A Cautionary Note on Estimating Effect Size

Don van den Bergh*[1], Julia M. Haaf[1], Alexander Ly[1,2], Jeffrey N. Rouder[3], and Eric-Jan Wagenmakers[1]

[1]University of Amsterdam

[2]Centrum Wiskunde & Informatica

[3]University of California Irvine

Abstract

An increasingly popular approach to statistical inference is to focus on the estimation of effect size while ignoring the null hypothesis that the effect is absent. We demonstrate how this common "null hypothesis neglect" may result in effect size estimates that are overly optimistic. The overestimation can be avoided by incorporating the plausibility of the null hypothesis into the estimation process through a "spike-and-slab" model.

Consider the following hypothetical scenario: a colleague from the biology department has just conducted an experiment and approaches you for statistical advice. The analysis yields $p < 0.05$ and your colleague believes that this is grounds to reject the null hypothesis. In line with recommendations both old (e.g., Grant, 1962; Loftus, 1996) and new (e.g., Cumming, 2014; Harrington et al., 2019) you convince your colleague that it is better to replace the p-value

*Correspondence concerning this article should be addressed to: Don van den Bergh, University of Amsterdam, Department of Psychological Methods, Nieuwe Achtergracht 129B, 1018VZ Amsterdam, The Netherlands. E-Mail should be sent to: donvdbergh@hotmail.com. This work was supported by a Research Talent grant from the Netherlands Organisation of Scientific Research (NWO) to DvdB and an Advanced ERC grant 743086 UNIFY to EJW.

with a point estimate of effect size and a 95% confidence interval (but see Morey et al., 2016). You also manage to convince your colleague to plot the data (see Figure 1). Mindful of the reporting guidelines of the *Psychonomic Society*[1] and *Psychological Science*[2], your colleague reports the result as follows: "Cohen's $d = 0.30$, CI $= [0.02, 0.58]$".

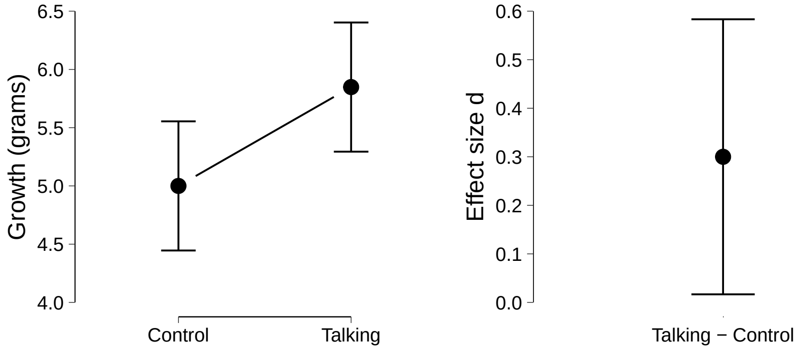

Figure 1: Standard estimation results for the fictitious plant growth example. Left panel: a descriptives plot with the mean and 95% confidence interval of plant growth in the two conditions. Right panel: point estimate and 95% confidence interval for Cohen's d.

Based on these results, what would be a reasonable point estimate of effect size? A straightforward and intuitive answer is "0.30". However, your colleague now informs you of the hypothesis that the experiment was designed to assess: "plants grow faster when you talk to them".[3] Suddenly, a population effect size of "0" appears eminently plausible. Any observed difference may merely be due to the inevitable sampling variability.

The example above raises the question: When are effect sizes overestimated? Standard point estimates and confidence intervals ignore the possibility that the effect is spurious (i.e., the null hypothesis $\mathcal{H}_0$). This is not problematic when $\mathcal{H}_0$ is deeply implausible, either because $\mathcal{H}_0$ was highly unlikely *a priori* or because the data decisively undercut $\mathcal{H}_0$. But when the data fail to undercut $\mathcal{H}_0$, or when $\mathcal{H}_0$ is highly likely *a priori* (i.e., "plants do not grow faster when you talk

[1] https://www.springer.com/psychology?SGWID=0-10126-6-1390050-0

[2] https://www.psychologicalscience.org/publications/psychological_science/ps-submissions#STAT

[3] This example is inspired by Berger and Delampady (1987).

to them"), then $\mathcal{H}_0$ is not ruled out as a plausible account of the data. Effect size estimates that ignore a plausible $\mathcal{H}_0$ are generally biased and overconfident: the fact that $\mathcal{H}_0$ provides an acceptable account of the data should shrink effect size estimates towards zero.

