Abstract

Bayes factors quantify the evidence in support of the null (absence of an effect) or the alternative hypothesis (presence of an effect). Based on commonly used cut-offs, Bayes factors between 1/3 and 3 are interpreted as evidentially weak and one typically concludes there is an absence of evidence. In this commentary on Warmelink, Subramanian, Tkacheva, and McLatchie (2019) we discuss how a Bayesian report can be made more informative. Firstly, this implies a departure from the labels provided by commonly used cut-offs when reporting Bayes factors. Instead, we encourage researchers to report the value of the Bayes factors, or to convert these values into nominal support for the hypotheses. Secondly, researchers can provide recommendations to design follow-up studies by examining the posterior distribution of the magnitude of the effect size. Lastly, we show how individual Bayes factors can be evaluated in the context of large-scale meta-analyses.

Keywords: Bayes factors, deception detection, sample size

What to do with all these Bayes factors:
How to make Bayesian reports in deception research more informative

A crucial distinction when interpreting the results of a study is the difference between the absence of an effect from the absence of evidence. This distinction is typically ignored within frequentist statistics, where "non-significant" *p* values (*p* > .05) are incorrectly interpreted as an absence of an effect (Altman & Bland, 1995). For this reason alone, Warmelink, Subramanian, Tkacheva, and McLatchie (2019) should be commended for reporting Bayes factors, which can quantify the evidence for the alternative (presence of an effect) over the null hypothesis (absence of an effect), and vice versa. For instance, a Bayes factor $BF_{10}=8$ implies that the observations are 8 times more likely under the hypothesis that the effect is present than under the hypothesis that the effect is absent. Similarly, $BF_{10}=0.125$ (equivalently, $BF_{01}=1/BF_{10}=8$) implies that the data are 8 times more likely under the null compared to the alternative.

To aid interpretation, cut-offs (e.g., Jeffreys, 1961) for Bayes factors have been proposed. These cut-offs suggest that Bayes factors between 1 and 3 (1/3 and 1) provide only weak evidence for the alternative hypothesis relative to the null hypothesis (or for the null hypothesis relative to the alternative; e.g., Jarosz & Wiley, 2014). Hence, when confronted with evidentially weak Bayes factors, the common conclusion is that there is an absence of evidence (e.g., Wagenmakers, Morey, & Lee, 2016), and the results are perceived as inconclusive.

The reporting of Bayes factors is still quite novel within deception research, though it appears to be gaining popularity (e.g., Kleinberg, Warmelink, Arntz, & Verschuere, 2018; Leal, Vrij, Deeb, & Kamermans, 2019). We believe this is a good thing, but it does bring with it new challenges for reviewers and editors. Using Warmelink, Subramanian, Tkacheva, and McLatchie (2019) as a concrete example, we elaborate on the continuous nature of Bayes factors and how to interpret them. We begin by warning against an over-reliance on cut-offs. Next, we give some suggestions on how to make a Bayesian report more informative.[1]

Risk of Cut-Offs and Reporting Alternatives

As noted above, cut-offs have been suggested to aid the interpretation of Bayes factors. These cut-offs, however, should be used with caution. Let us take the results of Warmelink et al. (2019) as a concrete example. In total, they report 22 Bayes factors. Nineteen of these are between 1/3 and 3, and, therefore, evidentially weak according to the cut-offs they themselves adopt. The remaining three analyses resulted in Bayes factors of BF_{10}=3.58, BF_{10}=3.64, and in opposite direction BF_{01}=4.76, which were interpreted to provide “moderate” evidence. This interpretation highlights the main risk of cut-offs: They encourage categorical rather than continuous thinking. Yes, a Bayes factor of 3.58 provides more evidence than a Bayes factor of 2.8, but to say that one provides moderate evidence, whereas the other provides weak or inconclusive evidence, seems somewhat misguided (Wagenmakers et al., 2018).

