


Reducing our dependence on null hypothesis testing: A key to enhance the reproducibility and credibility of our science



Author:

Kevin R. Murphy¹ 

Affiliation:

¹Department of Work and Employment Studies, Kemmy Business School, University of Limerick, Limerick, Ireland

Corresponding author:

Kevin Murphy,
Kevin.R.Murphy@ul.ie

Dates:

Received: 16 July 2019

Accepted: 27 Aug. 2019

Published: 05 Nov. 2019

How to cite this article:

Murphy, K.R. (2019). Reducing our dependence on null hypothesis testing: A key to enhance the reproducibility and credibility of our science. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde*, 45(0), a1717. <https://doi.org/10.4102/sajip.v45i0.1717>

Copyright:

© 2019. The Authors.
Licensee: AOSIS. This work is licensed under the Creative Commons Attribution License.

Read online:



Scan this QR code with your smart phone or mobile device to read online.

Problemification: Over-reliance on null hypothesis significance testing (NHST) is one of the most important causes of the emerging crisis over the credibility and reproducibility of our science.

Implications: Most studies in the behavioural and social sciences have low levels of statistical power. Because 'significant' results are often required, but often difficult to produce, the temptation to engage in questionable research practices that will produce these results is immense.

Purpose: Methodologists have been trying for decades to convince researchers, reviewers and editors that significance tests are neither informative nor useful. A recent set of articles published in top journals and endorsed by hundreds of scientists around the world seem to provide a fresh impetus for overturning the practice of using NHST as the primary, and sometimes sole basis for evaluating research results.

Recommendations: Authors, reviewers and journal editors are asked to change long-engrained habits and realise that 'statistically significant' says more about the design of one's study than about the importance of one's results. They are urged to embrace the ATOM principle in evaluating research results, that is, *accept* that there will always be uncertainty, and be *thoughtful*, *open* and *modest* in evaluating what the data mean.

Keywords: Significance Testing; Confidence Intervals; Questionable Research Practices; Null Hypothesis.

Introduction

There are many indications that several sciences, including psychology, have a reproducibility crisis in their hands (Ioannidis, 2005; McNutt, 2014; Pashler & Wagenmakers, 2012). Peer-reviewed research that is published in highly reputable journals has often failed to replicate; papers that attempt to replicate published research very often report smaller and non-significant effects (Open Science Collaboration, 2015). This persistent failure to replicate published findings calls the credibility and meaning of those findings and, by extension, of other published research into question. Efendic and Van Zyl (2019) provide an excellent summary of the challenges this crisis poses to Industrial and Organizational Psychology, and they outline several thoughtful responses this journal might make to increase the robustness and credibility of the research published in the *South African Journal of Industrial Psychology*. The changes they propose will not be easy to implement, in part because they require authors, reviewers and editors to change their perspectives and their behaviours. However, there are reasons to be optimistic about one of the major changes the authors propose.

Many of the problems with reproducibility can be traced to our field's long reliance on Null Hypothesis Significance Testing (NHST). Efendic and Van Zyl (2019) documented two of the most problematic aspects of the use of statistical tests of the null hypothesis in making decisions about study results. These were (1) the inadequate power of most studies and (2) the strong temptation to engage in a range of questionable research practices (ranging from *p* fishing [i.e. trying multiple statistical tests to find one in which $p < 0.05$] to outright fabrication [Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016; Neuroskeptic, 2012]) in search of a 'significant' ($p < 0.05$) result. While many of the changes proposed by Efendic and Van Zyl (2019) could help to address some of the problems caused by an over-reliance on NHST, I do not think they go far enough. As long as the pursuit of $p < 0.05$ remains central to our evaluation of research results, I do not think we will make a meaningful dent in addressing the reproducibility crisis. Fortunately, there are reasons to believe that the long reign of NHST is coming to an end.

