efpsa Journal of European Psychology Students

Mehler, D. M. A., Edelsbrunner, P. A., & Matić, K. (2019). Appreciating the Significance of Non-significant Findings in Psychology. *Journal of European Psychology Students, 10*(4), 1–7. Doi:10.5334/jeps.e2019a

LETTER FROM THE EDITORS

Appreciating the Significance of Non-Significant Findings in Psychology

* David M. A. Mehler[1,2], * Peter A. Edelsbrunner[3], Karla Matić[4]

Hypothesis tests for which the null hypothesis cannot be rejected ("null findings") are often seen as negative outcomes in psychology. Null findings can, however, bear important insights about the validity of theories and hypotheses. In addition, the tendency to publish mainly significant findings is considered a key reason for failures to replicate previous studies in various fields, including psychology. In this editorial, we discuss the relevance of non-significant results in psychological research and ways to render these results more informative. We discuss the possibility to test whether null findings provide evidence for the absence or negligible size of an effect, based both on frequentist and Bayesian statistical methods. We further discuss the role of adequate power analysis in obtaining informative evidence for null findings, with a special emphasis on student research. Lastly, we encourage researchers at all career stages to submit null findings for publication.

Keywords: equivalence testing; null hypothesis; Bayes factor; ROPE testing; Registered Reports

Imagine putting great care into the design of your thesis project, and eventually getting results that are not statistically significant and thus apparently do not support your main hypothesis. Such situations may be especially frustrating for students; they may question the theoretical background and foundations of their research, or even resort to restating their hypotheses in light of the non-significant results. However, this common assumption—that non-significant findings indicate flaws in a theory or undermine the value of a research project—is a misconception (Edelsbrunner & Thurn, 2018). In this editorial, we discuss the importance of publishing non-significant results. We illustrate why a non-significant finding alone does not indicate evidence for the absence of an effect and introduce statistical methods (frequentist and Bayesian) that allow to test whether null findings indicate absence or a negligible size of an effect. We further discuss the role of adequate power analysis in obtaining informative evidence for null findings, with a special emphasis on student research. We encourage researchers at all career stages to submit null findings for publication.

* Authors contributed equally. Corresponding author: David M. A. Mehler (mehlerdma@gmail.com) 1 University of Munster, Germany; 2 Cardiff University, United Kingdom; 3 ETH Zurich, Switzerland; 4 University of Leuven, Belguim

The Relevance of Non-Significant Results

Adequate reporting of non-significant and inconclusive findings improves the reliability of the scientific literature. Researchers who withhold their non-significant findings (called *file-drawer effect*) or journals that refuse to publish statistically non-significant findings (so-called *publication bias*) create—however inadvertently—a literature that is a distorted version of the scientific reality. One can think about the reported effects as only "the tip of the iceberg", with many of the reported effects likely being overestimates of the real effects (Algermissen & Mehler, 2018; Schäfer & Schwarz, 2019). This can result in difficulties to replicate previous work, because our estimates of the sample sizes needed to reliably find an effect are based on biased prior literature (Algermissen & Mehler, 2018; Open Science Collaboration, 2015). The adequate reporting and publishing of null findings is also informative because it can enable researchers to refute a theory, for instance if they repeatedly provide *evidence for the absence* of an effect (Fidler, Singleton Thorn, Barnett, Kambouris, & Kruger, 2018). Therefore, adequate reporting of non-significant findings renders scientific literature as a whole more complete, and allows for a better judgment about the replicability of scientific work.

What Does a Non-Significant Finding Mean?

Researchers talk about "null findings" when their statistical tests do not reach significance, and oftentimes they interpret such non-significant tests as conclusive evidence for the absence of the effect in question (Hoekstra, Finch, Kiers, & Johnson, 2006). However, this interpretation is misleading, because a non-significant effect can occur for at least two other reasons. First, the effect might exist with about the predicted size, but it could merely have been overlooked because the evidence in the given sample is not sufficiently strong. Second, the effect could be smaller than expected, perhaps even close to zero, and might thus be considered negligible or absent.