In this paper, we discuss the spike-and-slab model and its merits to avoid overestimating effect size. First, we formally introduce the spike-and-slab model. Second, we apply the spike-and-slab model to the example in the introduction and illustrate how it circumvents overestimation. Third, we illustrate some properties of the spike-and-slab model in a simulation. Fourth, we demonstrate the spike-and-slab model by reanalyzing the data of Heycke et al. (2018). Finally, we conclude with practical recommendations and a discussion on when not to use the spike-and-slab model.

A Spike-and-Slab Perspective

Here we formalize the conceptual ideas from the introduction in the spike-and-slab model (Clyde et al., 1996; Mitchell & Beauchamp, 1988; Rouder et al., 2018). The population effect size is typically approximated with a sample estimate. This sample estimate assumes that $\mathcal{H}_0$ is false and that the population effect size is nonzero. To formalize this, let δ denote the population effect size, let $\hat{\delta}$ denote an estimate, and let $\hat{\delta} \mid \mathcal{H}_1$ denote an estimate that assumes that $\mathcal{H}_1$ is true. Assuming that $\mathcal{H}_0$ is true leads to $\hat{\delta} \mid \mathcal{H}_0$, which is usually 0. Key is that both estimates, $\hat{\delta} \mid \mathcal{H}_1$ and $\hat{\delta} \mid \mathcal{H}_0$, are *conditional* on the hypotheses. For example, $\hat{\delta} \mid \mathcal{H}_1$ should be read as "the estimated effect size given that the alternative hypothesis is true".

The first component, the spike, corresponds to the position that talking to plants does not affect their growth (i.e., $\delta = 0$), whereas the second component, the slab, corresponds to the position that speaking to plants does affect their growth (i.e., $\delta \neq 0$). The spike and slab are analogous to $\mathcal{H}_0$ and $\mathcal{H}_1$ discussed above. Both components are deemed *a priori* equally likely, such that the prior probability for each component is $^1/_2$. After seeing the data, the prior probabilities of each component, $\Pr(\text{spike})$ and $\Pr(\text{slab})$, are updated to posterior probabilities, $\Pr(\text{spike} \mid \text{data})$ and $\Pr(\text{slab} \mid \text{data})$.

Quantifying the uncertainty about the two components using probabilities allows us to account for the uncertainty of each component. Furthermore, this allows us to consider the *marginal* estimate of effect size. The marginal estimate of the spike-and-slab model weighs the estimates of each component by its plausibility after seeing the data. That is, following the law of total probability, we sum the estimates conditional on each component weighed by the posterior probabilities of the components:

$$\hat{\delta} = \left(\hat{\delta} \mid \text{spike}\right) \Pr(\text{spike} \mid \text{data}) + \left(\hat{\delta} \mid \text{slab}\right) \Pr(\text{slab} \mid \text{data}).$$

To obtain the posterior probabilities, we need to choose an inferential paradigm. In a frequentist approach, one could compute an information criteria for each model, say the Akaike Information Criterion (AIC; Akaike, 1973), and define $\Pr(\text{spike} \mid \text{data}) = {}^{\text{AIC}|\text{spike}}/_{(\text{AIC}|\text{spike}+\text{AIC}|\text{slab})}$. This approach is taken by, for instance, Burnham and Anderson (2002). Another frequentist approach is penalized maximum likelihood. There a penalty term is added to the likelihood before maximizing it. A well-known example is LASSO regression (Tibshirani et al., 2005).

