Reply to Mac Giolla and Ly (2019):

On the reporting of Bayes Factors in Deception Research

Mac Giolla and Ly's (2019) commentary on Warmelink et al. (2019) discussed how deception researchers could improve their use and reporting of Bayes factors, which quantify relative evidence for competing hypotheses. Mac Giolla and Ly’s advice included:

I. Not to over-rely on cut-offs when interpreting Bayes factors.

II. To report alternatives to Bayes factors, including “nominal support”.

III. To report the posterior distribution.

We broadly agree, but in the first part of our reply we add a number of clarifications that we believe deception researchers should consider before following these suggestions. We also add two further suggestions:

1. Report *Robustness Regions* in order to demonstrate the sensitivity of conclusions to the different models of H1 used to calculate Bayes factors.

2. Design informative studies by estimating the sample size required to obtain a minimum level of evidence.

Response to Mac Giolla and Ly (2019)

In this section, we comment on the three suggestions made by Mac Giolla and Ly (2019) on the reporting of Bayesian analyses.

Do not over-rely on cut-offs

Mac Giolla and Ly begin by “warning against an over-reliance on cut-offs” (p. 2). A cut-off is the level of evidence that a researcher considers sufficient to conclude that their data supports either hypothesis, or that the data is inconclusive.

Although Mac Giolla and Ly suggest that cut-offs “encourage categorical rather than continuous thinking” (p. 2), we believe it is possible to interpret Bayes factors as a continuous measure while simultaneously acknowledging whether the Bayes factor meets a pre-specified threshold of evidence to enable a categorical interpretation; for example, to state whether an intervention was successful or not, or that the evidence was inconclusive. This judgment will influence subsequent behaviours, such as whether they encourage or discourage uptake of the intervention when disseminating their results. Thus, although the use of cut-offs should not prevent the interpretation of the Bayes factor as a continuous measure, cut-offs can nonetheless guide the researcher’s decision-making process.

Our purpose is not to discuss where the minimum cut-off should be set, which is a matter of ongoing debate in psychology (e.g. (Aczel et al., 2018; Benjamin et al., 2018; for an argument against cut-offs, see Morey, 2015). Instead, we urge researchers to clearly specify when they are using cut-offs, and what those cut-offs are. We believe that Warmelink et al. (2019) provide a reasonable and not over-reliant model of how cut-offs can be used: (I) introduce the Bayes factor as a continuous measure, (II) specify their interpretation of the Bayes factor cut-offs, (III) use labels that indicate the increasing strength of evidence, and (IV) always report the exact Bayes factor. Violating these guidelines might suggest an over-reliance on cut-offs.

Researchers should report “nominal support” (posterior odds)

The term “nominal support” in Mac Giolla and Ly’s commentary is not part of the standard Bayesian vernacular (see Box 1 for a short glossary), but appears to be equivalent to interpreting posterior odds. We would like to suggest three things that deception researchers should consider before reporting posterior odds.

1) If posterior odds are reported, they should be reported alongside Bayes factors. Van Ravenzwaaij and Wagenmakers (2019) recently argued that, as the posterior odds are the product of the Bayes factor and the prior odds, both should be reported and justified alongside posterior odds.

2) Mac Giolla and Ly suggest that the use of cut-offs is problematic for Bayes factors, and they then offer posterior odds as an alternative. Readers might assume that posterior odds therefore avoid the use of cut-offs in a manner that Bayes factors do not. However, this will again not be the case for researchers who either desire to or are required to categorically interpret their results. Again, the use of such thresholds to interpret either Bayes factors or posterior odds does not diminish their value as continuous measures. Researchers can correctly interpret a Bayes factor as a continuous measure *and* acknowledge whether it met a pre-specified minimum threshold of evidence in order to draw certain categorical conclusions. The use of such thresholds can help with identifying the required sample size and limit researcher flexibility when interpreting results.

3) Further guidance is required on how deception researchers can specify scientifically-informed prior odds. Mac Giolla and Ly specify 8% “for the sake of argument”, but not all experimental hypotheses are equally (im)plausible, and the degree of scepticism will differ between researchers. Meanwhile, reporting the Bayes factor allows readers to calculate their own posterior odds based on their prior odds.

