frontiers in Psychology

OPINION
published: 02 July 2018
doi: 10.3389/fpsyg.2018.01125

Why You Should Report Bayes Factors in Your Transcranial Brain Stimulation Studies

Anna Lena Biel and Elisabeth V. C. Friedrich*

Biological Psychology, Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

Keywords: Bayesian statistics, null results, reproducibility, tACS, TBS, tDCS

In this commentary, we argue that it is essential to determine whether a non-significant sample effect really indicates that a particular application of transcranial brain stimulation (TBS) had no effect. We point out that non-significant results do not necessarily support a non-effect and show why reporting Bayesian statistics can help answering whether there is good enough evidence for the null hypothesis in your TBS data.

TBS aims to modulate or probe neural activity. However, reports on physiological and behavioral changes often failed to show conclusive results (Hill et al., 2016; Mancuso et al., 2016). There are many possible reasons for such inconsistencies. Recently it has been demonstrated that sufficiently large samples are essential in designing TBS experiments (Minarik et al., 2016). However, a-priori power-analyses are often skewed due to publication bias, where large or statistically significant effects get published more often. Therefore, the actual efficacy of TBS might be overestimated. While it is possible to adjust overestimated effect size for publication bias, insights about ineffective TBS duration, intensity, frequency or montage cannot be taken into account when unpublished. Thus, initiatives such as this Research Topic should encourage researchers to publish their non-significant outcomes in order to make relevant contributions to the field as well.

However, conventional significance testing cannot determine whether non-significant outcomes really indicate that a TBS protocol had no effect. In conventional significance testing, a research hypothesis assuming a certain population effect (H1), is compared against the null hypothesis assuming a non-effect in the population (H0). The probability for getting an observed sample effect is evaluated based on the significance level. If the outcome is below-threshold, one can provide evidence *against* the null hypothesis and accept the research hypothesis – whereas it is never possible to state evidence *for* the null hypothesis.

Bayes factors (BFs) are a powerful tool for evaluating evidence both for the research hypothesis and for the null hypothesis (e.g., Rouder et al., 2009; Dienes, 2011; Kruschke, 2011). In case of a conventional non-significant test, the observed sample effect either truly supports the null hypothesis or was too weak to yield evidence against it. Bayes factor tests, however, are highly useful to inform whether the data do or do not favor the null hypothesis over the alternative. We demonstrate this by simulating a series of fictional TBS experiments.

We assumed that N participants performed a task under two conditions, namely sham and real TBS. Task performance in these TBS conditions would differ by a true population effect dz. This difference in task performance was simulated by selecting N observations from a normal distribution with a mean of dz and a standard deviation of 1. We repeated this fictional experiment 1000 times. Each time, we tested for the effect of condition by comparing the research hypothesis assuming an increase of task performance during real TBS relative to sham TBS conditions (H1: $dz > 0$), against the null hypothesis assuming a non-effect (H0: $dz = 0$). First, we calculated a one-sided one-sample t-test, which is conventionally considered as significant (i.e., H0 is rejected) if p-values fall below 0.05. Next, we calculated a corresponding Bayes factor test which yields a

Edited by:
Domenica Veniero, University of Glasgow, United Kingdom

Reviewed by:
Zoltan Dienes, University of Sussex, United Kingdom

****Correspondence:***
Anna Lena Biel
anna.lena.biel@psy.lmu.de

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

Received: *22 February 2018*
Accepted: *12 June 2018*
Published: *02 July 2018*

Citation:
Biel AL and Friedrich EVC (2018) Why You Should Report Bayes Factors in Your Transcranial Brain Stimulation Studies. Front. Psychol. 9:1125. doi: 10.3389/fpsyg.2018.01125