Maybe it makes more sense to refer to ridge regression as there the shrinkage is also smooth.

An alternative is a Bayesian approach. Here we assign probability distributions to unknown parameters, such as effect size. This requires us to specify additional prior distributions, however, the posterior model probabilities follow directly from Bayes theorem.Furthermore, we obtain a posterior distribution for effect size which fully captures the uncertainty over possible effect sizes and allows for easy computation of uncertainty intervals. In the remainder of this paper, we adopt the Bayesian approach for the spike-and-slab model. For a full derivation of the posterior distribution, see Appendix A.

TODO: Actually write this appendix/ box.

To illustrate both the overestimation and the spike-and-slab model we reanalyze the fictitious data from Figure 1. R code for the analysis is available at https://osf.io/uq8st/. In almost all current empirical work, an estimate of effect size is based solely on the second component, which yields a point estimate and an uncertainty interval (for frequentists, $\delta = 0.30$, 95% CI: [0.02, 0.58]; for Bayesians $\delta = 0.29$, 95% CRI: [0.02, 0.57]). The spike-and-slab model, however, also considers the possibility that an effect can be absent; consequently, the overall estimate from the spike-and-slab model is a weighted average of the two components, shrinking the estimate towards zero. Figure 2 contrasts the traditional slab-only estimation against the spike-and-slab estimation.

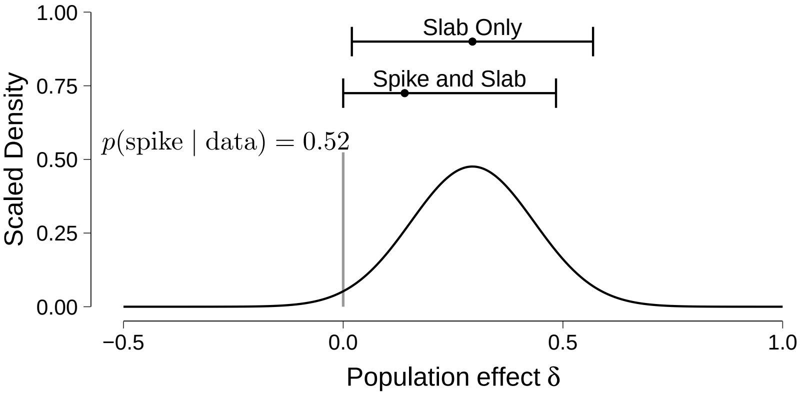

Figure 2: The spike-and-slab model. The black line represents the posterior distribution of effect size given the slab (i.e., the effect is non-zero). The posterior is scaled so that its mode ($\delta = 0.29$) equals the posterior probability of the alternative model (i.e., $p(\mathcal{H}_1 \mid \text{data}) = 0.48$). The grey line represents the posterior probability of the spike (i.e., $\mathcal{H}_0$: the effect is absent). The error bars and dots above the density show 95% credible intervals and the posterior mean for the slab-only model and for the spike-and-slab model.

Compared to the traditional results based only on the slab, the posterior mean and central 95% credible interval of the spike-and-slab model are shrunken towards 0 (i.e., 0.14 (95% CRI: [0.00, 0.48]) vs. 0.29 (95% CRI: [0.02, 0.57])). This shrinkage is due to the non-negligible probability that the effect is absent. The spike-and-slab posterior represents the plausibility that the effect is absent by the height of the spike, and the uncertainty about the effect's magnitude, given that it is present, by the width of the slab. Note that as the posterior probability of $\mathcal{H}_0$ decreases, the spike-and-slab results approach those of the slab-only model.

The Influence of the Spike

In the fictitious example, the spike-and-slab model avoids overestimation by shrinking estimates of effect size towards zero. That result may not be surprising, as the effect was small. However, it makes one wonder to what extent the spike-and-slab model helps with estimation. What differs between a slab only model and the spike-and slab? In this section, we illustrate in a simulation how the estimated effect size shrinks towards zero as a function of the observed effect size, the prior on effect size, the sample size, and the prior probability of the spike. We chose these parameters because the posterior distribution only depends on these quantities.

