What to do with all these Bayes factors:

The issue of weak evidence in deception research

Erik Mac Giolla
Department of Psychology,

University of Gothenburg,

Gothenburg, Sweden

[b]Department of Behavioral Sciences,

University West,

Trollhättan, Sweden

Alexander Ly

Psychological Methods Department,
University of Amsterdam,
The Netherlands

Machine Learning Group,
Centrum Wiskunde & Informatica,
The Netherlands

Author Note

We would like to thank Timothy J. Luke and Eric-Jan Wagenmakers for their helpful comments on earlier versions of this paper.

Correspondence concerning this article should be addressed to Erik Mac Giolla, Department of Psychology, University of Gothenburg, P.O. Box 500, 405 30 Gothenburg, Sweden. E-mail: Erik.Mac.Giolla@psy.gu.se; Phone: +46 31 786 1647; Fax: +46 31 786 4628

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 what to do when we as researchers, reviewers, and editors, are faced with evidentially weak Bayes factors. Firstly, we warn against an over-reliance on cut-offs when interpreting Bayes factors. Secondly, we argue that when Bayes factors are evidentially weak researchers should nonetheless clearly outline what can or cannot be learned from the data. Lastly, we give some suggestions on how to make a Bayesian report more informative. Details of these suggestions will require a discussion at the level of the research community.

Keywords: Bayes factors, deception detection, sample size

What to do with all these Bayes factors:

The issue of weak evidence in deception research

"[A Bayes Factor of 5.33] would interest a gambler, but would be hardly worth more than a passing mention in a scientific paper" Harold Jeffreys (1961, p. 256-257)

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.. 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 and it appears to be gaining popularity. We believe this is a good thing, but it does bring with it new challenges for reviewers and editors. In this commentary, we raise the issue of what we, as reviewers and editors, should do when faced with evidentially weak Bayes factors. We begin by

warning against an over-reliance on cutoffs. Next, we give some suggestions on how to make a Bayesian report more informative.

As noted above, cut-offs have been suggested to aid the interpretation of Bayes factors. These cutoffs, 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 3.58, 3.64, and 4.76, which were interpreted to provide "moderate" evidence (Wagenmakers et al., 2018). 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.

To highlight this point, the *relative evidence* quantified by Bayes factors can be transformed into *nominal support* for the tested hypotheses. To understand what a Bayes factor of 3.58 says about the hypotheses, we suppose that—by lack of better choice—H_1 is as likely as H_0 before data observation. Hence, the prior probability of there being an effect is 50%, and we put the remaining 50% on the hypothesis that the effect is absent. This set-up leads to prior model odds of 1-to-1, which we can update to posterior model odds by a simple multiplication with the Bayes factor. In this case, the posterior model odds are 3.58-to-1. These odds can be visualized using a simple pie-chart known as a pizza plot (Wagenmakers et al., 2018). The prior belief for H_1 of 50% is updated by the data to a posterior belief of approximately 78% (3.58 x 0.50 / (3.58 x 0.50 + 0.50)). This leaves a probability of 22% in support of H_0.

(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 support. Our guess is you would not be that surprised—there was after all a 22% probability for H_0. 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.

A key issue here is to be fair to the uncertainty implied by the Bayes factors. This uncertainty is lost when we only report the labels such as "weak" and "moderate", and when the relative evidence is not converted to nominal support. 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. Various reasons are then provided for these results. Such a discussion, however, only tells half the story. The hypotheses may have lacked support, but the evidentially weak Bayes factors suggest that there is not much support for the null hypotheses either. At the very least, the discussion should reflect this uncertainty.