How often do we overlook an effect although it really exists? We can estimate this by computing statistical power—the probability to obtain a significant result for an effect of a certain size that really exists, given a specific statistical model and sample size. An example of statistical power for a commonly used statistical test, and how it relates to effect sizes, is depicted in Figure 1.

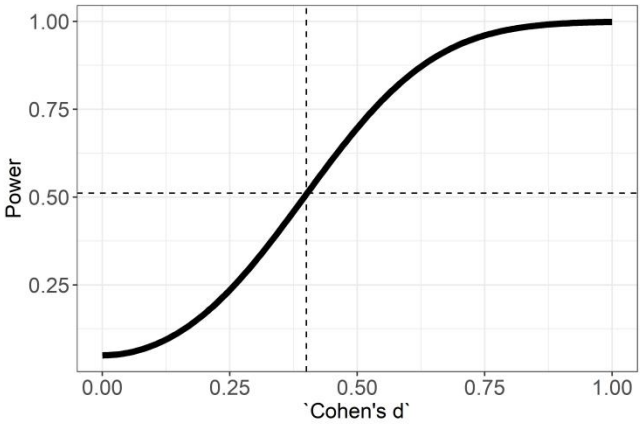

Figure 1. Power of an independent samples t-test with n = 50 per group, for different effect sizes indicating real difference between groups in Cohen's d. Power at Cohen's d = 0.4 indicated by dashed lines. Code available at https://osf.io/d5zyt

Specifically, we can see that, if we use a t-test to compare group means between two groups with a fixed size of Cohen's *d* = .4 (which is considered a typical effect size in psychology; Kühberger, Fritz, & Scherndl, 2014; see also Szucs & Ioannidis, 2017 and Schäfer & Schwarz, 2019 for more conservative estimates), the statistical power is just about .5. In such situation, we would only find a significant effect in about 50% of cases. In other words, with such low power, data collection becomes as effective as a mere coin toss. If the real effect is even smaller

(Figure 1 further left on x-axis), this percentage drops substantially and the effects will be easily missed. In contrast, if the effect is larger, it becomes easier to detect it reliably.

It is usually suggested to estimate the statistical power for a specific study design and statistical test before the study is conducted, and adapt the sample size to achieve at least 80–90% statistical power. However, it has been shown that many studies in psychology do not achieve sufficient statistical power (Szucs & Ioannidis, 2017). This tells us that a non-significant p-value does not indicate that an effect is absent; it could have just been overlooked.

Frequentist Equivalence Testing

So how can we establish whether we overlooked a true effect, or if a non-significant result indicates rather that the effect is really absent or of negligible size? We can start by defining, before we look at the data, which size the specific effect would have to exceed in order to count as *meaningful*. This decision should be based on careful consideration of prior research, theory, and research question: at which minimal size is the effect large enough to have meaningful consequences from a theoretical or practical perspective? We then accept that effects below this boundary can be considered negligible. Such an effect size is called SESOI—the *Smallest Effect Size of Interest*.

We can use equivalence testing to determine if we have sufficient evidence to consider the effect in question negligible (Lakens, Scheel, & Isager, 2018), as depicted in Figure 2. We estimate a confidence interval around the effect size of interest. If the confidence interval lies fully within the area of effect sizes below our SESOI, as it does in Figure 2, we have a significant result in favor of equivalence—the effect in question is negligible. More specifically, the null hypothesis and alternative hypothesis are

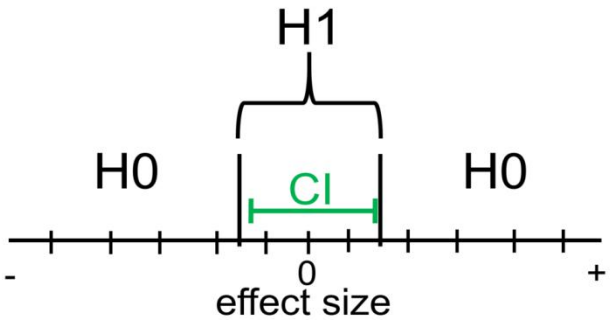

Figure 2. Illustration of a significant result of an equivalence test. Confidence interval lies fully within the equivalence bounds defined by H0 that the effect has a meaningful size, supporting the decision to reject the hypothesis and instead accepting the hypothesis H1 that the effect is negligible.

