 doi:10.1111/spsr.12376

Debate

The Use and Usefulness of *p*-Values in Political Science: Introduction

Daniel Bischof[1] and Mariken van der Velden[2]
[1]Department of Political Science, University of Zurich (CH)
[2]Department of Communication Science, Vrije Universiteit Amsterdam (NL)

P-values are the most frequently employed metric to assess the significance of statistical findings in the social sciences. Since the earliest years of their usage the meaning and usefulness of *p*-values were topics of heated discussion (Berkson 1942; Fisher 1935). Lately the reproduction/replication crisis resuscitated this debate (Benjamin et al. 2018; Gelman 2018; Lakens et al. 2018; McShane et al. 2019; Nuzzo 2014; Trafimow and Marks 2015). Meanwhile, the skepticism has not stopped at the gates of political science. Most prominently the journal "Political Analysis" banned *p*-values "in regression tables or elsewhere" after the new editor took over the board of editors in 2017 (Gill 2018: 1).[1] Also political scientists contributed to a swelling debate suggesting to lower the threshold for *p*-values to 0.005 (Benjamin et al. 2018; Esarey 2017).

This present debate seeks to contribute to the discussion on *p*-values by summarizing the main arguments of it, providing an encompassing discussion of *p*-values – also from an epistemological perspective – as well as advice for the discipline about the Do's and Don'ts for *p*-values. In doing so it contributes to two ongoing debates: First, the actual meaning of *p*-values – the mathematical definition. Second, the potential to misuse *p*-values – even if correctly understood and defined. In February 2018 the Department of Political Science at the University of Zurich held a workshop containing public lectures about the use and usefulness of *p*-values. The goal of the workshop was to cover a wide spectrum of opinions on *p*-values, covering frequentist, bayesian and more epistemological views. The present contribution summarizes these public lectures but also goes beyond them by giving each author the opportunity to engage with the position of the remaining authors on the one hand and by adding a critical discussion of all lectures in the conclusion on the other hand.

Our introduction first gives a brief history of *p*-values and their definition before summarizing the discussion about the use and "misuse" of *p*-values in the discipline and situates this into the larger debate on the replication crisis; Vera Troeger (2019) discusses the logic of statistical inference and significance testing and the implications of significance testing for empirical research; Susumu Shikano (2019) provides a detailed insight into two Bayesian approaches to hypothesis testing; Marco R. Steenbergen (2019) takes a stronger epistemological view on to *p* or not to *p*; and finally, Simon Hug (2019) critically discusses

[1] Later, this statement was changed to allow *p*-values in specific cases.

the contributions of this debate by emphasizing that none of the discussed approaches appear to provide a "silver bullet" for the issues they seek to address.

A Brief History of *p*

So, what are *p*-values and why do we as researchers care so much about them? *p*-values date back to the 18th century where Pierre Simon-Laplace came-up with first ad hoc definitions of *p*-values. Later on in the 20th century Pearson formalized them in his χ^2 test and then Fisher popularized them by also proposing the threshold of 0.05 for statistical significance. The first references from Fisher on the threshold of 0.05 stems from his well-known "lady testing tea" experiment. Muriel Bristol, a phycologist and enthusiastic tea drinker, claimed to be able to differentiate tea which was poured on milk from milk which was poured on tea – of course keeping the amount of tea and milk constant.

Fisher and his friend William Roach decided to test Bristol's tea tasting skills with a simple experiment: Muriel Bristol was provided eight cups of tea (four prepared by first adding milk; four prepared by first adding tea). Bristol then was asked to name the four cups prepared by her be-liked method. Thus, the null hypothesis was that Muriel Bristol did not have the ability to distinguish the preparation of tea. Given $n = 8$ cups and $k = 4$ chosen cups the experiment results in 70 possible combinations.[2] In order to reject the null hypothesis Fisher suggested that Bristol needed to get four out of four cups right. The combination of four correctly classified cups has a chance to occur in one out of 70 combinations. Bristol eventually got all eight cups correct.