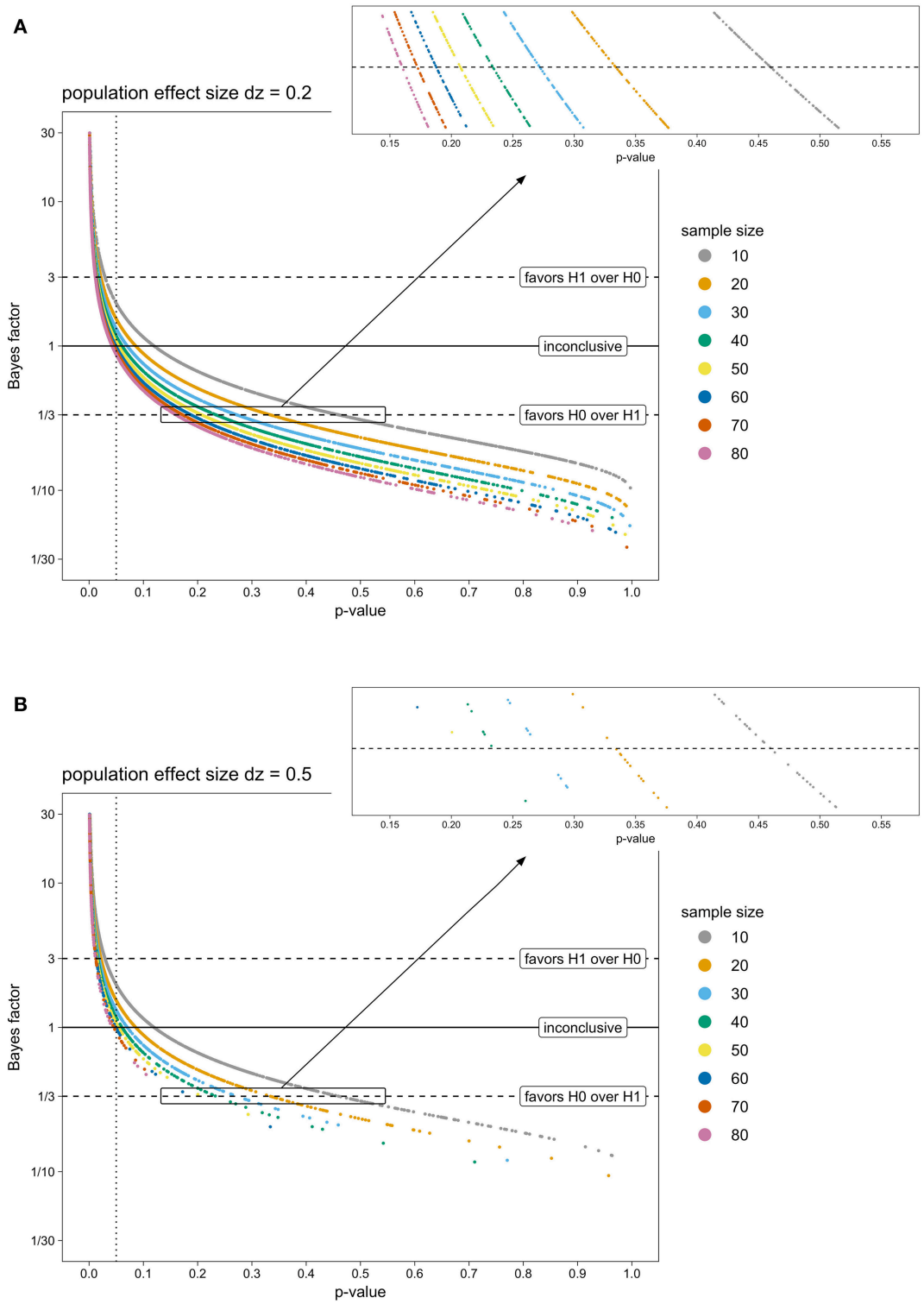

FIGURE 1 | *P*-values from a one-sided one-sample *t*-test and corresponding Bayes factors of simulated TBS experiments, for eight sample sizes (colored points) and two exemplary population effect sizes (**A**: $dz = 0.2$; **B**: $dz = 0.5$). *T*-tests with a *p*-value below 0.05 (dotted line) are conventionally considered as significant and H0 is rejected. BFs above 3 (upper dashed line) indicate evidence for H1 being more likely than H0. BFs below 0.33 (lower dashed line) yield evidence for H0 being favored over H1. BFs between 0.33 and 3 (area between the two dashed lines) are considered as inconclusive, or not more than anecdotal evidence for one of the hypotheses. *Note. BF, Bayes factor; H0, null hypothesis; H1, research hypothesis; TBS, Transcranial Brain Stimulation*.

BF quantifying how well H1 predicts the empirical data relative to H0 (BF_{10}). Here, BFs above 1 indicate evidence for H1 over H0, whereas BFs below 1 suggest the exact opposite. By convention (Jeffreys, 1961; Lee and Wagenmakers, 2014), the strength of evidence for one hypothesis compared to its competing hypothesis is regarded as noteworthy if BFs are above 3 or below 0.33. Thus, BFs between 0.33 and 3 are considered as inconclusive, or only anecdotal evidence for any hypothesis. We conducted this simulation for eight samples differing in sample size ($N = 10$, 20, 30, 40, 50, 60, 70, or 80) and six TBS protocols differing in population effect size compared to sham ($dz = 0$, 0.1, 0.2, 0.3, 0.4, or 0.5). The simulation was run using R (version 3.2.4; R Core Team, 2016) where BFs were computed using default priors by the R package BayesFactor (version 0.9.12-2; Morey and Rouder, 2015), modeling H1 as a Cauchy distribution scaled in standardized effect sizes with scale factor = 0.7 Cohen's *dz* units.