Report the Posterior Distribution

Mac Giolla & Ly encourage deception researchers to make their studies more informative by reporting the posterior distribution. A distinction exists in statistical inference between estimation (e.g. using posterior distributions) and hypothesis testing (e.g. using

Bayes factors) (see Dienes, 2020). To calculate posterior distributions, researchers use vague "default" prior models, which allow the posterior distribution to be primarily determined by the data. However, vague prior models are not very useful for testing theories, as it is unlikely that the vague prior model represents a scientific theory of interest. In contrast, Warmelink et al. (2019) used prior models that represented their scientific theory (i.e., the prior model is a 'model of H1', see Dienes, 2019, box 4) to calculate *informed Bayes factors*.

Researchers should exercise caution before automatically calculating posterior distributions using the model of H1 used to calculate informed Bayes factors. Consider the Bayes factor reported on p. 7 of Warmelink et al (2019): $B_{H(0,\,7.05)} = 3.64$. The model of H1 was specified using a half-normal distribution (indicated by the subscript H in the notation and the solid black line in Figure 1A and 1B). The half-normal distribution represents several aspects of the theory: (I) only effects in one direction are considered plausible, these are represented as positive in the analysis, (II) smaller effect sizes are considered more plausible than larger effect sizes, and (III) effect sizes up to twice the approximate scale-of-effect are considered plausible (the approximate scale-of-effect was 7.05, based on the results of Warmelink, 2012).

To see why caution should be exercised before calculating a posterior distribution, consider Figure 1. The red dotted line in Panel A shows the actual posterior distribution for the data in this example. There is nothing particularly disconcerting about Panel A – the most plausible effect sizes according to the posterior distribution lie midway between the most plausible effect sizes according to the model of H1 (i.e., those close to zero), and the most plausible effect sizes according to the data (i.e., those close to the observed mean difference of 9.41).

Now consider Panel B. The red dotted line in Panel B shows the posterior distribution assuming that Warmelink et al. (2019) had observed the same magnitude of effect, but in the opposite direction predicted by theory. The posterior distribution is clearly not an accurate estimation of effect sizes in light of the data. The model of H1 assigned zero plausibility to results in the opposite direction to the prediction and consequently so does the posterior distribution.

In sum, posterior distributions can be informative for estimation purposes when a vague prior model has been used, but researchers should exercise caution before automatically calculating posterior distributions when reporting informed Bayes factors.

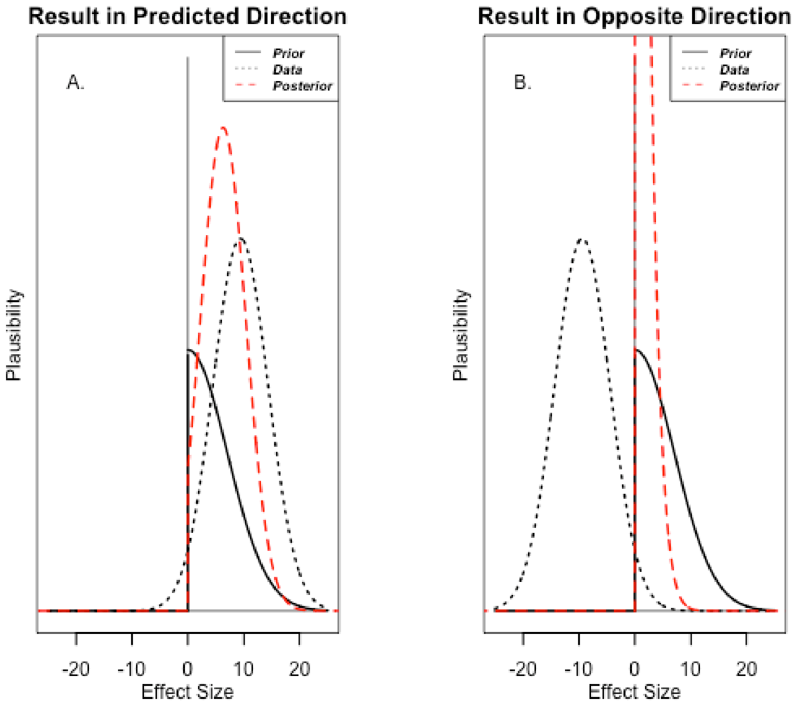

Figure 1. Both panels present the model of H1 used by Warmelink et al (solid black line), the model of the data (dotted black line) and the posterior distribution (dashed red line). Panel A shows the actual results. Panel B shows the posterior distribution for counterfactual data with the same effect size, but in the opposite direction to what was predicted.