different when working with a SESOI compared to traditional significance tests: the null hypothesis in an equivalence test states that the effect is either below or above the threshold of negligibility, and the alternative hypothesis states that the effect is negligible. Hence, a significant result of the equivalence test supports the decision to reject the null hypothesis that the effect is larger than the SESOI. Instead, we may accept the hypothesis that the effect is smaller than the SESOI, and therefore negligible. However, if the confidence interval does not lie fully within the equivalence bounds, the equivalence test result remains inconclusive, suggesting that more data would be required to yield a conclusive result.

Overall, equivalence testing re-emphasizes the importance of statistical power and thus the study design, and foremost sample size planning. Power analyses are ideally informed by pre-defined SESOIs to increase both the chance to detect small effects and to show equivalence in case of a non-significant finding. We recommend interested readers to consult published tutorials by Lakens and colleagues (2018a) and Lakens and colleagues (2018b).

Bayesian Approaches: Credible Intervals and Bayes Factors

There are two common methods to evaluation of null findings based on Bayesian statistics, an approach that has recently gained traction (Wagenmakers et al., 2018).

The first is analogous to frequentist equivalence testing. Again, equivalence bounds (which some Bayesian statisticians call ROPE—"region of practical equivalence", see Kruschke, 2011) are determined. However, instead of examining whether the confidence interval of an effect lies within, above, or below the equivalence bounds, here the Bayesian equivalent called the credible interval (Kruschke, 2011) is estimated and examined. Given certain conditions, the credible interval often covers a similar area as a frequentist confidence interval, but it has a different interpretation (Kruschke, 2011; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). For example, if the Bayesian 90% credible interval lies fully within the SESOI region, we can be 90% sure that the parameter representing the effect we are interested in actually lies within the equivalence region. In contrast, if the frequentist confidence interval lies fully within the equivalence region, it supports the decision to reject the null hypothesis of non-equivalence and accept the alternative hypothesis of a negligible effect (Lakens et al., 2018b).

The second Bayesian approach is hypothesis testing based on Bayes factors. The Bayes factor is often seen as the Bayesian counterpart to frequentist hypothesis testing via p-values. In light of collected data, it informs us how well different hypotheses can predict the data compared to each other. Taking into account prior odds, this can be translated into the extent to which the null hypothesis (e.g., the effect of interest does not differ from zero) is more or less likely than an alternative hypothesis (e.g., that there actually is a relevant effect; Kruschke, 2011). To get a grasp on Bayesian statistics, we suggest to consult Edelsbrunner (2017) for a reader-friendly blog post; Etz, Gronau, Dablander, Edelsbrunner, and Baribault (2018) for an overview of introductory literature; Wagenmakers and colleagues (2018a and 2018b) for a theoretical primer and basic practical introduction; and Kruschke (2011), Lakens and colleagues (2018a) for tutorials and discussions of the approaches presented here and more in-depth discussions.

Conclusion: Making the Most of Non-Significant Findings

The issues outlined here show that a well-conducted power analysis is key to a good study design. However, students might lack the prior knowledge required for conducting an appropriate power analysis that allows a specific statistical test to detect (or reject) a smallest effect size of interest with a sufficiently large probability. Moreover, student projects are often constrained by limited time and monetary resources that make it difficult to collect sufficiently large data sets. We have four suggestions how to make student research under these conditions as informative as possible.

1) An adequate power analysis and subsequent adjustment of the sample size is required to make a specific statistical test informative. We therefore recommend students to inform themselves about power analysis. Helpful resources include Harms and Lakens (2018) who provide a conceptual explanation, Greenland and colleagues (2016) offering a guide to misconceptions, Judd, Westfall, and Kenny (2017) who cover more complex experimental designs, and Arend and Schäfer (2019) who elaborate on situations that involve multilevel data.

2) Collaboration with others, or making use of available open data, may help mitigate issues around statistical power (Allen & Mehler, 2019). Collecting the data for student projects in small teams rather than individually can ensure adequate sample sizes and thus provide sufficient statistical power.

