UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Nalborczyk, L., et al. (2019). Pragmatism should Not be a Substitute for Statistical Literacy, a Commentary on Albers, Kiers, and Van Ravenzwaaij (2018). *Collabra: Psychology,* 5(1): 13. DOI: https://doi.org/10.1525/collabra.197

ORIGINAL RESEARCH REPORT

Pragmatism should Not be a Substitute for Statistical Literacy, a Commentary on Albers, Kiers, and Van Ravenzwaaij (2018)

Ladislas Nalborczyk[*,†], Paul-Christian Bürkner[‡] and Donald R. Williams[§]

Based on the observation that frequentist confidence intervals and Bayesian credible intervals sometimes happen to have the same numerical boundaries (under very specific conditions), Albers et al. (2018) proposed to adopt the heuristic according to which they can usually be treated as *equivalent*. We argue that this heuristic can be misleading by showing that it does not generalise well to more complex (realistic) situations and models. Instead of pragmatism, we advocate for the use of parsimony in deciding which statistics to report. In a word, we recommend that a researcher interested in the Bayesian interpretation simply reports credible intervals.

Keywords: Bayesian statistics; Bayes; confidence interval; credible interval

Context

The main line of reasoning of Albers et al. (2018) seems to be the following: as frequentist confidence intervals and Bayesian credible intervals sometimes happen to be similar, we can usually interpret them the same way. More precisely, they argue that because confidence intervals and credible intervals do sometimes have the same numerical boundaries (and because when they do, they have similar consequences on the inference being made), then, from a pragmatic perspective, they should be treated as *equivalent*.

While we agree with their main observation (i.e., that confidence intervals and credible intervals obtained with uninformative priors might sometimes coincide), we disagree with their main conclusion (i.e., that confidence intervals can be interpreted as credible intervals). We think the examples presented in Albers et al. (2018) are overly simplistic and actually quite rare. Moreover, the pragmatic stance tends to blur the distinction between credible intervals and confidence intervals, whereas we think it would be more efficient, at a pedagogical level, to emphasise these differences.

* University of Grenoble Alpes, CNRS, LPNC, 38000, Grenoble, FR

† Department of Experimental Clinical and Health Psychology, Ghent University, BE

‡ Department of Psychology, University of Münster, DE

§ Department of Psychology, University of California, Davis, US

Corresponding author: Ladislas Nalborczyk (ladislas.nalborczyk@univ-grenoble-alpes.fr)

Rebuttals

Conditioning on impossible values

The debate between the frequentist and the Bayesian schools of inference has been firing for many decades and we do not wish to reiterate all the arguments here (we refer the interested reader to the introduction of Albers et al., 2018). Bayesian statistics rest on the use of Bayes' rule, which states that:

$$p(\theta \mid y) \propto p(y \mid \theta) \times p(\theta)$$

In other words, the posterior probability of some parameter (or vector of parameters) θ is proportional to the product of its prior probability $p(\theta)$ and the likelihood $p(y|\theta)$. Noteworthy here is that the posterior probability $p(\theta|y)$ can be interpreted as a *conditional* probability, *given* the data *and* the model (including the prior information).

This highlights a first undesirable consequence of Albers et al.'s (2018) proposal. Using confidence intervals (or credible intervals with flat priors) to make probability statements can lead to nonsensical situations. For instance, let's say you're fitting a simple linear regression model to estimate the average reaction time in some cognitive task.[1] Using a confidence interval to make a probability statement (under the pretence that it is numerically similar to a credible interval) is akin to implicitly assuming a uniform prior over the reals. It means assuming that all values between $-\infty$ and ∞ are equally plausible, including negative values. This is an inappropriate assumption when dealing with reaction times, proportions, scales scores, most physical

measurements (e.g., weight, height), or anything else that has a restricted range of definition.