The slow demise of null hypothesis significance testing

For decades, methodologists (e.g. Cohen, 1962, 1988, 1994; Fidler, Thomason, Cumming, Finch, & Leeman, 2004; Meehl, 1978; Murphy, Myers, & Wolach, 2014; Schmidt, 1996; Sterling, 1959) have been warning the research community about the perils of relying on NHST as a method for evaluating research results. Despite decades of criticism, null hypothesis tests are still widely used for evaluating the results of research. For example, Bakker, Van Dijk and Wicherst (2012) cited several reviews that together suggest that over 95% of papers in psychology use null hypothesis testing as one criterion for evaluating results. Authors who find that their results are not significant may decide to abandon that hypothesis or to not submit their work (self-censorship); reviewers and editors who see that the principal results of a study are not statistically significant may decide not to accept that study for publication. Several studies (e.g. Bakker et al., 2012; Ioannidis, 2005) have modelled the effects of reliance on null hypothesis testing to show how it biases the published literature and how it contributes to the reproducibility crisis in the social sciences.

There are many critiques of NHST, but in my view, two are paramount. Firstly, NHST tests a hypothesis that very few people believe to be credible – the hypothesis that treatments have *no* effect whatsoever or that variables are *completely* uncorrelated (Cohen, 1994). It is indeed plausible that treatments might have a very small effect (perhaps so small that they can safely be ignored) or that the correlations between two variables might be quite small, but the hypothesis that they are *precisely* zero is not a credible one (Murphy et al., 2014). The null hypothesis is a point hypothesis that divides the set of possible outcomes into two zones, that is, either that a specific hypothesis about the value of some statistic is true (i.e. ρ is precisely equal to zero) or the entire range of alternatives (ρ is precisely equal to any value other than zero, to the millionth decimal place) contains the truth. It is well known that the likelihood of *any* point hypothesis being precisely true is exceedingly small, and that it does not matter whether a value of zero or some other specific value for a statistic is assumed. Because a point hypothesis is infinitely precise and the range of alternatives to a point hypothesis is infinitely large, the likelihood that any point hypothesis – including the classic null that the effect is precisely zero – is true will tend to approach zero. This undermines the whole architecture of NHST. For example, if H_0 is never, or essentially never true, it is impossible to make a type I error and all of the statistical tools designed to minimise these errors (e.g. stringent alpha levels, Bonferroni corrections) become meaningless. Several alternate approaches have been developed to test the more credible hypothesis, such as the hypothesis that the effects of treatments fall within a range of values that are all trivially small (Murphy et al., 2014; Rouanet, 1996; Serlin & Lapsley, 1985, 1993), but their uptake has been limited.

Secondly, the outcomes of NHST are routinely misinterpreted. If you fail to reject H_0 , you are likely to conclude that your

treatments did not work or that the variables you are interested in are not related to one another. That is, you are very likely to interpret NHST as telling you something about your results. This is wrong (Murphy et al., 2014). The failure to reject H_0 tells you something about the design of your study, in particular, that you did not build a study with sufficient statistical power. One of the lessons learnt by scanning power tables is that given a sufficiently large sample, you can reject virtually *any* null hypothesis, no matter what treatments or variables you are studying (Cohen, 1988; Murphy et al., 2014). Conversely, if your sample is small enough, you can be virtually certain that you will not reject H_0 , regardless of the treatments or variables being studied. Null hypothesis significance testing is essentially an assessment of whether or not your study was powerful enough to detect whatever effect you are studying, and it is very little else.