Report Robustness Regions to demonstrate the robustness of conclusions

In line with Jeffrey's platitude that "the answer should depend on the question", the result of the Bayes factor depends on the specified models of H1. APA guidelines state that one should therefore report how sensitive the inference is to the specification of the model of

H1 whenever reporting Bayes factors (Appelbaum et al., 2018, p. 20)[1]. This can be done by reporting robustness regions alongside Bayes factors (for examples see Lakens et al., 2020; Lin, et al., in press). Robustness regions communicate the extent to which the inference drawn from a Bayes factor is dependent on the model of H1 used to calculate the Bayes factor. If all plausible effect sizes that could be used to model H1 fall within the robustness region, the inference is robust; if plausible effect sizes that could be used to model H1 fall outside the robustness region, the inference is not robust.

Robustness regions are reported as *RR*[*LL, UL]* where LL corresponds to the smallest scale-of-effect and UL corresponds to the largest scale-of-effect that draw the same conclusion as the model of H1 used to calculate the Bayes factor assuming the specified minimum cut-off for evidence is constant. Consider again the Bayes factor discussed above: $B_{\mathrm{H}(0,\ 7.05)} = 3.64$. Warmelink et al. (2019) concluded that they had obtained moderate evidence for an effect, when specifying their model of H1 with a scale-of-effect of 7.05 and a minimum cut-off of $B = 3$.

To create robustness regions, we calculated a series of Bayes factors, each time manipulating the scale-of-effect to identify the smallest and largest scale-of-effect that still produces $B > 3$ (see DeBruine and Dienes, 2020, for R code that calculates Robustness Regions non-iteratively). In this case, the robustness region is *RR*[4.44, 15.74]. Whether this can be considered robust or not depends on the scientific context. If effects outside the robustness region are scientifically plausible, the inference is not robust. If the robustness regions are judged to cover all theoretically plausible effect sizes, the inference is robust. In this example, we conclude that the results are not robust. Many deception researchers may argue that an effect smaller than 4.44 details would be too small to be of interest. However,

[1] At the time of writing Warmelink et al. (2019), we did not know sensitivity measures were required, hence they were not included.

an effect larger than 15.74 can be considered plausible and interesting (e.g. Vrij et al., 2018, report a mean difference of 29 details between two interviewing conditions)[2].

Estimate the sample size required to obtain conclusive evidence

As Mac Giolla and Ly highlight, 19 of the 22 Bayes factors reported by Warmelink et al. (2019) were inconclusive. Here we provide a simple method (with corresponding R script available at https://osf.io/34nb6/, based on original by Palfi & Dienes, 2019) that researchers can adopt to estimate the number of participants required in order to obtain a pre-specified level of evidence (originally outlined by Dienes, 2015).

As sample size increases, and the standard error reduces towards zero, Bayes factors are better able to distinguish between evidence for two competing hypotheses. The process requires one to (i) estimate an approximate scale-of-effect if H1 were true (for a tutorial on how to estimate what your theory predicts, see Dienes, 2019), (ii) estimate the standard error for a given sample size (e.g., the standard error from a past study), and (iii) specify a minimum level of evidence one wishes to obtain. The R script then simulates the standard error in light of increasing sample size, and calculates Bayes factors for each sample size. It does this assuming the evidence supports H1 and H0 separately. The script also outputs a graph to show the Bayes factor for H0 and H1 at each tested sample size. Figure 2 shows the graph created assuming a researcher wanted to replicate the result from Warmelink et al. (2019), $B_{H(0,\,7.05)} = 3.64$, with a minimum cut-off of 3. In this case, a replication would require 130 participants to obtain a Bayes factor of 3 in favour of H1 assuming the study supported H1, or 328 participants to obtain a Bayes factor of 3 in favour of H0 assuming the study supported H0[3]. It is common that more participants are required to obtain evidence for the

[2] To interpret robustness regions, it can be useful to consider the smallest effect of interest. Dienes (2020) provides guidelines on how to do so. If one is uncertain whether a finding is robust, however, one can continue to collect data until the robustness region extends below any plausible smallest effect size of interest.