Further, there are examples where numerically equivalent intervals do not necessarily reflect the most probable parameter values (given all available information), but could still have valid frequentist properties. Indeed, whereas both Bayesian and frequentist intervals could have nominal coverage probabilities (Albers et al., 2018), the additional requirement for (meaningful) probabilistic inference is compatibility with previous information. Rather, in addition to the data, the probabilities are also conditional on all assumptions including the prior distribution. To make this point, we use a recent example from a registered replication report (Verschuere et al., 2018). The original effect was reported as $d = 1.45$, 95% [0.29, 2.61] (Mazar, Amir, & Ariely, 2008). Following the argument of Albers et al. (2018), we could state there is a 50% chance the effect is greater than 1.45. Although this would be mathematically correct for the posterior distribution (Gelman 2013), this does not mean it accurately reflects the most probable values. Indeed, based on the priming literature, it would be unreasonable to make such a probability statement. On the other hand, we could envision such a wide interval (Bayesian or frequentist) covering the population value 95% of the time. Thus, interpretive exchangeability is not a given and can lead to misleading inferences when conditioning on impossible values. We now move to a discussion of two concrete examples examining the generalisability of the heuristic suggested by Albers et al. (2018) in regards to the coverage properties (and the numerical boundaries) of confidence intervals and credible intervals.

Frequentist properties of Bayesian credible intervals

A simple regression example

In **Figure 1**, we present some simulation results showing that Bayesian credible intervals (obtained with weakly informative priors) do have the same properties as frequentist confidence intervals in the case of a simple regression model. Indeed, on repeated sampling, X% of the constructed intervals will contain the population value of θ (as expressed by the coverage proportion displayed in **Figure 1**).

Bayesian credible intervals with non-informative or weakly informative priors may have the same frequentist characteristics as confidence intervals, but also allow for conditional probability statements (e.g., given the prior and the information contained in the data, we can say that there is a X% probability that the population value of θ lies in the interval).[2] Therefore, in simple situations, the principle of parsimony would lead to use and report the most inclusive (general) statistics. Thus, we suggest that the researcher interested in the Bayesian