Recent developments in the scientific literature have finally given some reason for optimism that our over-reliance on NHST is coming to an end. A series of recent articles (Amrhein, Greenland, & McShane, 2019; Wasserstein & Lazar, 2016; Wasserstein, Schirm, & Lazar, 2019) in high-profile journals (e.g. *Nature* and *American Statistician*) have accomplished three things that decades of research papers and chapters of previous critics of NHST had not been able to accomplish. Firstly, they have described in clear and largely non-technical language the deficiencies of NHST as a method for making decisions about the meaning of results. Secondly, they have gathered the support of hundreds of signatories (there were over 800 signatories from over 50 countries within a week of the distribution of Amrhein et al.'s 2019 draft) to statements calling for an end to mechanical reliance on significance testing for evaluating findings. Thirdly, and most importantly, they (particularly Wasserstein et al., 2019) have presented constructive alternatives. I believe this recent wave of papers presents an opportunity for researchers in virtually all disciplines to improve the methods they apply to make sense of data and to advance the cause of reproducible science.

A call for action

I therefore urge the *South African Journal of Industrial Psychology* to adopt the core principles articulated by Wasserstein et al. (2019). Firstly, it is not sufficient to simply say 'don't use significance testing'; it is critical to help researchers and readers to understand and empower them to utilise the available alternatives. For example, Serlin and Lapsley (1985, 1993) developed methods for assessing whether or not the hypothesis of no effect was a good enough approximation of reality to serve as a working description of one's result. Rather than testing whether the effect was *precisely* zero, their methods allowed one to evaluate the possibility that the effects were close enough to zero to be treated as such. Building on these concepts, Murphy et al. (2014) showed how the entire body of methods subsumed under the general linear model (e.g. correlation, *t*-tests, Analysis of Variance (ANOVA), Analysis

of Covariance (ANCOVA) and Multiple Regression) could be adapted to test the hypothesis that the effects of interventions fell within a range of values that were all so trivially small that we could conclude with confidence that whatever effect interventions might have, they were too small to care about. Alternatively, one might move away from the binary (significant vs. non-significant) classification systems these methods imply to simply describe the range of plausible values for key parameters, using tools such as confidence intervals (Fidler et al., 2004). I will return to this suggestion below.

Secondly, it is important to change the way in which we view and report our results. In essence, we need to change our collective 'scientific language'. Readers often appear to interpret the phrase 'this result is statistically significant' to mean 'this result is important'. This has to stop. Thirdly, accept uncertainty. We sometimes compute confidence intervals or some similar measure, but we do this mainly to see whether or not our confidence interval includes zero (or whatever value is used to define a particular point hypothesis). It is much better to be aware of, and to understand, the implications of uncertainty in estimating population statistics from samples and to keep this uncertainty in mind when interpreting the results. Fourthly, be thoughtful in the design, analysis and interpretation of studies. In particular, we should use statistics and data analysis as a tool for helping us understand what the data mean. The use of NHST as a tool for making binary decisions (significant vs. non-significant) discourages thoughtful analysis; as we move away from a mechanical procedure for evaluating data, we will be forced to move towards methods of data analysis that compel us to think about what the data mean. Fifthly, be open. If we get rid of a simple widely known (but not widely understood) procedure for making sense of data, we are going to have to fall back on informed judgement for evaluating results. That is, we will have to present and defend criteria for evaluating data rather than falling back on a familiar but ultimately meaningless procedure such as NHST.

Finally, Wasserstein et al. (2019) encourage scientists to be open and modest in the interpretation of their data. They advocate that we operate using the ATOM principle, that is, *accept* that there will always be uncertainty, and be *thoughtful, open* and *modest*. This strikes me as a very useful piece of advice for the *South African Journal of Industrial Psychology*, its contributing authors and incoming editorial board.

Conclusion

Will reducing our reliance on NHST solve all of the problems that have come to light in recent research on the credibility and reproducibility of our results? Certainly not! However, it is a good start towards enhancing the credibility, reproducibility and meaning of the research published in our journals. Does this mean that we should abandon significance tests altogether? Probably not; this would be a

step too far for many of the journal's readers, authors, editors and reviewers. However, I believe that it is a realistic step to require all studies to start with a realistic power analysis and to report the results of this analysis. This would have two salutatory effects. Firstly, it would allow authors to think concretely about what sort of effects they expect to observe and to justify their assumptions about effect size. Secondly, it would discourage authors from attempting to use small sample studies to address important problems. Unless the effect you expect to observe (and can articulate a realistic basis for this expectation) is quite large, power analyses will encourage you to collect larger sample than you might otherwise settle for. Larger samples would enhance the stability and reproducibility of the results reported in our journals. Serious attention to statistical power would also render most NHST tests essentially moot. The whole point of performing a power analysis before you collect data is to help you design a study where you will easily reject the null hypothesis H_0 .