3) Changes in study design, for example adding information from an additional measurement point such as a pretest to a study design, can substantially increase power (Venter, Maxwell, & Bolig, 2002).

4) If the sample size suggested by a power analysis is difficult to achieve (e.g., due to limited time or monetary resources), it can still be informative to focus on precise parameter estimation, and on presenting informative descriptive statistics as well as data visualizations, rather than on hypothesis testing (Valentine, Aloe, & Lau, 2015). While this strategy does not really circumvent the issue of hypothesis testing (Morey, Rouder, Verhagen, & Wagenmakers, 2014), descriptive statistics and data visualizations can be very informative both as an addition to hypothesis tests and on their own (Valentine et al., 2015). Alternatively, Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2017) and Schönbrodt and Wagenmakers (2018) present sample planning based on Bayesian statistics that can be helpful in case of limited resources, and Allen and Mehler (2019) discuss how constraints and benefits might be weighted in the design of studies conducted by early career researchers. Whichever approach is taken, however, constraints in the sampling plan should be described transparently in the manuscript to permit a fair evaluation.

In fact, transparent documentation of planned experiments helps to ensure that non-significant findings are published and provides an opportunity for external feedback before data collection. Noteworthy, the publishing format Registered Report, which has been introduced at the Journal of European Psychology Students in 2016 (King et al., 2016), includes a Stage 1 peer-review for which researchers submit a documentation of their methodology, hypotheses, analysis plan, and power calculation for peer-review before data collection starts (for an example, see Kvetnaya, 2018). After successful submission, researchers are guaranteed that their study will be published independently of the statistical outcome, thus mitigating the publication bias (Allen & Mehler, 2019). Registered Reports offer a wide range of benefits to students, but they also require more time for planning and piloting experiments than traditional experimental formats. Therefore, in case of tight time constraints, researchers can still preregister their main hypotheses, methodology and analysis plan before data collection and make it openly accessible on a public repository such as the Open Science Framework (Crüwell et al., 2018).

Overall, we conclude that adequate reporting and follow-up of non-significant test results (using frequentist or Bayesian statistics) as well as properly conducted power analyses (that are ideally informed by pre-defined SESOIs) render traditionally disregarded "null findings" informative. The Journal of European Psychology Students is committed to support researchers in following best research practices, and therefore fully encourages authors to submit studies resulting in non-significant findings, employ follow-up analyses of non-significant results as described here, and consider submitting Registered Reports.

Acknowledgements

We would like to thank Johannes Algermissen, Sophia Crüwell, Fabian Dablander, Kelsey MacKay, and Anne Scheel for helpful feedback on an earlier draft of this manuscript.

Conflicts of Interest

The authors have no conflicts of interest to declare.

References

Algermissen, J., & Mehler, D. M. A. (2018). May the power be with you: are there highly powered studies in neuroscience, and how can we get more of them? *Journal of Neurophysiology*, *119*(6), 2114–2117. Doi:10.1152/jn.00765.2017

Allen, C. P. G., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS Biology*, *17*(5), e3000246. Doi:10 .1371/journal.pbio.3000246

Arend, M. G., & Schäfer, T. (2019). Statistical Power in Two-Level Models: A Tutorial Based on Monte Carlo Simulation. *Psychological methods*, *24(1)*, 1–19. Doi:10.1037/met0000195

Crüwell, S., Doorn, J. V., Etz, A., Makel, M. C., Niebaum, J. C., Orben, A., ... Schulte, M. (2018). 8 Easy Steps to Open Science: An Annotated Reading List. *PsyArXiv Preprints*, 1–32. Doi:10.31234/osf.io/cfzyx

Edelsbrunner, P. A. (2017). Bayesian Statistics: What Is It and Why Do We Need It [Blog post]. Retrieved from http://blog. efpsa. org/2014/11/17/bayesian-statistics-what-is-it-and-why-do-we-need-it-2

Edelsbrunner, P. A., & Thurn, C. (2018). Misinterpretations of Non-significant p-values: Estimating Their Frequency and Potential Consequences for Educational Theory and Policy. In *Proceedings of the 51st Congress of the German Society for Psychology*. Germany: Frankfurt am Main. Doi:10.17605/osf.io/g5paq

