Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Experimentale. 2003;**57**(3):189-202

[22] Gill J. The insignificance of null hypothesis significance testing. Political Research Quarterly. 1999;**52**(3):647-674

[23] Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review. 2009;**16**(2):225-237

[24] Christensen R. Testing Fisher, Neyman, Pearson, and Bayes. The American Statistician. 2005;**59**(2):121-126

[25] Goodman SN. P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. American Journal of Epidemiology. 1993;**137**(5):485-496

[26] Fisher RA. Two new properties of mathematical likelihood. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character. 1934;**144**(852):285-307

[27] Fisher RA. Statistical Methods for Research Workers. Edinburgh: Genesis Publishing Pvt Ltd; 1925

[28] Fisher RA. Statistical methods and scientific induction. Journal of the Royal Statistical Society. Series B (Methodological). 1955;**17**:69-78

[29] May RB, Masson MJ, Hunter MA. Application of Statistics in Behavioral Research. NY: Harper & Row; 1990

[30] Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. 1933;**A231**:289-337

[31] Singh VB. Don't Confuse Type I and Type II errors. 2015. Available from: https://www.linkedin.com/pulse/dont-confuse-type-i-ii-errors-bhaskar-vijay-singh-frm?articleId=6077308381431951360 [Accessed: June 21, 2017]

[32] Nix TW, Barnette JJ. The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. Research in the Schools. 1998;**5**(2):3-14

[33] Gigerenzer G. The superego, the ego, and the id in statistical reasoning. In: A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues. Hillsdale, NJ: L. Erlbaum Associates; 1993. pp. 311-339

[34] Sedlmeier P, Gigerenzer G. Teaching Bayesian reasoning in less than two hours. Journal of Experimental Psychology General. 2001;**130**(3):380-400

[35] Nickerson RS. Null hypothesis significance testing: A review of an old and continuing controversy. Psychological Methods. 2000;**5**(2):241-301

[36] Badenes-Ribera L, Frias-Navarro D, Iotti B, Bonilla-Campos A, Longobardi C. Misconceptions of the p-value among Chilean and Italian Academic Psychologists. Frontiers in Psychology. 2016;**7**:1247

[37] Kline RB. Beyond Significance Testing, Statistics Reform in the Behavioral Sciences. 2nd ed. Washington, DC: American Psychological Association; 2013

[38] Cohen J. The earth is round ($p < .05$). American Psychologist. 1994;**49**:997-1003

[39] Carver R. The case against statistical significance testing. Harvard Educational Review. 1978;**48**(3):378-399

[40] Rozeboom WW. The fallacy of the null-hypothesis significance test. Psychological Bulletin. 1960;**57**(5):416

[41] Wetzels R, Raaijmakers JG, Jakab E, Wagenmakers E-J. How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. Psychonomic Bulletin & Review. 2009;**16**(4):752-760

[42] Cohen J. The statistical power of abnormal-social psychological research: A review. The Journal of Abnormal and Social Psychology. 1962;**65**(3):145

[43] Shaver JP. What statistical significance testing is, and what it is not. The Journal of Experimental Education. 1993;**61**(4):293-316

[44] Carver RP. The case against statistical significance testing, revisited. The Journal of Experimental Education. 1993;**61**(4):287-292

[45] Krueger J. Null hypothesis significance testing: On the survival of a flawed method. American Psychologist. 2001;**56**(1):16

[46] Meehl PE. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology. 1978;**46**(4):806-834

[47] Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers E-J. Statistical evidence in experimental psychology an empirical comparison using 855 t tests. Perspectives on Psychological Science. 2011;**6**(3):291-298

[48] Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas H. Yes, Psychologists Must Change the Way They Analyse Their Data: Clarifications for Bem, Utts, and Johnson (2011). 2011. Available from: http://web.stanford.edu/class/psych201s/psych201s/papers/ClarificationsForBemUttsJohnson.pdf [Accessed: July 26, 2017]

[49] Bem DJ. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. Journal of Personality and Social Psychology. 2011;**100**(3):407-425

[50] Bem DJ, Utts J, Johnson WO. Must psychologists change the way they analyze their data? Journal of Personality and Social Psychology. 2011;**101**(4):716-719

[51] Bernardo JM. A Bayesian analysis of classical hypothesis testing. Trabajos de estadística y de investigación operativa. 1980;**31**(1):605-647

[52] Lindley DV. The philosophy of statistics. Journal of the Royal Statistical Society: Series D (The Statistician). 2000;**49**:293-337

[53] Jeffreys H. Theory of Probability. Oxford: Clarendon Press; 1961

[54] O'Hagan A, Forster JJ. Kendall's Advanced Theory of Statistics. Vol. 2B. Bayesian Inference. London: Arnold; 2004

