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. Yet, this approach is implicitly based on the assumption that there is an effect 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. We illustrate the implications of this approach and provide an empirical example.

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

*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.

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 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 $= [0.02, 0.58]$".

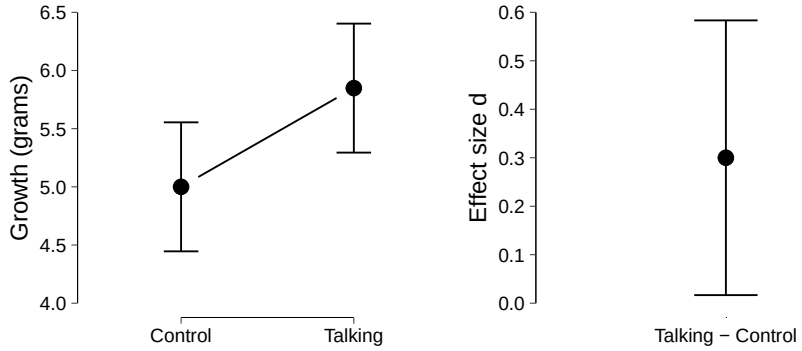

Figure 1: Standard estimation results for the fictitious plant growth example. Left panel: a descriptives plot with the mean and 95% confidence interval of plant growth in the two conditions. Right panel: point estimate and 95% confidence interval for Cohen's d.

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 is rhetorical but serves to underscore the potential conflict between standard reporting guidelines and common sense. The example raises the question: When are effect sizes overestimated? Standard point esti-

[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).

mates 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 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 overly optimistic and overly confident: the fact that $\mathcal{H}_0$ provides an acceptable account of the data should shrink effect size estimates towards zero. The statistical benefits of shrinkage are described in Efron and Morris (1977; see also Davis-Stober et al., 2018; Rouder and Lu, 2005; Shiffrin et al., 2008); the benefits of shrinking estimates towards zero are discussed for instance in George and McCulloch (1993) and Iverson et al. (2010), and van Erp et al. (2019).

In this paper we discuss a statistical approach to ameliorate the overestimation of effect size: the so-called "spike-and-slab" model. 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 tempers the estimated effect size. Third, we visualize how the spike-and-slab model may shrink the estimated effect size toward zero in general. 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 to use the spike-and-slab model.

A Spike-and-Slab Perspective

The spike-and-slab approach has been widely discussed in the statistical literature (e.g., Clyde et al., 1996; Geweke, 1996; Ishwaran, Rao, et al., 2005; Mitchell & Beauchamp, 1988; O'Hara, Sillanpää, et al., 2009) and in the psychological literature (e.g., Bainter et al., 2020; Iverson et al., 2010; Rouder et al., 2018; Yu et al., 2018). Conceptually, the approach is relatively straightforward.

As usual, the statistical goal is to infer the population effect size from a set of sample observations. Let δ denote the population effect size, let $\hat{\delta}$ denote a point estimate, and let $\hat{\delta} \mid \mathcal{H}_1$ denote a point estimate assuming the alternative hypothesis, $\mathcal{H}_1$. Assuming the null hypothesis $\mathcal{H}_0$ leads to $\hat{\delta} \mid \mathcal{H}_0$, which usually equals 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 under the alternative hypothesis that the effect exists". To the best of our knowledge, all existing guidelines for reporting effect size estimates recommend that researchers provide $\hat{\delta} \mid \mathcal{H}_1$; implicitly, the guidelines suggest to ignore $\mathcal{H}_0$, resulting in the notion that the population effect size is nonzero. In contrast, in the spike-and-slab model, the estimate of effect size is determined by both $\mathcal{H}_1$ and $\mathcal{H}_0$.

As the name suggests, the spike-and-slab model consists of two components. 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 commonly deemed *a priori* equally likely, such that the prior probability for each component is $1/2$. One can assign prior probabilities other than $1/2$, if this is motivated by prior research, prior data, or existing theories (e.g., Wilson & Wixted, 2018). After observing the data, the prior probabilities of both components, $\Pr(\text{spike})$ and $\Pr(\text{slab})$, are updated to posterior probabilities, $\Pr(\text{spike} \mid \text{data})$ and $\Pr(\text{slab} \mid \text{data})$.