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to Become a Bayesian in Eight Easy Steps: An Annotated Reading List. *Psychonomic Bulletin & Review*, *25*(1), 219–234. Doi:10.3758/s13423-017-1317-5

Fidler, F., Singleton Thorn, F., Barnett, A., Kambouris, S., & Kruger, A. (2018). The Epistemic Importance of Establishing the Absence of an Effect. A*dvances in Methods and Practices in Psychological Science*, *1*(2), 237–244. Doi:10.1177/2515245918770407.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical Tests, p Values, Confidence Intervals, and Power: A Guide to Misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. Doi:10.1007/s10654-016-0149-3

Harms, C., & Lakens, D. (2018). Making 'Null Effects' Informative: Statistical Techniques and Inferential Frameworks. *Journal of Clinical and Translational Research*, *3*(2), 382–393. Doi:10.18053/jctres.03.2017S2.007

Hoekstra, R., Finch, S., Kiers, H. A., & Johnson, A. (2006). Probability as certainty: Dichotomous Thinking and the Misuse of p-values. *Psychonomic Bulletin & Review*, *13*(6), 1033–1037. Doi:10.3758/BF03213921

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments With More Than One Random Factor: Designs, Analytic Models, and Statistical Power. *Annual Review of Psychology*, *68*, 601–625. Doi:10.1146/annurev-psych-122414-033702

King, M., Dablander, F., Jakob, L., Agan, M., Huber, F., Haslbeck, J., & Brecht, K. (2016). Registered Reports for Student Research. *Journal of European Psychology Students*, *7*(1). Doi:doi.org/10.5334/jeps.401.

Kruschke, J. K. (2011). Bayesian Assessment of Null Values via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. Doi:10.1177/1745691611406925

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication Bias in Psychology: A Diagnosis Based on the Correlation Between Effect Size and Sample Size. *PloS One*, *9*(9), e105825. Doi:10.1371/journal.pone.0105825

Kvetnaya, T. (2018). Registered Replication Report: Testing Disruptive Effects of Irrelevant Speech on Visual-Spatial Working Memory. *Journal of European Psychology Students*, *9*, 10–15. Doi:10.5334/jeps.450

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018). Improving Inferences about Null Effects with Bayes Factors and Equivalence Tests. *The Journals of Gerontology: Series B, gby065*. Doi:10.1093/geronb/gby065

Lakens, D., Scheel, A., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. A*dvances in Methods and Practices in Psychological Science, 1*(2), 259–269. Doi:10.1177/2515245918770963

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*(1), 103–123. Doi:10.3758/s13423-015-0947-8

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why Hypothesis Tests Are Essential For Psychological Science: A Comment on Cumming. *Psychological Science*, *25*(6), 1289–1290. Doi:10.1177/0956797614525969

Open Science Collaboration. (2015). Estimating the Reproducibility of Psychological Science. *Science*, *349*(651). Doi:10.1126/science.aac4716.

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, *10*. Doi:10.3389/fpsyg.2019.00813

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes Factor Design Analysis: Planning for Compelling Evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. Doi:10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences. *Psychological Methods*, *22*(2), 322. Doi:10.1037/met0000061

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature. *PLOS Biology*, *15*(3), e2000797. Doi:10 . 1371/journal.pbio.2000797

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. Doi:10.1080/01973533.2015.1060240

Venter, A., Maxwell, S. E., & Bolig, E. (2002). Power in Randomized Group Comparisons: The Value of Adding a Single Intermediate Time Point to a Traditional Pretest-Posttest Design. *Psychological*

Methods, *7*(2), 194–209. Doi:10.1037/1082-989X.7.2.194

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... Morey, R. D. (2018). Bayesian Inference For Psychology. Part II: Example Applications With JASP. *Psychonomic Bulletin & Review*, *25*, 58–76. Doi:10.3758/s13423

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. D. (2018). Bayesian Inference For Psychology. Part I: Theoretical Advantages and Practical Ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57. Doi:10.3758/s13423-017-1343