Figure 1 depicts *p*-values and Bayes factors obtained from two exemplary population effect sizes (**Figure 1A**: $dz = 0.2$, **Figure 1B**: $dz = 0.5$). Unsurprisingly, with increasing sample size, more *t*-tests were significant ($p > 0.05$) and more corresponding Bayes factors indicated at least moderate evidence for H0 over H1 ($BF < 0.33$). Similarly, fewer *t*-tests were non-significant and fewer Bayes factors favored the H0 with increasing population effect size.

Interestingly, critical *p*-values, where corresponding BFs fell below 0.33 (i.e., indicating at least moderate evidence for H0 over H1), decreased when sample size increased. For example, for samples of 10 participants, *p*-values as high as 0.45 were associated with BFs being inconclusive ($0.33 > BF > 1$). Only *p*-values beyond 0.45 were indicative for at least moderate evidence supporting H0 ($BF < 0.33$). In contrast, for samples of 80 participants, tests with *p*-values around 0.15 or more could be considered to favor H0 according to BFs. Thus, when small samples were tested, non-significant *p*-values had to be much larger for corresponding BFs to indicate at least moderate evidence for H0 than in the case of larger samples.

This relation between *p*-values, BFs and sample size stayed the same across population effect sizes: Population effect size only influenced how many *t*-tests were non-significant or BF-tests favored the H0 overall, but did not influence the range of non-significant *p*-values where corresponding BFs remained inconclusive.

In line with these described observations from simulated TBS experiments, similar associations between *p*-values and BFs have been established for other statistical tests and other models of H1 (e.g., Dienes, 2014, 2015). Taken together, they illustrate the following: First, non-significant tests with a high *p*-value do not automatically prove the null hypothesis to be true, but might indicate inconclusive evidence. Second, sample size heavily influences the threshold of critical *p*-values where Bayes factors indicate that the null hypothesis is more likely than the research hypothesis.