[3] Assumptions include that the effect size and sample error obtained by Warmelink et al (2019) are accurate.

point-null hypothesis than for the alternative hypothesis. In this example, to ensure enough participants are recruited that the study could obtain sufficient evidence for either hypothesis, a replication study should look to recruit at least 328 participants.

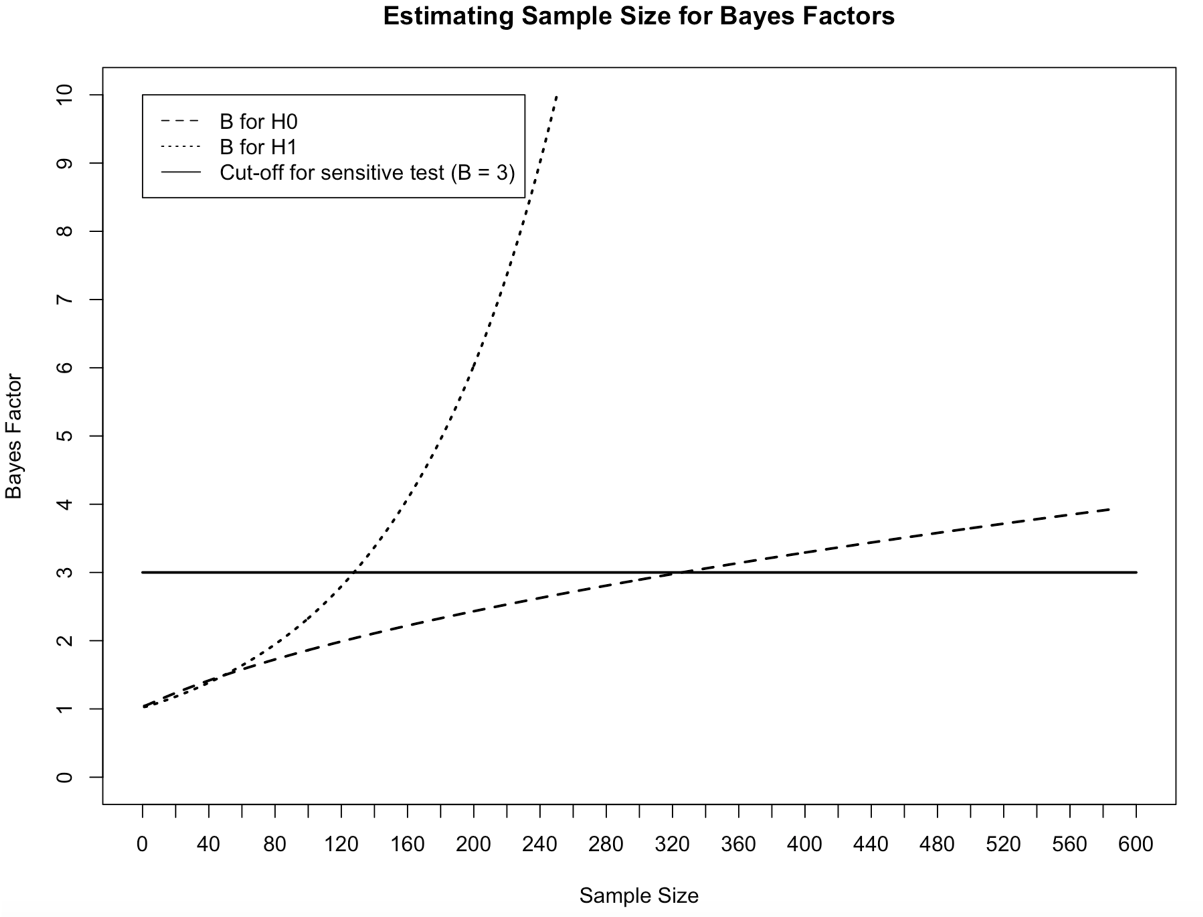

Figure 2: The output from the supplemental R script showing the estimated sample size required to obtain differing levels of evidence, based on the data of Warmelink et al. (2019).