Figure 3 shows the relation between the observed effect size and the estimated effect size for the slab and for the spike-and-slab for 40 observations and 100 observations. All plots show that a smaller prior standard deviation induces more shrinkage towards zero. This makes sense as a small prior standard deviation implies there is more prior mass near the mean of the prior, which is zero. The influence of the prior standard deviation is intrinsic to a Bayesian approach, but not to the spike-and-slab model. Comparing the plots between the two columns illustrates the influence of the spike; whenever the observed effect size is near zero, the estimate is shrunken towards zero in the right column but not in the left column. However, when the observed effect size is far from zero, there is little shrinkage. The shrinkage can be explained in the following way. Whenever the observed effect size is small, the data are well described by an effect size of zero and thus the posterior probability of the spike is substantial. In contrast, when the observed effect size is large the data are poorly described by an effect size of zero and the posterior probability of the spike is negligible. As a consequence, the estimate of the spike-and-slab is practically equivalent to

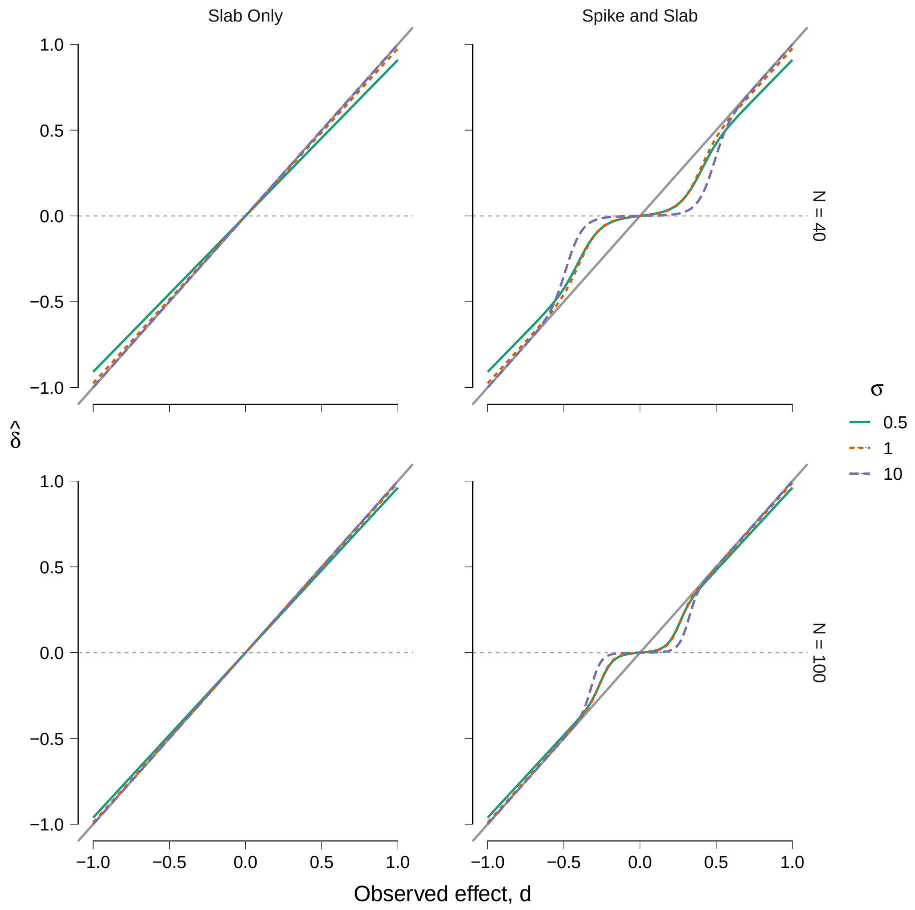

Figure 3: Observed effect size versus posterior mean for different model components and prior standard deviations. The left column shows inference based on the slab only model while the right column shows inference based on the spike-and-slab model. In the top row, the sample size was 40 while in the bottom row the sample size was 100. Different lines represent different standard deviations for the prior distribution on δ. The prior probability of the spike was $1/2$. Inspired by Figure 5 of Rouder et al. (2018).