Thus, we conclude that any non-significant findings from conventional significance testing should be supported with evidence from Bayes Factor analyses. This is especially essential in the case of small samples. Of course, Bayesian alternatives to conventional hypothesis testing are not restricted to this case but may be advantageous in many situations. Without entering the debate whether inferential decisions should be based on a purely Bayesian approach (e.g., Dienes and Mclatchie, 2018), we argue that Bayes factor tests may be highly useful for the TBS community by distinguishing between evidence for an (un-)successful TBS protocol and inconclusive evidence. The approach of using Bayes factors to get the most out of non-significant results (Dienes, 2014) is therefore most attractive for the field: Showing the absence of a particular effect of TBS by means of Bayes factor tests may impact on the choice of stimulation parameters more positively than merely reporting conventional non-significant tests.

PRACTICAL RECOMMENDATIONS

- The absence of a particular effect of TBS compared to sham TBS can be demonstrated by reporting Bayes factors favoring H0 (there is no condition difference between sham and real TBS) over H1.
- Similarly, the specificity of an observed TBS effect can be shown by reporting Bayes factors favoring H0 (the control condition does not differ from zero) over H1.
- For non-significant *t*-tests, corresponding Bayes factors for *p*-values as high as 0.45 may indicate inconclusive evidence for either H0 or H1 when testing small samples around 10 subjects.
- For other standard statistical tests (*t*-tests, ANOVAs, regressions, etc.), there is easy-to use open-source software (JASP Team, 2018) available, providing both conventional tests as well as their Bayesian alternatives.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

Both authors were involved in the conceptualization of the topic. AB conducted the simulation, and prepared and wrote the manuscript. EF edited the manuscript.

FUNDING

This work was funded by Deutsche Forschungsgemeinschaft (DFG) SA 1872/2-1.

ACKNOWLEDGMENTS

We thank Marleen Haupt, Carola Romberg-Taylor and Paul Sauseng for helpful comments.

REFERENCES

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspect. Psychol. Sci.* 6, 274–290. doi: 10.1177/1745691611406920

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Front. Psychol.* 5:781. doi: 10.3389/fpsyg.2014.00781

Dienes, Z. (2015). "How Bayesian statistics are needed to determine whether mental states are unconscious," in *Behavioural Methods in Consciousness Research,* ed M. Overgaard (New York, NY: Oxford University Press), 199–220.

Dienes, Z., and Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychon. Bull. Rev.* 25, 207–218. doi: 10.3758/s13423-017-1266-z

Hill, A. T., Fitzgerald, P. B., and Hoy, K. E. (2016). Effects of anodal transcranial direct current stimulation on working memory: a systematic review and meta-analysis of findings from healthy and neuropsychiatric populations. *Brain Stimul.* 9, 197–208. doi: 10.1016/j.brs,.2015.10.006

JASP Team (2018). *JASP (Version 0.8.5)[Computer software]*. JASP Team.

Jeffreys, H. (1961). *Theory of Probability, 3rd Edn*. Oxford, UK: Oxford University Press.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspect. Psychol. Sci.* 6, 299–312. doi: 10.1177/1745691611406925

Lee, M. D., and Wagenmakers, E. J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press.

Mancuso, L. E., Ilieva, I. P., Hamilton, R. H., and Farah, M. J. (2016). Does transcranial direct current stimulation improve healthy working memory? A meta-analytic review. *J. Cogn. Neurosci.* 7, 1–27. doi: 10.1162/jocn_a_00956

Minarik, T., Berger, B., Althaus, L., Bader, V., Biebl, B., Brotzeller, F., et al. Sauseng, P. (2016). The importance of sample size for reproducibility of tDCS effects. *Front. Hum. Neurosci.* 10:453. doi: 10.3389/fnhum.2016.00453

Morey, R. D., and Rouder, J. N. (2015). *{BAYESFACTOR}: Computation of Bayes Factors for Common Designs*. R package version 0.9.12-2. Available online at: https://cran.r-project.org/web/packages/BayesFactor/

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available online at: http://www.R-project.org/

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16, 225–237. doi: 10.3758/PBR.16.2.225

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.