Better still is to relate the results to the broader research field. For example, it is likely that only a small fraction of deception studies lead to the detection of an actual effect (Luke, in press). For the sake of argument, let us say the base rate of finding an actual deception cue is 8%, which leaves 92% for H_0. Hence, before data collection, the odds of finding evidence for H_1 is 8-to-92. A "moderate" Bayes factor of 3.58, then updates the prior model odds of 8-to-92 to 28.64-to-92. The nominal support for H_1 of 8% is then increased to 24% (3.58 x 0.08/(0.92 + 3.58 x 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 > 1, but based on the posterior model probabilities we cannot

conclude that there is more *nominal support* for H_1 than H_0. Note that now a Bayes factor of 11.5 (92/8) is needed for posterior model odds of 1-to-1, which is equivalent to posterior model probabilities of 50% for both H_0 and H_1. In other words, only if the Bayes factors is larger than 11.5 will we favor H_1 over H_0. Hence, the base rate of 8-to-92 moved the threshold of *preference* of H_1 over H_0 from 1 to 11.5. The idea to change the threshold of preference for H_1 over H_0 to correct for false discoveries 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 $p < 5.8 \times 10^{-8}$. Our suggested base rate of 8% for H_1, however, was pulled out of thin air and it is not meant as a standard, as this matter should be discussed at the level of the research community.

What else can be learned from the Bayesian analysis? Another option is to study the posterior distribution of the effect size parameter, which can provide us with an estimate of the magnitude of the effect, given that it exists, or provides 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. Using a default prior, this will 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 an extremely pedagogical online tool to visual this point see https://rpsychologist.com/d3/bayes/). With the estimate of the effect in hand, one can can then plan appropriate follow up studies.

However, Bayes factors between 1/3 and 3 computed from studies with small samples, will have posterior distributions of the effect size parameter that will be wide. Fortunately, most of the default Bayes factors allow for optional continuation (Grünwald, de Heide & Koolen, 2018). This implies that during the review process, reviewers can request the authors to continue sampling without increasing the type 1 error, which is not the case if the authors had instead used p values. In situations when researchers cannot collect more data, a similar reasoning can still be applied. That is, once published, and the data are made public, the same principle allows for cumulative Bayesian updating (e.g., Ly et al., in press). For example, follow-up studies can build on the findings of Warmelink et al. (2019) without discarding their data.

This, however, is not an open invitation for small-sample studies. Due to the inherent variability in small-sample studies, some researchers question their scientific value altogether (Yarkoni & Westfall, 2017). Regardless of one's preferred statistical approach, small-sample

studies delay scientific progress—a point recently brought to light by Luke (in press). 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 conclusive results (for guidelines on how to design more informative studies, see Schönbrodt & Wagenmakers, 2018). Since DePaulo et al.'s meta-analysis, deception researchers' focus on traditional cues to deceit has waned. Instead, there is a new focus on active cues, where interviewers actively try to elicit cues to deceit (Vrij & Granhag, 2012), of which Warmelink et al. (2019) is one example. 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.

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

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.

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

Grünwald, P. D., de Heide, R. & Koolen, W. M. (2018). Safe testing *Manuscript in preparation.*

Luke, T. J. (in press). Lessons from Pinocchio: Cues to deception may be highly exaggerated. *Perspectives on Psychological Science*.

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.

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.

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.

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

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

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122.

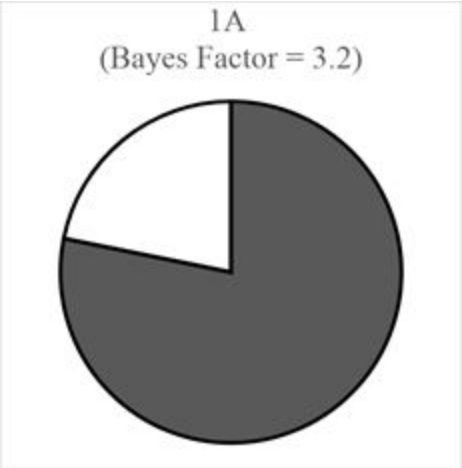

Figure 1. Pizza plots, or proportion wheels, visualizing Bayes factors as the proportion of a circle. The dark portion provides the evidence in support of H1.