[55] Gamerman D, Lopes HF. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Boca Raton: CRC Press; 2006

[56] Gilks WR, Richardson S, Spiegelhalter DJ. Introducing Markov Chain Monte Carlo, Markov Chain Monte Carlo in Practice. London: Chapman & Hall; 1996

[57] Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. Statistics in Medicine. 2009;**28**(25):3049-3067

[58] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS-a Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing. 2000;**10**(4):325-337

[59] Kruschke JK. Introduction: Credibility, Models, and Parameters, Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. Boston: Academic Press; 2015. pp. 15-30

[60] Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Vienna: TU Wien; 2003. p. 125

[61] Love J, Selker R, Marsman M, Jamil T, Dropmann D, Verhagen A, Wagenmakers E. JASP (Version 0.7) [Computer Software]. Amsterdam, the Netherlands: JASP Project; 2015

[62] Griffiths TL, Tenenbaum JB, Kemp C. Bayesian inference. In: Holyoak K, Morrison R, editors. The Oxford Handbook of Thinking and Reasoning. New York: Oxford University Press; 2012. pp. 22-35

[63] Samaniego F. A Comparison of the Bayesian and Frequentist Approaches to Estimation. New York: Springer; 2010

[64] Bayes T. An essay toward solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London. 1763;**53**:370-418

[65] Lee MD, Wagenmakers E-J. Bayesian Cognitive Modeling: A Practical Course. Cambridge, New York: Cambridge University Press; 2014

[66] Berger JO, Moreno E, Pericchi LR, Bayarri MJ, Bernardo JM, Cano JA, De la Horra J, Martín J, Ríos-Insúa D, Betrò B. An overview of robust Bayesian analysis. Test. 1994;**3**(1): 5-124

[67] Jackman S. Bayesian Analysis for the Social Sciences. West Sussex: Wiley Chichester; 2009

[68] Myung IJ, Pitt MA. Applying Occam's razor in modeling cognition: A Bayesian approach. Psychonomic Bulletin & Review. 1997;**4**(1):79-95

[69] Bigler ED. Symptom validity testing, effort, and neuropsychological assessment. Journal of the International Neuropsychological Society. 2012;**18**(04):632-640

[70] Lakens D. Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. Social Psychological and Personality Science. 2017;March 4:1-21

[71] Barbey AK, Sloman SA. Base-rate respect: From ecological rationality to dual processes. Behavioral and Brain Sciences. 2007;**30**(03):241-254

[72] Gigerenzer G, Hoffrage U, Mellers BA, et al. How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review. 1995;**102**:684-704

[73] Fiedler K, Brinkmann B, Betsch T, Wild B. A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. Journal of Experimental Psychology: General. 2000;**129**(3):399-418

[74] Lesage E, Navarrete G, De Neys W. Evolutionary modules and Bayesian facilitation: The role of general cognitive resources. Thinking & Reasoning. 2013;**19**(1):27-53

[75] Ayal S, Beyth-Marom R. The effects of mental steps and compatibility on Bayesian reasoning. Judgment and Decision Making. 2014;**9**(3):226-242

[76] Lindley DV. Bayesian statistics: A review. Society for Industrial and Applied Mathematics; 1972

[77] Gliner JA, Leech NL, Morgan GA. Problems with null hypothesis significance testing (NHST): What do the textbooks say? The Journal of Experimental Education. 2002;**71**(1): 83-92

[78] Morey RD, Rouder JN. Bayes Factor: Computation of Bayes Factors for Common Designs. R package version 0.9.12-2. 2015. Available from: https://cran.r-project.org/package=BayesFactor [Accessed: June 21, 2017]

[79] Berry DA. Bayesian clinical trials. Nature Reviews Drug Discovery. 2006;**5**(1):27-36

[80] Briggs AH. A Bayesian approach to stochastic cost-effectiveness analysis. Health Economics. 1999;**8**(3):257-261

[81] Ortega A, Wagenmakers E-J, Lee MD, Markowitsch HJ, Piefke M. A Bayesian latent group analysis for detecting poor effort in the assessment of malingering. Archives of Clinical Neuropsychology. 2012;**27**(4):453-465

[82] Stegmueller D. How many countries for multilevel modeling? A comparison of Frequentist and Bayesian approaches. American Journal of Political Science. 2013;**57**(3):748-761

[83] Morey RD. On verbal categories for the interpretation of Bayes factors. 2015. Available from: http://bayesfactor.blogspot.cl/2015/01/on-verbal-categories-for-interpretation.html [Accessed: June 21, 2017]