The boundaries between the categories further fade away, when we convert a Bayes factor into posterior model probabilities. A Bayes factor only measures the strength of the *relative evidence* provided by the data, whereas the posterior model probabilities quantify the *nominal support* for the hypotheses after data observation. To calculate the posterior model probabilities, we combine the prior plausibility of the hypotheses with the Bayes factor using the following formula

$$P(H_1 \,|\, \text{data}) = \frac{\text{BF}_{10} \times P(H_1)}{\text{BF}_{10} \times P(H_1) + P(H_0)} \text{ and } P(H_0 \,|\, \text{data}) = 1 - P(H_1 \,|\, \text{data}) \tag{1}$$

For instance, if the plausibility of the presence and the absence of an effect are equal before data observation, then we set $P(H_1)=0.5$ and $P(H_0)=0.5$. Observations leading to a Bayes factor of $BF_{10}=3.58$ then yield a nominal support of 78%, $P(H_1 \mid data)=3.58 \times 0.5 /(3.58 \times 0.5 + 0.5)$, for the hypothesis that there is an effect, leaving a posterior probability of 22%, $P(H_0 \mid data) = 1-0.78$, for the hypothesis that the effect is absent[2]. These probabilities can in turn be visualized using a simple pie chart known as a pizza plot (Wagenmakers et al., 2018; see Figure 1).

(insert Figure 1 about here)

Now look at Figure 1. The dark portion of the pie chart corresponds to the 78% in support of H_1. Imagine that this chart was on the table in front of you. You close your eyes, spin the chart, and at random place your finger down. Now open your eyes. How surprised would you be to find your finger on the white portion of the pie chart? Your degree of surprise is an intuitive measure of the strength of the nominal support for H_1 (Wagenmakers et al., 2018). Our guess is you would not be that surprised—there was after all a 22% probability of your finger landing on the white portion. Similar computations show that the "evidentially weak" Bayes factor of 2.8 would lead to a posterior probability of 74% in support of H_1 and 26% for H_0. In terms of nominal support for the hypothesis, the distinction between "evidentially weak" and "moderate" Bayes factors in these examples is negligible. Hence, the categorizations are not that informative, and we strongly encourage researchers to discuss their results using the value of the Bayes factors or the posterior probabilities for each hypothesis instead.

Again, we can use Warmelink et al. (2019) as a concrete example. In their discussion, they conclude that Hypothesis 2 was partly supported and that Hypothesis 3 was not supported.

We worry that such language exaggerates the strength of evidence of the data. Ultimately, researchers should be fair to the uncertainty implied by the Bayes factors. This uncertainty is lost when we focus on labels such as “weak” and “moderate”, and when the relative evidence is not converted to nominal support.

However, simply reporting that results are uncertain will likely leave many reviewers and readers wanting. We see at least two options to help researchers provide more informative Bayesian reports: first, by focusing on the posterior distribution of the effect size, and second by relating Bayes factors to the broader research field by discussing prior model probabilities. We will discuss each approach in turn.

Beyond Bayes Factors: Examining the Posterior Distribution of the Effect Sizes

A Bayesian report can be made more informative by studying the posterior distribution of the effect size parameter.[3] This can provide us with an estimate of the magnitude of the effect, given that it exists, or provide us with information on how to design a follow-up study. For instance, say we conduct a large sampled between-group study ($N = 1{,}000$) and find a standardized between-groups difference (Cohen’s d) = 0.1. With a default Bayesian t-test, this would produce a Bayes factor around 1.5 in support of the null hypothesis, that there is no difference between the two groups. Although this result helps little in determining whether the null or alternative is true, by examining the posterior distribution of the parameter, and the associated credible interval (the Bayesian equivalent to the frequentist confidence interval), one can still learn something from the data. Specifically, that *if* there is an experimental effect, it is likely to be very small indeed (for a pedagogical online tool to visual this point see https://rpsychologist.com/d3/bayes/). Such a conclusion is considerably more valuable than what one can glean by solely focusing on the Bayes factor. Furthermore, with the estimate of the effect in hand, one can then efficiently plan follow-up studies.

The Potential Role of Prior Model Probabilities

Another way to move beyond a simple reporting of Bayes factors is to use more informed prior model probabilities when calculating the nominal support for hypotheses. Above, for lack of better choice, we had equal prior model probabilities: $P(H_0)=0.5$ and $P(H_1)=0.5$. More informed prior model probabilities can be derived from large-scale field-wide meta-analyses, or better still can be agreed upon by the research community. In particular, if one has reason to be sceptical for the presence of an effect, the prior model probabilities can be weighted in favour of H_0, thereby correcting for false positives.[4]