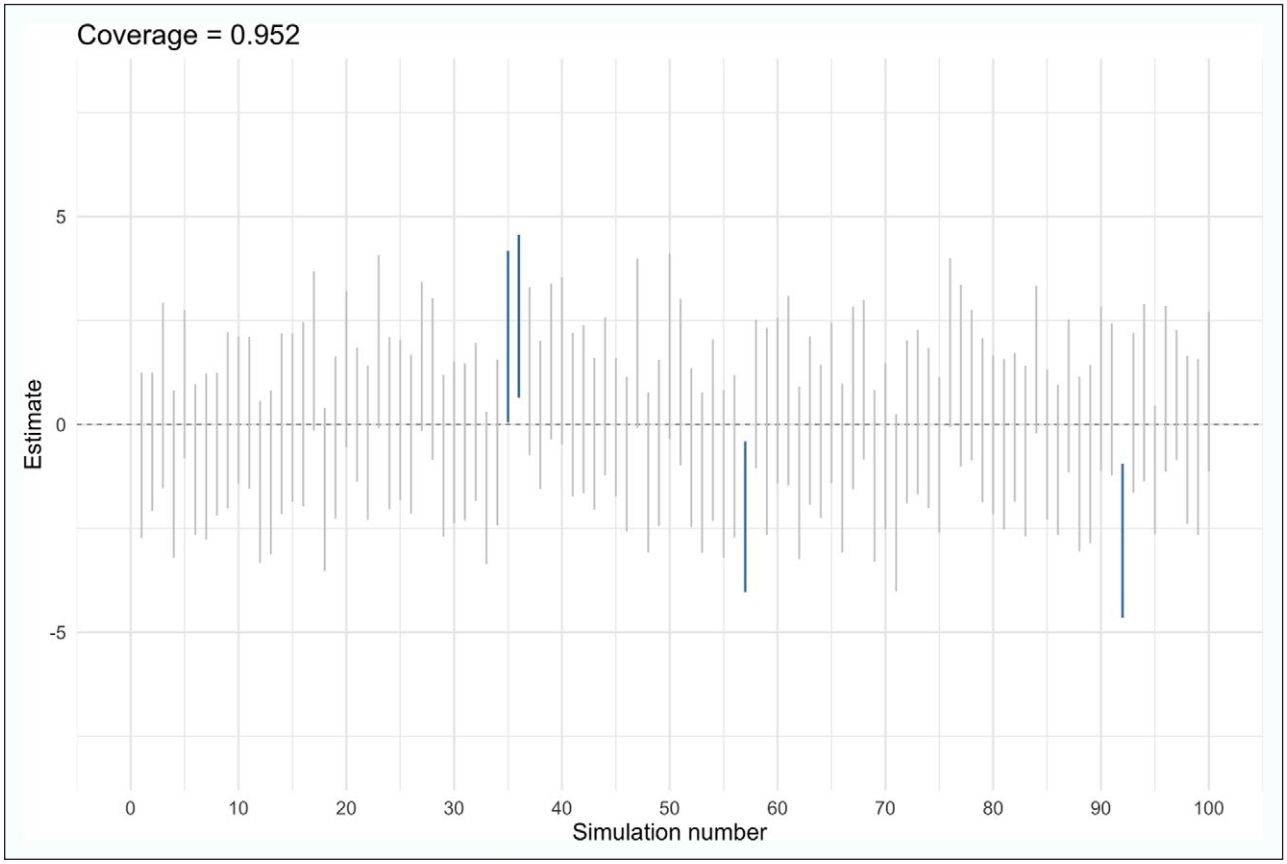

Figure 1: Coverage properties of Bayesian credible intervals when using weakly informative priors. Blue vertical credible intervals represent intervals that "missed" the population value of the parameter (whose value is represented by the horizontal dashed line), while grey intervals represent intervals that contained the population value. Note: for readability, only the first 100 simulations are plotted.

interpretation should use and report Bayesian credible intervals.

What about more complex models?

In this section, we report simulation results of the coverage properties of both confidence and credible intervals around the amount of heterogeneity τ in random-effects meta-analysis models.

The effect sizes to be combined in meta-analyses are often found to be more variable than it would happen because of sampling alone. The usual way to take into account this heterogeneity is to use random-effects models (also known as multilevel models). Several methods have been proposed to obtain confidence intervals around the point estimate of τ in such models (for a discussion, see Williams, Rast, & Bürkner, 2018). The method developed by Paule and Mandel (1982) and implemented in the `metafor` package (Viechtbauer 2010) guarantees nominal coverage probabilities of confidence intervals computed with this method, even in small samples, given that model assumptions are satisfied. Below we compare the coverage properties of confidence intervals (computed with this method) and credible intervals for a simple random-effects meta-analysis model of 6 studies, with a population value of $\tau = 0.1$ (see code in supplementary materials for more details).

As shown in **Figure 2**, the coverage proportion of confidence intervals is close to the nominal 95% value. However, the credible intervals (wider than the confidence intervals) appear to contain the population value of τ in almost all 10.000 simulations, resulting in a coverage proportion close to 1.

Thus, even when using non-informative priors (we used $\tau \sim$ HalfCauchy(1000)), the numerical boundaries as well as the coverage properties of confidence intervals and credible intervals can differ considerably. More generally, we feel that using simplistic examples to make general claims is highly problematic in that there is no guarantee that this generalises well to more complex models.

Differences matter

Albers et al. (2018) write: "In the present paper, we have demonstrated by means of various examples that confidence intervals and credible intervals, in various practical situations, are very similar and will lead to the same conclusions for many practical purposes when relatively uninformative priors are used".

Contrary to what the authors postulate, differences between confidence intervals and credible intervals are observable in a large variety of situations (actually, all but one). For instance (but non exhaustively), i) when samples are small, ii) when the space of the outcome is multi-modal or non-continuous, iii) when the range of the outcome is restricted, or iv) when the prior is at least weakly informative. Combining these four possibilities, we argue that confidence intervals and credible intervals actually almost never give similar results. Moreover, as we previously demonstrated, numerical estimates can be