the estimate of the slab. The plots in the right column of Figure 3 shows the effect of sample size on the shrinkage. For the bottom right plot, $N = 100$, if the observed effect size is small then the estimate is still shrunken towards 0, but as the observed effect size grows the shrinkage decreases much more quickly than in the top right plot where $N = 40$. This makes sense from a signal-detection perspective. If the observed effect size is, for example, 0.3 after 40 observations, the posterior probability of the spike is substantial. However, after collecting 60 additional observations while the observed effect size remains 0.3, the posterior probability of the spike decreases as it becomes increasingly less probable that the data generating model had an effect size of zero.

Next, we explore the relationship between shrinkage and the prior probability of the spike. Figure 4 shows the shrinkage for various prior probabilities. The smaller the prior probability of the spike, the less the effect size is shrunken towards 0. If the prior probability is small then the spike was a-priori implausible and less evidence is needed to make its influence negligible.

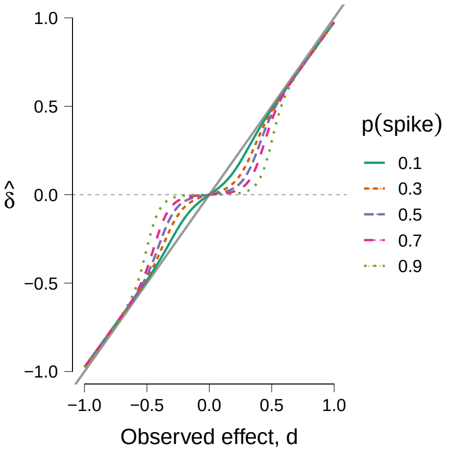

Figure 4: Observed effect versus posterior mean (x-axis) versus the posterior mean of the spike-and-slab model (y-axis). The different lines represent different standard prior probabilities of the spike. The figure is based on 40 observations with a prior standard deviation of 1.

Empirical Example: Reanalysis of Two Minds

To improve upon the previous fictional example, we highlight how the spike-and-slab approach can be used in psychological practice by reanalyzing the results of Heycke et al. (2018). Heycke et al. (2018) conducted two registered replications of Rydell et al. (2006). We first briefly explain the design of the study before reanalyzing the results, for a detailed description see the "Procedure" section in Heycke et al. (2018). Afterward, we reanalyze the Explicit Evaluation and Implicit Evaluation analyses with a spike-and-slab model. Finally, we provide a robustness analysis.

Participants were shown a positive or negative prime on a computer screen followed by an image of a person. Next, several behavioral descriptions that were either negative or positive appeared underneath the image of the person. Participants were asked to indicate whether the behavioral description was characteristic or uncharacteristic for the person shown. After the first block of trials, participants explicitly evaluated the target person and performed an implicit association task (IAT). In total, data of 51 participants was analyzed.

Explicit Evaluation In the analysis of the explicit evaluations, Heycke et al. (p. 10; 2018) conducted a paired t-test and concluded that the rating of the target character is more positive if positive information is shown before negative information: $t(27) = 11.52$, $p < .001$; $\text{BF}_{10} = 1.37 \times 10^9$, $d = 2.09$, 95% HDI $[1.41, 2.79]$. The magnitude of the effect is large and thus a spike-and-slab reanalysis yields practically the same results: $\hat{\delta} = 2.10$, 95% CRI: [1.74, 2.47].[4]

Implicit Evaluation In the analysis of the IAT, Heycke et al. (p. 10; 2018) conducted a paired t-test and concluded that when negative primes were presented before positive primes there was some indication that the IAT rating became more negative: $t(27) = -2.54$, $p = .017$, $\text{BF}_{10} = 2.92$, $d = -0.44$, 95% HDI $[-0.83, -0.06]$. Here, the magnitude of the effect is smaller and as a consequence the results from the spike-and-slab reanalysis are more conservative: $\hat{\delta} = -0.35$, 95% CRI: $[-0.75,\ 0.00]$. The estimate of effect size is shrunken towards 0 because the spike provides a reasonable account of the data, $\Pr(\text{spike} \mid \text{data}) = 0.25$.