In the context of deception detection, we have strong reason to be highly sceptical to most cues to deceit. It seems that the research on deception cues, prior to 2003, are compatible with a world in which there are no reliable cues to deceit at all (Luke, 2019). Hence, in the best case, Luke's findings suggest that only a small fraction of deception studies lead to the detection of an actual effect—a genuine cue to deception. For the sake of argument, let us say that only 8% of the published articles report an actual deception cue, leaving 92% of the published articles reporting a false positive finding. Thus, $P(H_1)=0.08$ and $P(H_0)=0.92$. This interpretation of Luke's finding can now be used as a context to evaluate the moderate $BF_{10}=3.58$ as reported in Warmelink et al. (2019) using the Eq. (1). A direct calculation shows that the nominal support for H_1 of 8% is then increased to 24%, $P(H_1 \mid \text{data}) = 3.58 \times 0.08/(0.92 + 3.58 \times 0.08)$, which leaves a posterior probability of 76% in support for H_0. Hence, there is indeed *relative evidence* for H_1, because the Bayes factor is larger than 1, but based on the posterior model probabilities we cannot conclude that there is more *nominal support* for H_1 than H_0. In other words, the relative evidence of BF_{10}=3.58 is not enough to overcome the initial scepticism brought about by Luke's findings.

By changing $P(H_0)$=0.5 and $P(H_1)$=0.5 to the more stringent $P(H_0)$=0.92 and $P(H_1)$=0.08, we now require BF_{10} to be larger than 11.5 (=0.92/0.08) before the posterior model probability for H_1 exceeds that of H_0.[5] In other words, we changed the threshold of preference—the strength of the evidence needs to be at least 11.5 before we *begin* to consider a found deception cue feasible. Moreover, we now need a Bayes factor of 40.77, which can be verified using Eq. (1), before we get nominal support of $P(H_1 \mid \text{data}) = 0.78$.

Our aim here was to demonstrate how prior model probabilities can let us draw important conclusions about the hypotheses and thereby provide more informative results than simply reporting Bayes factors. We wish to stress however that our suggested base rate of 8% for H_1 is only meant for illustrative purposes. This matter should be discussed at the level of the research community.

Concluding Remarks

We provided some suggestions on how to make a Bayesian report more informative. Firstly, we warn against an over-reliance on the standard Bayes factor cut-offs. Instead, we encourage researchers to report the value of the Bayes factors, or convert it to the nominal support for the hypotheses. Secondly, we encourage researchers to go beyond a simple reporting of Bayes factors. This can be achieved, for example, by studying the posterior distribution of the effect size or by evaluating Bayes factors in the context of a large-scale meta-analysis using Eq (1).

Ultimately, however, there is no quick fix for making studies more informative after the fact. Greater care must be taken before data collection to increase the chances of producing compelling results (for guidelines see Schönbrodt & Wagenmakers, 2018 and see Ly et al., in press, on how to quantify replicability). Luke (2019) brought the consequences of ignoring this advice to light. Luke outlines how the first 80 years of research on cues to deceit may have been little more than a fool's errand. It seems that the extant research on deception cues, prior to 2003, are compatible with a world in which there are no reliable cues to deceit at all. Individual small-sample studies, whose effects may well be nothing more than sampling variation, were over-interpreted, published, and, in worst case scenarios, used for policy recommendations. It took some 60 years for this issue to come to light in DePaulo et al.'s (2003) meta-analysis. And almost another 20 for us to understand the true consequences of this practice in Luke's re-analysis of the data. One can rightly wonder how much quicker we would have come to this conclusion if studies from the outset had been designed to provide more informative and compelling results.

Since DePaulo et al.'s meta-analysis, deception researchers' focus on traditional cues to deceit has waned. Instead, there is a new focus on interviewers actively eliciting cues to deceit, (Vrij & Granhag, 2012), of which Warmelink et al. (2019) is one example. Rather than passively observing truth tellers and liars, these methods consist of asking questions in a strategic manner in order to increase differences in the statements or behaviors between these groups. This new approach may be a genuine path to a viable method of deception detection, or it may be another dead end. If we do not start designing our studies to produce more informative results, we may need to wait another 80 years to find out.