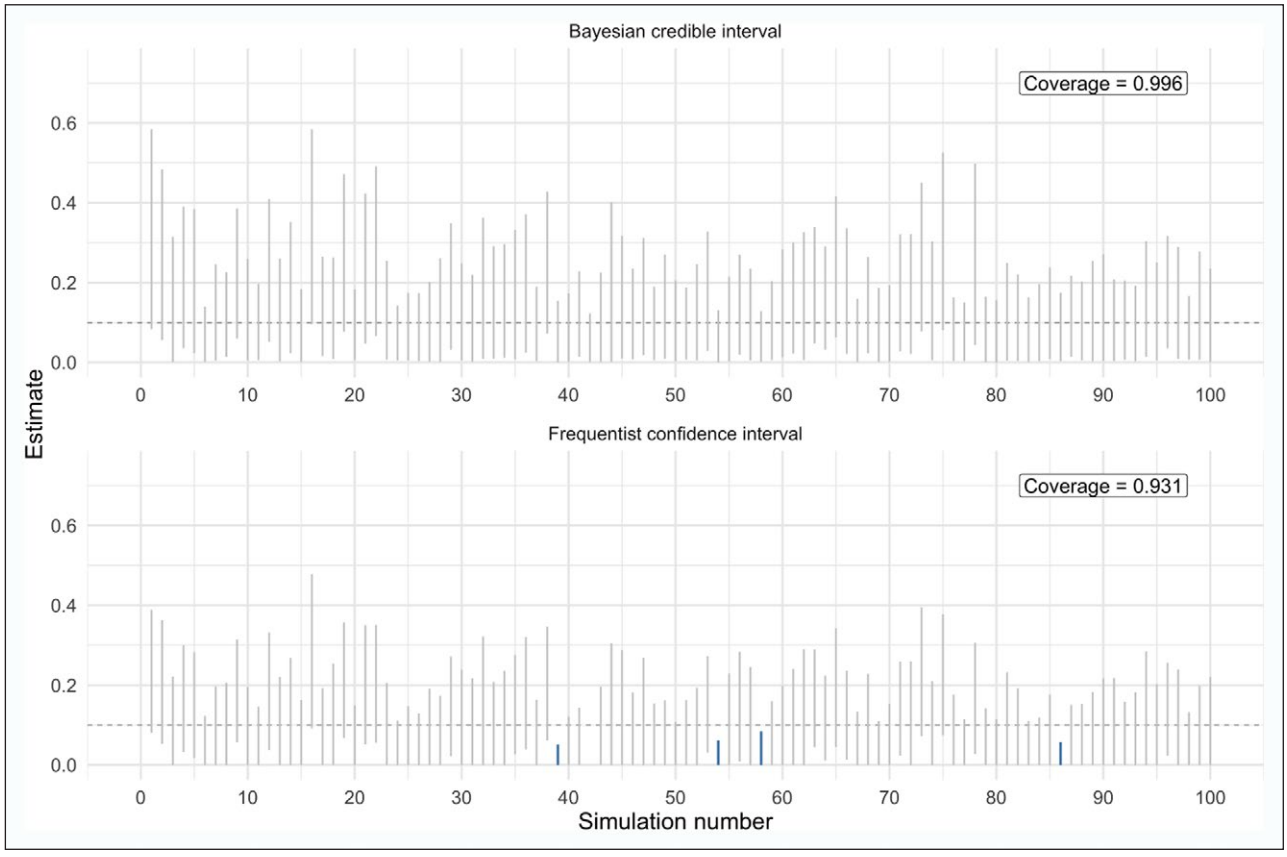

Figure 2: Coverage properties of 95% confidence intervals and 95% credible intervals for recovering the amount of heterogeneity in random-effects meta-analysis models. Note: For readability, only the first 100 simulations are plotted.

similar, but it does not entail that the conclusion we can draw from it (i.e., the inference being made) should be similar.

In the previous sections, we discussed why we think the heuristic suggested by Albers et al. (2018) can be misleading. In the following, we suggest an alternative to pragmatism which does not preclude statistical literacy.

An alternative to pragmatism

Applying parsimony in scientific and statistical practise

Albers et al. (2018) write: "By recognizing the near-equivalence between Bayesian and frequentist estimation intervals in 'regular cases', one can benefit from both worlds by incorporating both types of analysis in their study, which will lead to additional insights."

Confidence intervals can sometimes (i.e., under specific conditions) be identified with a special case of credible intervals for which priors are non-informative. Thus, one could ask, in consideration of the parsimony principle, why reporting redundant statistics? Would not it be easier to use the more general and flexible case? The parsimonious stance that we adopt here leads to the conclusion that the researcher interested in one specific interpretation should report the statistics that corresponds to this goal.[3] If a researcher is interested in the sampling distribution of the statistics under study (or in reaching a nominal coverage proportion), she should report confidence intervals. If she is rather interested in making conditional probability statements from the data, then s·he should report credible intervals (or ideally, the full posterior distribution).

A brief note on the frequentist properties of Bayesian procedures

Albers et al. (2018) quote Bayarri and Berger (2004) that wrote: "Statisticians should readily use both Bayesian and frequentist ideas".

We could not agree more with this statement. In addition, we recognise that both statistical traditions have their own advantages and drawbacks, and have been built to answer somehow different questions. Therefore, pretending that a statistic issued from one school of inference can be interpreted as a statistic issued from another school because they sometimes (under very restricted conditions) give the same numerical estimates is confusing and misleading.