Finally, a wholehearted adoption of power analysis will help to wean authors away from dependence on significance tests. If these tests become a foregone conclusion, their apparent value as evidence is likely to decline. This will not totally solve the problem of creating credible and reproducible science, but it strikes me as the best first step.

Acknowledgements

Competing interests

The author declares that they have no financial or personal relationships which may have inappropriately influenced them in writing this article.

Author's contributions

K.R.M. is the sole contributor to this article.

Ethical considerations

I confirm that ethical clearance was not needed or required for the study.

Funding information

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data availability statement

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Disclaimer

The views and opinions expressed in this article are those of the author and do not necessarily reflect the official policy or position of any affiliated agency of the author.

References

- Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*, 305–307. <https://doi.org/10.1038/d41586-019-00857-9>
- Bakker, M., Van Dijk, A., & Wicherts, J.M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 534–554. <https://doi.org/10.1177/1745691612459060>
- Banks, G.C., Rogelberg, S.G., Woznyj, H.M., Landis, R.S., & Rupp, D.E. (2016). Editorial: Evidence on questionable research practices: The good, the bad and the ugly. *Journal of Business and Psychology*, *31*, 323–338. <https://doi.org/10.1007/s10869-016-9456-7>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Social Psychology*, *65*, 145–153. <https://doi.org/10.1037/h0045186>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>
- Efendic, E., & Van Zyl, L.E. (2019). On reproducibility and replicability: Arguing for open science practices and methodological improvements at the South African Journal of Industrial Psychology. *SA Journal of Industrial Psychology*, *45*(0), a1607. <https://doi.org/10.4102/sajip.v45i0.1607>
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think. *Psychological Science*, *15*, 119–126. <https://doi.org/10.1111/j.0963-7214.2004.01502008.x>
- Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS Medical*, *2*, e124. <https://doi.org/10.1371/journal.pmed.0020124>
- McNutt, M. (2014). Reproducibility. *Science*, *343*, 229. <https://doi.org/10.1126/science.1250475>
- Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834. <https://doi.org/10.1037/0022-006X.46.4.806>
- Murphy, K., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (4th edn.). New York: Taylor & Francis.
- Neuroskeptic. (2012). The nine circles of scientific hell. *Perspectives on Psychological Science*, *7*, 643–644. <https://doi.org/10.1177/1745691612459519>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. <https://doi.org/10.1126/science.aac4716>
- Pashler, H., & Wagenmakers, E.J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. <https://doi.org/10.1177/1745691612465253>
- Rouanet, H. (1996). Bayesian methods for assessing the importance of effects. *Psychological Bulletin*, *119*, 149–158. <https://doi.org/10.1037/0033-2909.119.1.149>
- Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129. <https://doi.org/10.1037/1082-989X.1.2.115>
- Serlin, R.A., & Lapsley, D.K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 73–83. <https://doi.org/10.1037/0003-066X.40.1.73>
- Serlin, R.A., & Lapsley, D.K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – Or vice versa. *Journal of the American Statistical Association*, *54*, 30–34. <https://doi.org/10.1080/01621459.1959.10501497>
- Wasserstein, R.L., & Lazar, N.A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. <https://doi.org/10.1080/00031305.2016.1154108>
- Wasserstein, R.L., Schirm, A.L., & Lazar, N.A. (2019). Moving to a world beyond 'p < 0.05'. *The American Statistician*, *73*(sup1), 1–19. <https://doi.org/10.1080/00031305.2019.1583913>