A Cautionary Note on Estimating Effect Size

van den Bergh, D.; Haaf, J.M.; Ly, A.; Rouder, J.N.; Wagenmakers, E.-J.

DOI
10.31234/osf.io/h6pr8
10.1177/2515245921992035

Publication date
2021

Document Version
Final published version

Published in
Advances in Methods and Practices in Psychological Science

License
CC BY-NC

Citation for published version (APA):
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 \( p < .05 \), 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 with a point estimate of effect size and a 95% confidence interval (CI; but see Morey et al., 2016). You also manage to convince your colleague to plot the data (see Fig. 1). Mindful of the reporting guidelines of the Psychonomic Society and Psychological Science, your colleague reports the result as follows: “Cohen’s \( d = 0.30 \), 95% CI = [0.02, 0.58].”

Given 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.” Suddenly, a population effect size of zero 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 following 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, \( H_0 \)). This is not problematic when \( H_0 \) is deeply implausible, either because \( H_0 \) was highly unlikely a priori or because the data decisively undercut \( H_0 \). But when the data fail to undercut \( H_0 \) or when \( H_0 \) is highly likely a priori (i.e., “plants do not grow faster when you talk to them”), then \( H_0 \) is not ruled out as a plausible account of the data. Effect size estimates that ignore a plausible \( H_0 \) are generally overly optimistic and overly confident: The fact that \( H_0 \) provides an acceptable account of the data should shrink effect size estimates toward zero. The statistical benefits of shrinkage are described in Efron and Morris (1977; see also Davis-Stober et al., 2018; Rouder & Lu, 2005; Shiffrin et al., 2008); the benefits of shrinking estimates toward zero are discussed, for instance, in George and McCulloch (1993), Iverson et al. (2010), and van Erp et al. (2019).

The above point estimate, 0.30, may seem purely data-driven, but it is based on a model that assumes an effect
size different from zero. In this article, we propose an alternative model to estimate 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, 2005; Mitchell & Beauchamp, 1988; O’Hara & Sillanpää, 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 \( \delta \) denote the population effect size, let \( \hat{\delta} \) denote a point estimate, and let \( \hat{\delta}|H_1 \) denote a point estimate assuming the alternative hypothesis, \( H_1 \). Assuming the null hypothesis \( H_0 \) leads to \( \hat{\delta}|H_0 \), this usually equals 0. Key is that both estimates, \( \hat{\delta}|H_1 \) and \( \hat{\delta}|H_0 \), are conditional on the hypotheses. For example, \( \hat{\delta}|H_1 \) 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 \( \hat{\delta}|H_1 \); implicitly, the guidelines suggest to ignore \( H_0 \), 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 \( H_1 \) and \( H_0 \).

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 \( H_0 \) and \( H_1 \) discussed above. Both components are commonly deemed a priori equally likely such that the prior probability for each component is one half. One can assign prior probabilities other than one half if this is motivated by prior research, prior data, or existing theories (e.g., Wilson & Wixted, 2018). After observing the data, the prior probabilities (Pr) of both components, Pr (spike) and Pr (slab), are updated to posterior probabilities, Pr (spike | data) and Pr (slab | 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 Appendix in the Supplemental Material available online). 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 respective posterior probabilities:

\[
\hat{\delta} = (\hat{\delta} | \text{spike}) \Pr(\text{spike} | \text{data}) + (\hat{\delta} | \text{slab}) \Pr(\text{slab} | \text{data}).
\]  

Marginalizing across model components according to their posterior plausibility is a uniquely Bayesian operation, and this is the statistical framework we adopt in this article (for an accessible introduction to Bayesian
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 toward zero. The result may not be surprising given that 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 model? In this section, we illustrate how the estimated effect size shrinks toward 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 Appendix in the Supplemental Material).

Figure 3 shows the relation between the observed effect size and the estimated effect size for the slab and for the spike-and-slab models for 40 observations and 100 observations. All plots show that a smaller prior standard deviation of the slab induces some shrinkage toward zero. This effect is most obvious in the top left panel, and it makes sense because 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 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 toward 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 toward the spike’s estimate, zero. 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 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 shrunk toward zero, 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 shrunk toward zero. If the prior probability is small, then the spike was a priori implausible, and less evidence is needed to make its influence negligible.

**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
Explicit evaluation

In the analysis of the explicit evaluations, Heycke et al. (2018, p. 10) 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: \(t(27) = 11.52, \ p < .001; \ BF_{10} = 1.37 \times 10^7, \ d = 2.09, \ 95\% \ higher\ density\ interval \ (HDI) = [1.41, 2.79].\) The magnitude of the effect is large, and thus a spike-and-slab reanalysis yields practically the same results: \(\hat{\delta} = 2.10, \ 95\% \ CI = [1.74, 2.47].\)

Implicit evaluation

In the analysis of the IAT, Heycke et al. (2018, p. 10) 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: \(t(27) = -2.54, \ p = .017, \ BF_{10} = 2.92, \ d = -0.44, \ 95\% \ HDI = [-0.83, -0.06].\) Here, the magnitude of the effect is smaller, and as a consequence, the results from the spike-and-slab reanalysis are more conservative: \(\hat{\delta} = -0.35, \ 95\% \ CI = [-0.75, 0.00].\) The estimate of effect size is shrunked toward zero because the spike provides a reasonable account of the data, \(Pr(\text{spike} | \text{data}) = 0.25.\)

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 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 model, that is, the amount of shrinkage depends on the posterior plausibility of the spike. Note that in the right panel, the 95% CI becomes asymmetric as the prior, and therefore also the posterior probability of the spike, increases. It may appear that the CI is bounded by zero; however, this is a property of this particular data set. Had the observations been closer to zero, then the CI would have also contained negative values (e.g., the posterior mass in Fig. 2 is not zero for negative values of effect size).

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 instil.
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, 1961, p. 365; Ly & Wagenmakers, 2020); nonetheless, it has been largely ignored in empirical practice, statistical education, and 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 toward zero. This data-driven statistical skepticism is appropriate regardless of whether 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 mind-set, in which 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 null hypotheses are exactly true: In the spike-and-slab model, a narrow interval around zero will shrink estimates toward 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
A Cautionary Note on Estimating Effect Size

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.

Transparency

Action Editor: Brent Donnellan
Editor: Daniel J. Simons
Author Contributions

D. van den Bergh and J. M. Haaf drafted an initial version of the manuscript. D. van den Bergh conducted the reanalysis. J. N. Rouder shared and helped with R code for the spike-and-slab model. All authors reviewed and revised the manuscript jointly and approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by a Research Talent grant from the Netherlands Organisation of Scientific Research (NWO) to D. van den Bergh, a NWO VICI Grant 170.083 to E.-J. Wagenmakers, and an Advanced European Research Council Grant 743086 UNIFY to E.-J. Wagenmakers.

Open Practices

Open Data: not applicable
Open Materials: https://osf.io/79baq
Preregistration: not applicable
All materials have been made publicly available via OSF and can be accessed at https://osf.io/79baq. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Don van den Bergh https://orcid.org/0000-0002-9838-7308
Julia M. Haaf https://orcid.org/0000-0001-5122-706X
Jeffrey N. Rouder https://orcid.org/0000-0003-2023-3891
Eric-Jan Wagenmakers https://orcid.org/0000-0003-1596-1094

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245921992035

Notes

1. https://www.springer.com/psychology?SGWID=0-10126-6-1390050-0
3. This example is inspired by Berger and Delampady (1987).
4. These are the statistics reported by Heycke et al. (2018). BF stands for Bayes’s factor (see also the Appendix in the Supplemental Material). HDI is short for highest density interval, a type of credible interval.
5. The difference between the point estimate and the CIs is possibly caused by the difference in prior distributions for effect size. Heycke et al. (2018) used a Cauchy prior, whereas we used a normal prior.

References

van den Bergh et al.


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.