Robustness analysis In the reanalyses above the prior probability of the spike was set to 0.5. One might wonder how robust or how volatile the results

[4] The difference between the credible intervals is possibly caused by the difference in prior distributions for effect size. Heycke et al. (2018) use a Cauchy prior whereas we use a normal prior.

are to changes in the prior probability of the spike. Figure 5 visualizes the influence of the prior. In the left panel that shows the explicit evaluation data, the different estimates for different prior probabilities are practically identical. For this analysis, the data dominate the prior. In contrast, in the right panel that shows the explicit evaluation data, the prior probability of the spike has a large impact on the results. Here, the data are less informative and the prior has more influence.

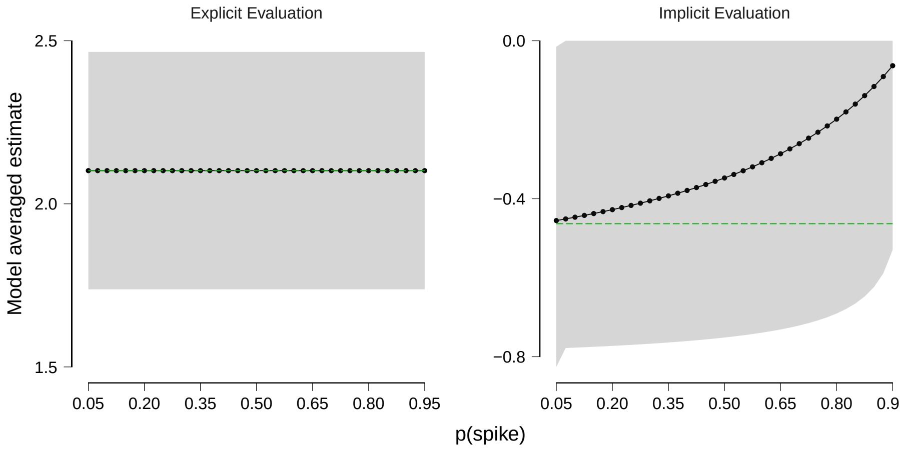

Figure 5: Robustness analysis that shows the prior probability of the spike (x-axis) versus spike-and-slab estimates (y-axis) for the explicit evaluation (left panel) and the implicit evaluation (right panel). Solid points show the point estimate of the spike-and-slab and the gray area represents the accompanying 95% credible interval. The green horizontal dashed line shows the estimate of the slab.

Discussion

Standard estimates of effect size ignore the null hypothesis and are therefore overconfident, that is, biased away from zero. The spike-and-slab model remedies this problem by explicitly considering the possibility that an effect is absent (Robinson, 2019; Rouder et al., 2018). The core idea dates back to Jeffreys (1939); nonetheless, it is ignored in empirical practice, in statistical education, and in journal guidelines.

This last sentence might be a bit too strong imo..

What if All Null Hypotheses Are False?

The spike-and-slab approach clashes with the popular estimation mindset, where it is argued that statistical significance should be abandoned in favor of estimation (Cumming, 2014; Cumming & Calin-Jageman, 2016; McShane et al., 2019; Valentine et al., 2015). One argument to forgo hypothesis testing is that all null hypotheses are false (Cohen, 1990; Meehl, 1978) and therefore there is no need to consider a component that states that an effect is exactly zero. The statistical counterargument is that, even if point null hypotheses are false, they are still mathematically convenient approximations to more complex hypotheses that allow mass on an interval close to zero (i.e., perinull hypotheses; Berger & Delampady, 1987; Kiers & Tendeiro, 2019). Thus, from a pragmatic perspective it is irrelevant whether or not null hypotheses are exactly true: in the spike-and-slab model, a narrow interval around zero will shrink estimates towards zero almost as much as the point null spike component will.