By applying the spike-and-slab model we learn about the relative plausibility of the two components; in addition, the spike-and-slab model produces a *marginal* estimate of effect size – a weighted combination of effect sizes from the spike and from the slab (for mathematical detail see the online Appendix). In other words, the spike-and-slab model yields an overall effect size averaged across the spike and the slab, with averaging weights determined by the respec-

tive posterior probabilities:

$$\hat{\delta} = \left(\hat{\delta} \mid \text{spike}\right) \Pr\left(\text{spike} \mid \text{data}\right) + \left(\hat{\delta} \mid \text{slab}\right) \Pr\left(\text{slab} \mid \text{data}\right). \tag{1}$$

Marginalizing across model components according to their posterior plausibility is a uniquely Bayesian operation, and this is the statistical framework we adopt in this paper (for an accessible introduction to Bayesian inference see Vandekerckhove et al., 2018). Researchers who prefer a frequentist approach can accomplish shrinkage by using penalized maximum likelihood methods such as LASSO and ridge regression (Tibshirani et al., 2005). Another option open to frequentists is to marginalize across the spike and the slab for instance by using the Akaike Information Criterion (AIC; Akaike, 1973) and defining the averaging weights as follows. Let $\Delta\text{AIC} = (\text{AIC} \mid \text{spike}) - (\text{AIC} \mid \text{slab})$, the difference in AIC between the spike and the slab. Next we use the "Akaike weight" w_{spike} as a substitute for the posterior probability of the spike: $w_{\text{spike}} = \exp\left(-1/2\,\Delta\text{AIC}\right) / \left(1 + \exp\left(-1/2\,\Delta\text{AIC}\right)\right)$ (Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004). The substitute for the posterior probability of the slab is simply: $w_{\text{slab}} = 1 - w_{\text{spike}}$.

Note that when the spike is located at $\delta = 0$, as is usually the case, then $\left(\hat{\delta} \mid \text{spike}\right) \Pr\left(\text{spike} \mid \text{data}\right) = 0$, and consequently Equation 1 simplifies to

$$\hat{\delta} = \left(\hat{\delta} \mid \text{slab}\right) \Pr\left(\text{slab} \mid \text{data}\right). \tag{2}$$

This equation shows that the spike-and-slab estimate $\hat{\delta}$ equals the estimate that is generally recommended in reporting guidelines, $\left(\hat{\delta} \mid \text{slab}\right)$, but reduced by the posterior probability for $\mathcal{H}_1$. This shrinkage towards zero becomes negligible when the posterior probability for $\mathcal{H}_1$ approaches 1.

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/. Remember that the frequentist point estimate for the

effect size conditional on $\mathcal{H}_1$, or the slab, was $\hat{\delta} = 0.30$, with a confidence interval of 95% CI: [0.02, 0.58]. The Bayesian equivalent is $\hat{\delta} = 0.29$, with a credible interval of 95% CRI: [0.02, 0.57]). Figure 2 contrasts this Bayesian slab-only estimate against the spike-and-slab estimate.

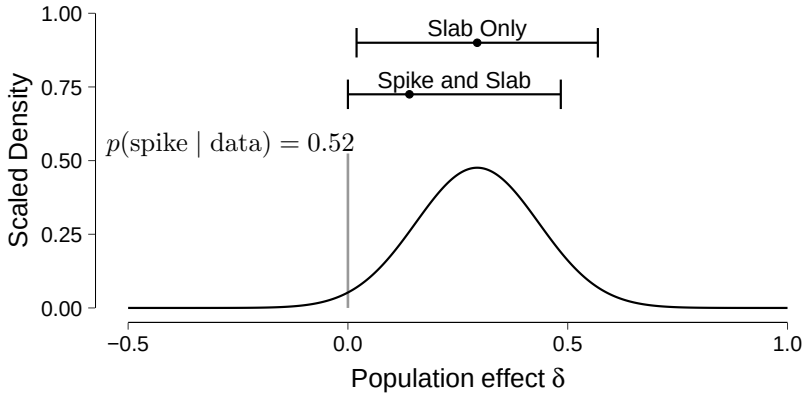

Figure 2: The spike-and-slab model. The black line represents the posterior distribution of effect size given the slab (i.e., the effect is non-zero). The posterior is scaled so that its mode ($\hat{\delta} = 0.29$) equals the posterior probability of the alternative model (i.e., $p(\text{slab} \mid \text{data}) = 0.48$). The grey line represents the posterior probability of the spike (i.e., $\hat{\delta} = 0$: 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. Here, the posterior probability of the spike after seeing the data, 0.52, is almost identical to its prior probability. In the figure, the plausibility that the effect is absent is represented 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 if the posterior probability of the spike was reduced, the spike-and-slab results would approach those of the slab-only model.

The Influence of the Spike

In the fictitious example, the spike-and-slab model reduces the estimated effect size by shrinking estimates of effect size towards zero. The 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 are the differences between a slab-only model and the spike-and-slab? In this section, we illustrate how the estimated effect size shrinks towards zero under various circumstances. We visualize the shrinkage as a function of the observed effect size, the prior on the standard deviation of effect size under the slab, the sample size, and the prior probability of the spike. We chose these parameters because the posterior distribution is fully determined by these quantities (see the online Appendix).

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 of the slab induces some shrinkage towards zero. This effect is most obvious in the top left panel, and it makes sense, as a small prior standard deviation implies there is more prior mass near the mean of the prior, which is zero. This influence of the prior standard deviation is typically referred to as *prior shrinkage*, and it 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 additional shrinkage to the prior shrinkage.

The shrinkage in the spike-and-slab model 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. As a result the marginal estimate is shrunken towards the spike's estimate, 0. In contrast, when the observed effect size is large the data are poorly described

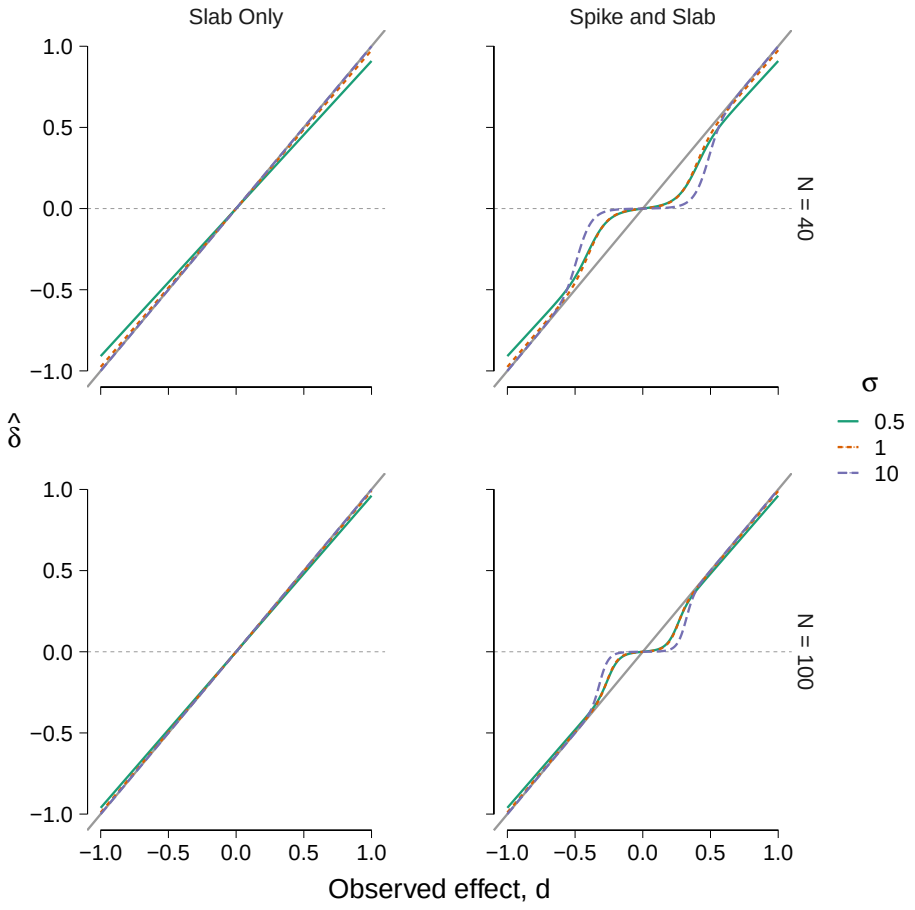

Figure 3: Observed effect size versus posterior mean for different model components and prior standard deviations. The left column shows inference based on the slab-only model while the right column shows inference based on the spike-and-slab model. In the top row, the sample size was 40 while in the 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).

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 the estimate of the slab. The plots in the right column of Figure 3 show 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.

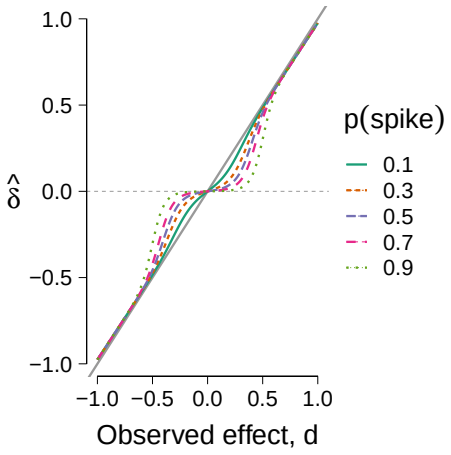

Figure 4: Observed effect (x-axis) versus the posterior mean of the spike-and-slab model (y-axis). The different lines represent different prior probabilities of the spike. The figure is based on 40 observations with a prior standard deviation of 1.

Empirical Example: Reanalysis of Two Minds

We now highlight how the spike-and-slab approach can be used in psychological practice by reanalyzing the results of Heycke et al. (2018), who conducted two registered replications of Rydell et al. (2006). We first briefly explain the design of the study before reanalyzing the Explicit Evaluation and Implicit Evaluation analyses with a spike-and-slab model. For a detailed description see the

"Procedure" section in Heycke et al. (2018). Finally, we provide a robustness analysis.

The goal of Heycke et al. (2018) was to replicate key evidence for implicit attitude formation. In the original study, Rydell et al. (2006) reported that attitudes induced by subliminal primes manifest when they are assessed by an implicit attitude measure, and attitudes induced by supraliminal cues manifest when they are assessed by an explicit attitude measure. This finding corresponds to a perhaps surprising dissociation of implicit and explicit attitude measures. In the Heycke et al. (2018) experiments participants were briefly flashed a positive or negative prime followed by an image of a person. Next, several behavioral descriptions that were either negative or positive appeared with the image of the person (e.g., "Bob cheated during a poker game"). Afterwards, participants explicitly evaluated the target person, and performed an implicit association task (IAT). In total, data of 51 participants were analyzed. Heycke et al. (2018) could not find the dissociation between explicit and implicit attitude measures. They found that while positive descriptions resulted in a more favorable explicit evaluation than negative descriptions, positive subliminal primes did not result in more favorable IAT scores than negative subliminal primes. In contrast, both explicit and implicit attitude measures were in line with the explicit descriptions they learned during the experiment.

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 $[1.41, 2.79]$.[4] 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].[5]

[4]These are the statistics reported by Heycke et al. (2018). BF stands for Bayes factor, see also the online Appendix. HDI is short for highest density interval, a type of credible interval.

[5]The difference between the point estimate and 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.

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 are to changes in the prior probability of the spike. Figure 5 visualizes the influence

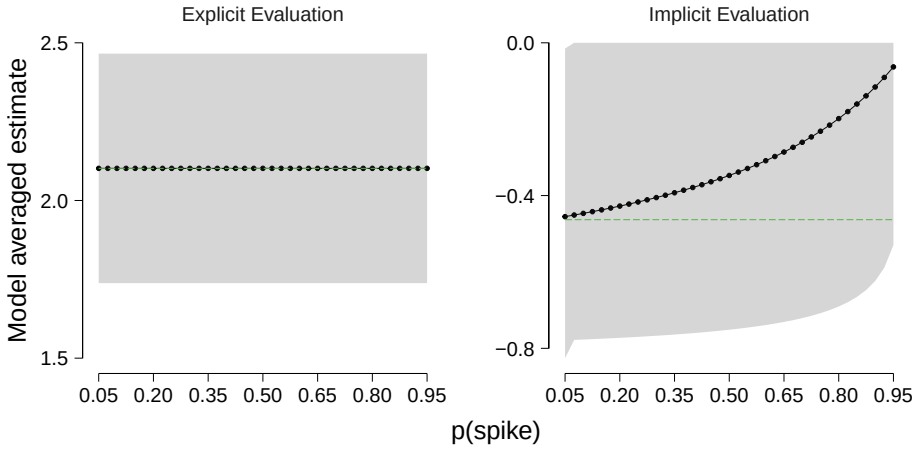

Figure 5: Robustness analysis that shows the prior probability of the spike (x-axis) versus spike-and-slab estimates (y-axis) for the explicit evaluation (left panel) and the implicit evaluation (right panel). Solid points show the point estimate of the spike-and-slab and the gray area represents the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

of the prior on the spike. 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 implicit 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. The adaptive shrinkage is a key feature of the spike-and-slab, that is, the amount of shrinkage depends on the posterior plausibility of the spike.