Conclusions and practical recommendations

To sum up, we feel that every proposal going in the direction of more fuzziness in the distinction between different kinds of intervals is misleading and should be rejected. Using a confidence interval as a credible interval or using a credible interval as a confidence interval seems inappropriate to us, as it tends to blur the distinction between essentially different statistical tools. Instead, we prefer to emphatically teach and discuss the differences between these tools and their domains of application. As Hoekstra, Morey, and Wagenmakers (2018), we believe that "the more pragmatic approach in which philosophically unsound interpretations of CIs are permitted and even endorsed is unhelpful, and should be replaced by a more principled one. If students are to learn a certain statistical technique, expecting from statistics teachers to guard them against quick-and-dirty versions seems very reasonable indeed".

Given the limitations of the pragmatic perspective offered by Albers et al. (2018) and the potentially harmful consequences of the heuristic they argued for, we rather suggest to use parsimony as a guiding principle in deciding which statistics to use and to report. In brief, we recommend that a researcher interested in the Bayesian interpretation of an interval simply reports credible intervals (or that a researcher interested in the coverage properties of confidence intervals simply reports confidence intervals).

Data Accessibility Statement

Reproducible code and figures are available at: https://osf.io/nmp6x/.

Notes

[1] Which is given by the intercept of the model, if no predictor is included, or if these predictors have been contrast-coded.

[2] Although, as we discussed earlier, this probability statement, while valid, makes little sense knowing that it is conditional on all possible values being equally likely a priori.

[3] Obviously, it is perfectly legitimate to be interested in several goals, but these goals should be clearly stated as such, and pursued using appropriate tools.

Acknowledgements

We thank Antonio Schettino and Ivan Grahek for helpful comments on a previous version of this manuscript, as well as the original authors and one anonymous reviewer for their suggestions during the review process.

Competing Interests

The authors have no competing interests to declare.

Author Contribution

LN wrote a first version of the manuscript and conducted the simulations for the regression example. DW wrote a part of the paper and conducted the simulations for the meta-analysis example. DW and PB critically commented on various versions of the manuscript. All authors contributed to writing of the final manuscript.

References

Albers, C., Kiers, H., & **van Ravenzwaaij, D.** (2018). Credible Confidence: A pragmatic view on the frequentist vs Bayesian debate. *Collabra: Psychology*, *4*(1), 31. DOI: https://doi.org/10.1525/collabra.149

Bayarri, M. J., & **Berger, J. O.** (2004). The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*, *19*(1), 58–80. DOI: https://doi.org/10.1214/088342304000000116

Gelman, A. (2013). P Values and Statistical Practice. *Epidemiology*, *24*(1), 69–72. DOI: https://doi.org/10.1097/EDE.0b013e31827886f7

Hoekstra, R., Morey, R. D., & **Wagenmakers, E.-J.** (2018). Improving the interpretation of confidence and credible intervals. In: *Looking back, looking forward.* Kyoto, Japan.

Mazar, N., Amir, O., & **Ariely, D.** (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*, 633–644. DOI: https://doi.org/10.1509/jmkr.45.6.633

Paule, R., & **Mandel, J.** (1982). Consensus Values and Weighting Factors. *Journal of Research of the National Bureau of Standards*, *87*(5), 377. DOI: https://doi.org/10.6028/jres.087.022

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., Kirchler, M., et al. (2018). Registered Replication Report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, *1*(3), 299–317. DOI: https://doi.org/10.1177/2515245918781032

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. DOI: https://doi.org/10.18637/jss.v036.i03

Williams, D. R., Rast, P., & **Bürkner, P.-C.** (2018). Bayesian Meta-Analysis with Weakly Informative Prior Distributions. *PsyArXiv*. Retrieved from: https://psyarxiv.com/7tbrm/. DOI: https://doi.org/10.31234/osf.io/7tbrm

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.197.pr

How to cite this article: Nalborczyk, L., Bürkner, P.-C., & Williams, D. R. (2019). Pragmatism should Not be a Substitute for Statistical Literacy, a Commentary on Albers, Kiers, and Van Ravenzwaaij (2018). *Collabra: Psychology,* 5(1): 13. DOI: https://doi.org/10.1525/collabra.197

Senior Editor: Victoria Savalei

Editor: Victoria Savalei

Submitted: 27 September 2018 **Accepted:** 03 Febuary 2019 **Published:** 26 March 2019

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS