# 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., Harrington et al., 2019; Cumming, 2014) 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, Hoekstra, Rouder, Lee, & Wagenmakers, 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 $= [0.02, 0.58]$".

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.[4]

*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.

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

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

[3] Specifically, imagine your colleague selected 100 plants and weighted them three times: at the start of the experiment, after one week, and after two weeks. The first week 50 plants were randomly selected and spoken to, while the others served as controls. The next week the roles were reversed: the previously spoken to plants served as controls while the control plants were now spoken to. The quantity of interest is the difference in weight between the two conditions. This example is inspired by Berger and Delampady (1987).

[4] Unless your colleague talked out loud, with consumption, and the plants were near.

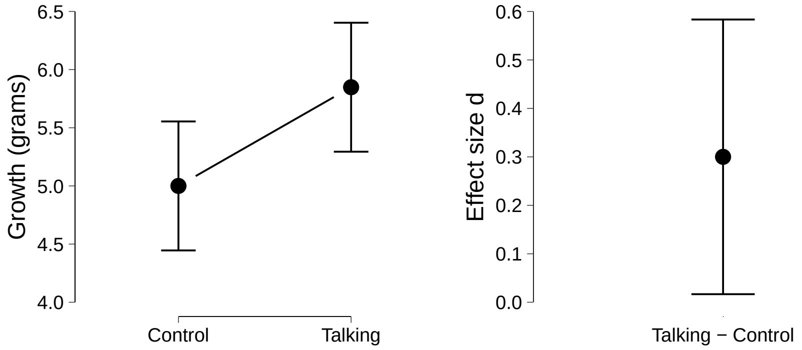

Figure 1: Standard estimation results for the fictitious plant growth example. Left panel: a descriptives plot with the mean and 95% confidence interval of plant growth in the two conditions. Right panel: point estimate and 95% confidence interval for Cohen's $d$.

## 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 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 overconfident: the fact that $\mathcal{H}_0$ provides an acceptable account of the data should shrink effect size estimates towards zero.

## A Spike-and-Slab Perspective

Here we illustrate both the overestimation and a remedy by reanalyzing the fictitious data from Figure 1.[5] We apply the spike-and-slab model (Rouder, Haaf, & Vandekerckhove, 2018; Clyde, Desimone, & Parmigiani, 1996; Mitchell & Beauchamp, 1988), which consists of two components. The first component corresponds to the position that talking to plants does not affect their growth (i.e., $\delta = 0$), whereas the second component corresponds to the position that speaking to plants does affect their growth (i.e., $\delta \neq 0$). Both components are deemed *a priori* equally likely, such that the prior probability for each component is $^1/_2$. Here we view the spike-and-slab setup as a single model, although it can also be viewed as a form of Bayesian model averaging (see Box 1 for details). 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, shrink-

[5] R code for the analysis is available at `https://osf.io/uq8st/`.

ing the estimate towards zero.[6] Figure 2 contrasts the traditional slab-only estimation against the spike-and-slab estimation. Compared to the traditional

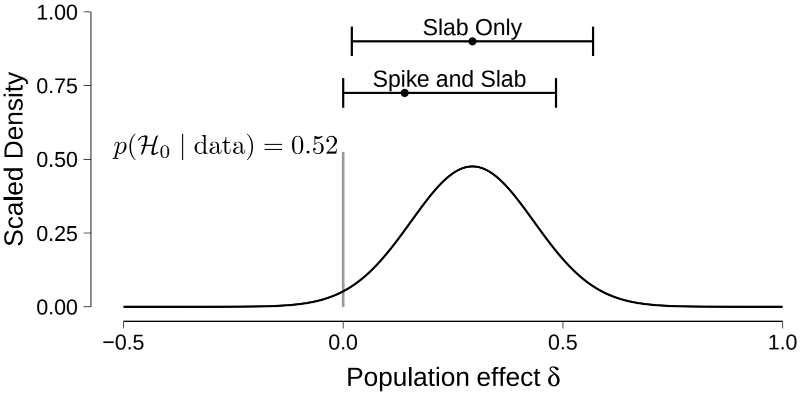

Figure 2: The spike-and-slab model. The black line represents the posterior distribution of effect size given the slab (i.e., the effect is non-zero). The posterior is scaled so that its mode ($\delta = 0.29$) equals the posterior probability of the alternative model (i.e., $p(\mathcal{H}_1 \mid \text{data}) = 0.48$). 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.

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.

# 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 both in empirical practice, in statistical education, and in journal guidelines.

### 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 (McShane, Gal, Gelman, Robert, & Tackett, 2019; Cumming &

[6] For the spike-and-slab model, the posterior distribution is constructed in the following manner: $p(\delta \mid \text{data}) = 1\{\delta = 0\}\Pr(\mathcal{M}_0 \mid \text{data}) + p(\delta \mid \text{data}, \mathcal{M}_1)\Pr(\mathcal{M}_1 \mid \text{data})$. Here, $1\{\delta = 0\}$ is the Dirac delta function which represents the spike under $\mathcal{H}_0$, Pr denotes probability of a model, and $p$ denotes density related to the magnitude of the effect.

Calin-Jageman, 2016; Valentine, Aloe, & Lau, 2015; Cumming, 2014). 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 point null hypotheses are merely mathematically convenient approximations to more complex perinull hypotheses that allow mass on an interval close to zero (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 perinull "stake" or "chimney" (Kiers & Tendeiro, 2019) component 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 (or perinull stake) 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 may provide overwhelming evidence that an effect is present. 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. A practical recommendation by Harold Jeffreys is to ignore the spike whenever sample sizes fall between 50 and 2000 and the maximum likelihood estimate deviates from the spike by more than three standard errors (Jeffreys, 1939, pp. 193–194; Jeffreys, 1980, p. 75).

## 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 & Vandaele, 1975, p. 640–641, as described in Zellner & Siow, 1980, p. 600–601; Haldane, 1932, p. 57; Iverson, Wagenmakers, & Lee, 2010; Rouder et al., 2018), and conceptually similar ideas date back much further (Wrinch & 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

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

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., ... Hamel, M. B. (2019). *New guidelines for statistical reporting in the journal.* Mass Medical Soc.

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, UK: Oxford University Press.

Jeffreys, H. (1980). Some general points in probability theory. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (pp. 451–453). Amsterdam, The Netherlands: North-Holland Publishing Company.

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

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

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). Valencia: 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). Amsterdam, The Netherlands: North-Holland Publishing Company.