Discussion

Standard estimates of effect size ignore the null hypothesis and are therefore overconfident, that is, farther away from zero than they should be. The spike-and-slab model tempers the enthusiasm that the standard estimates instill by explicitly considering the possibility that an effect is absent (Robinson, 2019; Rouder et al., 2018). The core idea dates back to Jeffreys (1939; see also Jeffreys, p. 365, 1961; Ly and Wagenmakers, 2020); nonetheless, it has been largely ignored in empirical practice, in statistical education, and in journal guidelines. We believe the spike-and-slab model is a useful statistical tool to make the interpretation of effect size estimates more robust. The spike-and-slab model optimally shrinks effect sizes with ambiguous statistical support towards zero. This data-driven statistical skepticism is appropriate regardless of whether or not researchers follow good research practices, for example, preregistering study design and analysis.

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; George & McCulloch, 1993; Ly et al., 2020). 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 Can the Spike be Ignored?

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 even 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: the inclusion of the spike offers no benefit but neither does it come with a statistical cost.

Conclusion

Standard methods for estimating effect size produce results that are overly optimistic. This tendency toward high estimates can be corrected by applying the spike-and-slab model that explicitly takes into account the possibility that the effect is absent. The spike-and-slab approach is not meant as a tool to downplay other researchers' findings that one disagrees with. Instead, it provides a more robust estimate of the size of an effect of high-quality studies whenever null and alternative hypothesis are plausible. We believe that the approach allows researchers a more nuanced interpretation of their own results taking into account the plausibility that there is no effect.

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.

Bainter, S. A., McCauley, T. G., Wager, T., & Losin, E. A. R. (2020). Improving practices for selecting a subset of important predictors in psychology: An application to predicting pain. *Advances in Methods and Practices in Psychological Science*, *3*(1), 66–80.

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

Clyde, M. A., Ghosh, J., & Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, *20*, 80–101.

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.

Davis-Stober, C. P., Dana, J., & Rouder, J. N. (2018). Estimation accuracy in the psychological sciences. *PloS one*, *13*(11), e0207239.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

George, E. I., & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.

Geweke, J. (1996). Variable selection and model comparison in regression. *In Bayesian Statistics 5 (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith)*, 609–620.

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

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.

Ishwaran, H., Rao, J. S. et al. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773.

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.

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

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.

Ly, A., Stefan, A., van Doorn, J., Dablander, F., van den Bergh, D., Sarafoglou, A., Kucharskỳ, Š., Derks, K., Gronau, Q. F., Komarlu Narendra Gupta, A. R., Boehm, U., van Kesteren, E.-J., Hinne, M., Matzke, D., Marsman, M., & Wagenmakers, E.-J. (2020). The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the p-value hypothesis test. *Computational Brain & Behavior*, (3), 153–161.

Ly, A., & Wagenmakers, E.-J. (2020). Bayes factors for peri-null hypotheses. *Manuscript in preparation*.

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.

O'Hara, R. B., Sillanpää, M. J. et al. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian analysis*, *4*(1), 85–117.

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.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

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.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

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.

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (Eds.). (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192– 196.

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science*, *1*(2), 186–197.

Yu, C.-H., Prado, R., Ombao, H., & Rowe, D. (2018). A Bayesian variable selection approach yields improved detection of brain activation from complex-valued fmri. *Journal of the American Statistical Association*, *113*(524), 1395–1410.

Online Appendix: Posterior Distribution for Effect Size under the Spike-and-Slab Model

The main text featured a paired samples t-test, both for the example and for the demonstration of regularities regarding the prior probability of the spike and the prior width of the slab. In this online Appendix we detail the prior distributions for this t-test and explain how the spike-and-slab shrinkage is related to Bayes factors. More generally, we show to derive the posterior distribution for effect size δ under the spike-and-slab model. We first derive the results for the slab and spike individually and combine them afterwards.