Conclusion

We whole-heartedly agree with Mac Giolla and Ly when they encourage researchers to design more informative studies and we are grateful for their commentary that sought to educate deception researchers on matters of reporting and interpreting Bayesian analyses. Here, we have sought to add a number of clarifications, and also specified two additional ways that the Bayesian analyses in Warmelink et al. (2019) could have been more informative. If researchers adopt these recommendations, we believe they will be better able to assess the evidence provided by their research.

Box 1

There is potential for confusion between the different uses of the word *prior* and *posterior* in Bayesian analyses. Here we provide an overview.

Prior and posterior odds: Measures of relative belief

Prior odds specify the relative degree of conviction a researcher has in two theories. If a person believes each hypothesis is equally plausible, that person has prior odds of P(H1) / P(H0) = .50/.50 = 1. Prior odds are updated by the level of evidence in the data (i.e., the Bayes factor) to derive the posterior odds: Prior odds x Bayes factor = Posterior odds. The posterior odds specify the relative degree of conviction a researcher has in two hypotheses following data collection.

Prior and posterior distributions: Specifying the plausibility of effect sizes

A prior model is a probability distribution that shows the plausibility of different population effect sizes (x-axis = possible effect sizes, y-axis = plausibility, see Figure 1). Warmelink et al (2019) used *informed Bayes factors*, meaning that the prior model was specified to represent the theory being tested (Dienes, 2019, Box 4). Their prior models were created by specifying the approximate effect size (i.e., the "scale-of-effect") that is plausible if the theory is correct. Prior models can be updated with the model of the data to create the posterior distribution. The posterior distribution demonstrates the plausibility of population effect sizes following data collection assuming the hypothesis is true. Note that "distribution" and "model" are used interchangeably.

In order to avoid confusion, we adopt Dienes's (2020) vernacular, referring to the prior model as the "model of H1", to reflect the fact that the prior model represents predictions made by the theory.

Bibliography

Aczel, B., Hoekstra, R., Wagenmakers, E.-J., Klugkist, I., Rouder, J. N., Vandekerckhove, J., Lee, M., Morey, R. D., Dienes, Z., & van Ravenzwaaij, D. (2018). *Expert opinions on how to conduct and report Bayesian inference*.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., & Camerer, C. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

DeBruine, L., & Dienes, Z. (2020). *Bayes Factors and Robustness Regions*. Bfrr. Retrieved 10 February 2020, from https://debruine.github.io/bfrr/

Dienes, Z. (2015). *How many participants might I need?* Youtube. https://www.youtube.com/watch?v=10Lsm_o_GRg&t=195s

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, *2*(4), 364–377.

Dienes, Zoltan. (2020). *How to use and report Bayesian hypothesis tests* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/bua5n

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, *75*(1), 45–57.

Lin, L., McLatchie, N., & Linkenauger, S. (In press). The influence of perceptual-motor variability on the perception of action boundaries for reaching. *Journal of Experimental Psychology: Human Perception and Performance*.

Mac Giolla, E., & Ly, A. (2019). What to do with all these Bayes factors: How to make Bayesian reports in deception research more informative. *Legal and Criminological Psychology*.

Morey, R. (2015, January 30). BayesFactor: Software for Bayesian inference: On verbal categories for the interpretation of Bayes factors. *BayesFactor*. https://bayesfactor.blogspot.com/2015/01/on-verbal-categories-for-interpretation.html

Palfi, B., & Dienes, Z. (2019). Why Bayesian "evidence for H1" in one condition and Bayesian "evidence for H0" in another does not mean good enough Bayesian evidence for a difference between conditions. *Advances in Methods and Practices in Psychological Science*.

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

Vrij, A., Leal, S., Jupe, L., & Harvey, A. (2018). Within-subjects verbal lie detection measures: A comparison between total detail and proportion of complications. *Legal and Criminological Psychology*, *23*(2), 265–279.

Warmelink, L. (2012). *Lying about intentions (Unpublished doctoral thesis)*. University of Portsmouth.

Warmelink, Lara, Subramanian, A., Tkacheva, D., & McLatchie, N. (2019). Unexpected questions in deception detection interviews: Does question order matter? *Legal and Criminological Psychology*, *24*(2), 258–272.