When to Ignore the Spike

There are two scenarios in which the presence of the spike can safely be ignored. First, the spike may be deeply implausible. This happens most often in problems of pure estimation, such as when determining the relative popularity of two politicians or the proportion of Japanese cars on the streets of New York. In such cases, no value or interval needs to be singled out for special attention. Second, the data, or data from prior studies, may provide overwhelming evidence that an effect is present, as in the reanalysis of the Explicit Evaluation data. When this happens, the results from a spike-and-slab model become virtually identical to those of a slab-only model, and the inclusion of the spike does not offer an added benefit albeit that the spike also does not hurt.

Conclusion

Standard methods for estimating effect size produce results that are overly optimistic. This bias toward high estimates can be corrected by applying the spike-and-slab model which explicitly accounts for the possibility that the effect is absent.

Box 1: The Spike-and-Slab Distribution as Bayesian Model Averaging

The spike-and-slab distribution can be viewed as a single model that consists of two components: the slab, which assumes that the effect is present, and the spike, which assumes the effect is absent. However, the spike-and-slab distribution can also be seen as a form of Bayesian model averaging. From that perspective, the spike and the slab are two individual models. The slab represents the unconstrained model that freely estimates effect size, and the spike represents the constrained model where the effect size is fixed to zero. Next, the results for each model are weighted by the posterior model probabilities and averaged, so that inference can be made using results from both models simultaneously. Such averaging over models yields optimal predictive performance (Zellner and Vandaele, 1975, p. 640–641, as described in Zellner and Siow, 1980, p. 600–601; Haldane, 1932, p. 57; Iverson et al., 2010; Rouder et al., 2018), and conceptually similar ideas date back much further (Wrinch and Jeffreys, 1921, p. 387; Jevons, 1874/1913). Note that these two perspectives —a two-component model or averaging of two models— differ in semantics but are mathematically equivalent.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Akademiai Kiado.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information–theoretic approach (2nd ed.)* Springer Verlag.

Clyde, M., Desimone, H., & Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, *91* (435), 1197–1208.

Cohen, J. (1990). Things I have learned (thus far). *American Psychologist*, *45*, 1304–1312.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, *69*, 54–61.

Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *28*, 55–61.

Harrington, D., D'Agostino Sr, R. B., Gatsonis, C., Hogan, J. W., Hunter, D. J., Normand, S.-L. T., Drazen, J. M., & Hamel, M. B. (2019). New guidelines for statistical reporting in the journal.

Heycke, T., Gehrmann, S., Haaf, J. M., & Stahl, C. (2018). Of two minds or one? a registered replication of rydell et al.(2006). *Cognition and Emotion*, *32* (8), 1708–1727.

Iverson, G. J., Wagenmakers, E.-J., & Lee, M. D. (2010). A model averaging approach to replication: The case of p_{rep}. *Psychological Methods*, *15*, 172–181.

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford University Press.

Jevons, W. S. (1874/1913). *The principles of science: A treatise on logic and scientific method*. MacMillan.

Kiers, H., & Tendeiro, J. (2019). With Bayesian estimation one can get all that Bayes factors offer, and more. *manuscript submitted for publication.* psyarxiv.com/zbpmy

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Atatistician*, *73*(sup1), 235–245.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

Robinson, G. K. (2019). What properties might statistical inferences reasonably be expected to have?—crisis and resolution in statistical inference. *The American Statistician*, *73*, 243–252.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*, 102–113.

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science*, *17*(11), 954–958.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108.

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life after nhst: How to describe your data without "p-ing" everywhere. *Basic and Applied Social Psychology*, *37*(5), 260–273.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 585–603). University Press.

Zellner, A., & Vandaele, W. (1975). Bayes–Stein estimators for k–means, regression and simultaneous equation models. In S. E. Fienberg & A. Zellner (Eds.), *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage* (pp. 627–653). North-Holland Publishing Company.