Following Rouder et al. (2018), we assume that the observed differences between the paired samples, denoted Z_i, are normally distributed with unknown mean δ and a variance of 1. As prior distribution for δ we use a normal distribution with mean 0 and variance σ^2. This implies $Z_i \sim \mathcal{N}(\delta, 1)$ for the data and $\delta \sim \mathcal{N}\left(0, \sigma^2\right)$ for the prior. The posterior distribution for δ is obtained through Bayes' theorem:

$$\overbrace{p\left(\delta \mid \boldsymbol{Z}\right)}^{\substack{\text{Posterior}\\ \text{distribution}}} = \overbrace{p\left(\delta\right)}^{\substack{\text{Prior}\\ \text{distribution}}} \times \frac{\overbrace{L(\boldsymbol{Z} \mid \delta)}^{\text{Likelihood}}}{\underbrace{p\left(\boldsymbol{Z}\right)}_{\substack{\text{Marginal}\\ \text{Likelihood}}}}.$$

The likelihood is given by:

$$\begin{aligned} L(\boldsymbol{Z} \mid \delta, \text{slab}) &= \prod_{i=1}^{N} \mathcal{N}\left(Z_i \mid \delta, 1\right) \\ &= (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{\boldsymbol{Z}} + s^2_{\boldsymbol{Z}} + \delta^2 - 2\bar{\boldsymbol{Z}}\delta\right)\right), \end{aligned}$$

where $\bar{\boldsymbol{Z}}$ and $s^2_{\boldsymbol{Z}}$ are the sample mean and sample variance of Z_i respectively. Next, we compute the marginal likelihood by integrating out the likelihood times prior with respect to δ:

$$
\begin{aligned}
p\left(\boldsymbol{Z} \mid \text{slab}\right) &= \int_{-\infty}^{\infty} L(\boldsymbol{Z} \mid \delta)\, p\left(\boldsymbol{Z}\right) \ \mathrm{d}\delta \\
&= (2\pi)^{-\frac{N+1}{2}} \exp\left(-\frac{N}{2}\left(\bar{\boldsymbol{Z}} + s_{\boldsymbol{Z}}^2\right)\right), \\
&\quad \times \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\delta^2\left(N + \frac{1}{\sigma^2}\right) - \delta\frac{2N\bar{\boldsymbol{Z}}}{\sigma^2}\right)\right) \ \mathrm{d}\delta.
\end{aligned}
$$

Here we may recognize a Gaussian integral and use the following identity:

$$
\int_{-\infty}^{\infty} \exp\left(-ax^2 + bx + c\right) \ \mathrm{d}x = \sqrt{\frac{\pi}{a}} \exp\left(\frac{b^2}{4a} + c\right).
$$

Filling in the identity and simplifying yields:

$$
p\left(\boldsymbol{Z} \mid \text{slab}\right) = (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{\boldsymbol{Z}} + s_{\boldsymbol{Z}}^2\right)\right) \frac{\exp\left(\frac{N^2\bar{\boldsymbol{Z}}^2}{2\left(N + \frac{1}{\sigma^2}\right)}\right)}{\sqrt{N + \frac{1}{\sigma^2}}}.
$$

Next, we can obtain an expression for the posterior distribution. However, often it suffices to write out the expression for the likelihood times prior and then identify the result as a known distribution. This is particularly common in Gibbs sampling where one is interested in the conditional posterior distributions. We also do this here, as it reduces inference about the posterior distribution (e.g., what is the mean or variance) to inference about a known distribution, in this case a normal distribution:

$$
\begin{aligned}
p\left(\delta \mid \boldsymbol{Z}, \text{slab}\right) &\propto (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{\boldsymbol{Z}} + s_{\boldsymbol{Z}}^2\right)\right) \exp\left(-\frac{N}{2}\left(\delta^2 - 2\bar{\boldsymbol{Z}}\delta\right)\right) \\
&\quad \times (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\delta^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\delta^2\left(N + \frac{1}{\sigma^2}\right) - \delta\frac{2N\bar{\boldsymbol{Z}}}{\sigma^2}\right)\right).
\end{aligned}
$$

We recognize a normal distribution with variance $\sigma_1^2 = \frac{1}{N + \frac{1}{\sigma^2}}$ and mean $\mu_1 = N\bar{\boldsymbol{Z}}\sigma_1^2$. Thus we have $p\left(\delta \mid \boldsymbol{Z}\right) \propto \mathcal{N}\left(\mu_1, \sigma_1^2\right)$.