References

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ, 311*(7003), 485. doi:10.1136/bmj.311.7003.485

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2), 121. Doi: 10.1038/nprot.2010.182

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological bulletin, 129*(1), 74-118. doi:10.1037/0033-2909.129.1.74

Grünwald, P. D., de Heide, R., & Koolen, W. M. (2019). Safe testing. *arXiv preprint arXiv:1906.07801*.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving, 7*(1), 2. doi: 10.7771/1932-6246.1167

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

Kleinberg, B., Warmelink, L., Arntz, A., & Verschuere, B. (2018). The first direct replication on using verbal credibility assessment for the detection of deceptive intentions. *Applied Cognitive Psychology*. Advance online publication. Doi: 10.1002/acp.3439

Leal, S., Vrij, A., Deeb, H., & Kamermans, K. (2019). Encouraging interviewees to say more and deception: The ghostwriter method. *Legal and Criminological Psychology*. Advance online publicaiton. Doi: 10.1111/lcrp.12152

Luke, T. J. (2019). Lessons from Pinocchio: Cues to deception may be highly exaggerated. *Perspectives on Psychological Science, 14*, 646-671.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E. J. (in press). Replication Bayes factors from evidence updating. *Behavior Research Methods*.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review, 25*(1), 128-142. Doi: 10.3758/s13423-017-1230-y.

Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition, 1*(2), 110-117. doi:http://dx.doi.org/10.1016/j.jarmac.2012.02.004

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., . . . Boutin, B. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic bulletin & review, 25*(1), 58-76. doi: 10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science, 25*(3), 169-176. Doi: 10.1177/0963721416643289

Warmelink, L., Subramanian, A., Tkacheva, D., & McLatchie, N. (2019). Unexpected questions in deception detection interviews: Does question order matter? *Legal and Criminological Psychology, Advance online publication*(0). doi:10.1111/lcrp.12151

Footnotes

[1] For a comprehensive introduction to Bayesian statistics see Wagenmakers et al. (2016) and for a primer on how to conducted Bayesian analyses with the free software package JASP see Wagenmakers et al. (2018).

[2] Probability in this context refers to the plausibility of the hypotheses. For instance, $P(H_1)=0.75$ and $P(H_0)=0.25$ implies that we believe that it is three (=0.75/0.25) times as plausible that there is an effect compared to no effect. Similarly, $P(H_1 \mid \text{data}) = 0.9$ and $P(H0 \mid \text{data}) = 0.1$ implies that we believe that it is 9 times as plausible that there is an effect compared to no effect after seeing the data.

This can be contrasted with "chance", which is a statement about the *potential* data. A popular example of such a statement is a p-value. For instance, t=1.9, p=0.06 means that there is a 6% chance to see *data* that lead to the observed t=1.9, and –more extreme, but not observed *potential data*– that lead to more extreme values of the test statistic, i.e., $t > 1.9$, and when testing two-sided also $t < -1.9$. Typically, this strict distinction between probability and "chance" is ignored outside of philosophy, and the reader has to infer the meaning of "probability" from the context.

[3] Note that the posterior distribution on the effect size is a continuous object, whereas the posterior model probability is discrete, since we only considered two models H_0 and H_1. The posterior model probabilities allow us to study whether the effect is present or absent. In contrast, the posterior distribution for the effect size allows us to study the magnitude of the effect, under the assumption that it exists.

[4] The idea to change the threshold of preference for H_1 over H_0 is not new. A similar idea was used in genome-wide association studies (e.g., Clarke et al., 2011), where corrections for

multiple comparisons led to changing the *p*-value threshold from $p < 0.05$ to $p < 5.8 \times 10^{-8}$. For more details on the relationship between the threshold of preference and type I error control, see Grünwald, de Heide and Koolen (2019).

[5] A direct application of Eq. (1) with $P(H_0)=0.92$ and $P(H_1)=0.08$ and $BF_{10}=11.5$ shows that the nominal support for H_1 is then $P(H_1 \mid data)=0.5$ and, thus, $P(H_0 \mid data)=0.5$. Similarly, when $BF_{10} > 11.5$ we then have $P(H_1 \mid data)$ larger than $P(H_0 \mid data)$.

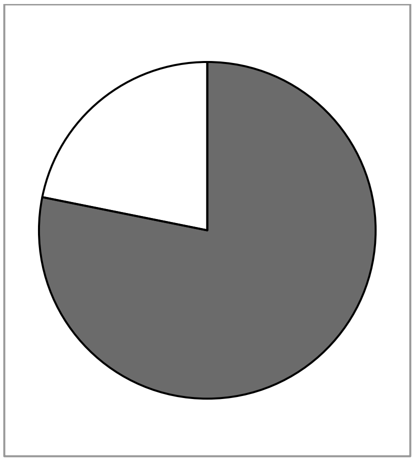

Figure 1. Pizza plot, or proportion wheel, visualizing a Bayes factor of 3.2 as the proportion of a circle. The dark portion provides the evidence in support of H1.