Next we compute the same for the spike. The spike states that $Z_i \sim \mathcal{N}\left(0, 1\right)$

and contains no parameters to estimate. Thus there are no prior distributions to specify and all that needs to be computed is the marginal likelihood:

$$p\left(\boldsymbol{Z} \mid \text{spike}\right) = (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{\boldsymbol{Z}} + s^2_{\boldsymbol{Z}}\right)\right).$$

Using both marginal likelihoods we can now obtain the Bayes factor in favor of the spike:

$$\text{BF}_{01} = \frac{p\left(\boldsymbol{Z} \mid \text{spike}\right)}{p\left(\boldsymbol{Z} \mid \text{slab}\right)} = \frac{\sqrt{N + \frac{1}{\sigma^2}}}{\exp\left(\frac{N^2 \bar{\boldsymbol{Z}}^2}{2\left(N + \frac{1}{\sigma^2}\right)}\right)}.$$

The posterior probability of the slab then equals:

$$\Pr\left(\text{slab} \mid \boldsymbol{Z}\right) = \frac{\Pr\left(\text{spike}\right)}{\Pr\left(\text{spike}\right) + \left(1 - \Pr\left(\text{spike}\right)\right)\text{BF}_{01}},$$

and the posterior probability of the spike is the complement. It then follows that the cumulative distribution function for the spike-and-slab posterior is given by:

$$P(\delta \leq x \mid \boldsymbol{Z}) = \begin{cases} \Pr\left(\text{slab} \mid \boldsymbol{Z}\right) \Phi(x; \mu_1, \sigma_1) & \text{if } x < 0, \\ \Pr\left(\text{spike} \mid \boldsymbol{Z}\right) + \Pr\left(\text{slab} \mid \boldsymbol{Z}\right) \Phi(x; \mu_1, \sigma_1) & \text{if } x \geq 0, \end{cases}$$

where $\Phi(x; \mu_1, \sigma_1)$ is the cumulative normal distribution. Due to the discontinuity at $x = 0$ there is no useful closed form expression for the posterior density. Nevertheless, the posterior mean of the spike-and-slab model is available in closed form. Using the law of total probability, we have:

$$p(\delta \mid \boldsymbol{Z}) = \Pr\left(\text{spike} \mid \boldsymbol{Z}\right) p(\delta \mid \text{spike}, \boldsymbol{Z}) + \Pr\left(\text{slab} \mid \boldsymbol{Z}\right) p(\delta \mid \text{slab}, \boldsymbol{Z}).$$

Computing the mean of left hand side yields:

$$\begin{aligned}\int_{-\infty}^{\infty} \delta\, p(\delta \mid \boldsymbol{Z})\ \mathrm{d}\delta &= \Pr\left(\text{spike} \mid \boldsymbol{Z}\right) \int_{-\infty}^{\infty} \delta\, p(\delta \mid \text{spike}, \boldsymbol{Z})\ \mathrm{d}\delta, \\ &\quad + \Pr\left(\text{slab} \mid \boldsymbol{Z}\right) \int_{-\infty}^{\infty} \delta\, p(\delta \mid \text{slab}, \boldsymbol{Z})\ \mathrm{d}\delta, \\ &= 0 + \Pr\left(\text{slab} \mid \boldsymbol{Z}\right) \left(\mu_\delta \mid \text{slab}, \boldsymbol{Z}\right).\end{aligned}$$

Here $(\mu_\delta \mid \text{slab}, \boldsymbol{Z})$ is the posterior mean of effect size under the slab. In a similar fashion, other statistics may be obtained. However, it is also possible to draw samples from marginal posterior distribution. To obtain a sample s, first draw u from a uniform distribution on $[0,\ 1]$. If $u < \Pr\left(\text{slab} \mid \boldsymbol{Z}\right)$ draw s from $p(\delta \mid \text{slab}, \boldsymbol{Z})$, otherwise s is zero. This approach is often used when the integrals become too unwieldy to compute analytically. For example, the R package BAS uses this procedure to compute credible intervals (Clyde et al., 2011).