Fisher discusses the threshold of 5% in close relation to the lady testing tea experiment. As outlined above Bristol's performance had a chance to occur in only 1.4 per cent, while if she had missed only a single cup the chances to observe such a performance would have increased drastically to 24.3 per cent.[3] In the latter case Fisher believed the likelihood of observing such a performance just by chance was too high. Future research built on Fisher's reasoning – without him necessarily having had the intention to be used as the default approach to hypotheses testing – and eventually stopped discussing the reasons for the 5% threshold entirely. As this example illustrates, what *p*-values then really tell us is how likely our data are, assuming that our H_0 (*Bristol not having the skills to tell the difference between the two tea preparation methods*) is true (Wasserstein and Lazar 2016). A standard for empirical testing was born and until today this standard guides social scientists' behavior, evaluations of research and most prominently publication standards.

But why have researchers recently 'seen the light' and paid attention to the shortcomings of *p*-values? In 2011, the renowned social-psychologist Diederik Stapel was found guilty of fraudulent research practices.[4] As it turned out Stapel had faked his entire data collection. He simply answered to his questionnaires himself and thereby created the data he and his research team then analyzed. His research fraud sparked a larger debate within psychology: To what extent was there a culture of "sloppy" science, in which some scientists did not understand the essentials of statistics, reviewers for journals encouraged researchers to leave unwelcome data out of their papers, and even the most prestigious

[2] $\frac{8!}{4!(8-4)!} = 70$

[3] $\frac{16+1}{70} = 0.243$

[4] http://www.apa.org/science/about/psa/2011/12/diederik-stapel.aspx

journals printed results that were obviously too good to be true?[5] The Open Science Collaboration (2015) replicated 100 studies published in psychology journals (Aarts et al. 2015). Using high-powered designs, they found that their mean effect size was approximately half of the size of the original articles.[6] Moreover, while 97% of the original studies had demonstrated significant results ($p < 0.05$), only 47% of the replicated studies had significant results – indicating that 53% of the studies could not be replicated. This is not only a problem of psychology, where the norm is to publish based on experimental studies. In economics, also half of the studies could not be replicated (Chang and Li 2015). Chang and Li (2015) replicate 67 original articles published in 13 well-regarded macro-economics and general interest economic journals and demonstrate that replication issues are not tied to using experimental data, but equally apply to studies using publicly available data sets. In political science the debate caught fire with the Mike LaCour case. LaCour did not only follow "sloppy" research practices but committed fraud by inventing data he never had collected in the first place (Broockman et al. 2015).

These examples emphasize that even if we understand p-values correctly from a mathematical perspective, they can be misused by overconfident claims about findings surpassing the arbitrary $p < 0.05$ threshold. They suggest that practices of 'p-hacking' are more likely to occur in an environment focusing so much on the question of whether $p < 0.05$. To some extent the idea of relying on such a threshold guides and motivates fraud in science and are then amplified by the human tendencies of *apophenia* - seeing patterns in random data - and of *conformation bias* - focusing on evidence that is in line with our (favored) explanation. These human tendencies are likely to affect how we walk through the 'garden of forking paths' when conducting analysis (Gelman and Loken 2013) and how we interact with the question of 'researcher degrees of freedom' (Simmons et al. 2011). And in many instances making the threshold is just one tiny step away – e.g. by adding/dropping a control, an interaction term or dropping some unfavorable outliers.

Thus, from our point of view not only the practices of how we engage and interpret p-values need to change, but eventually the environment under which we conduct research needs to adapt as well. As all contributions to this issue show: lowering the threshold for significance is unlikely to achieve this goal (Benjamin et al. 2018). A design-based derivation of the threshold might be better-equipped to achieve this goal (Lakens et al. 2018), but similar to the issue of the 'garden of forking paths' leaves researchers potentially with too many 'degrees of freedom'. Proposals calling for a purely Bayesian approach to questions of significance tend to ignore that eventually we will run in very similar questions and issues irrespective if we choose a Bayesian or Frequentist perspective. Thresholds and guidelines are also the likely outcome if we go Bayesian – as Simon Hug's discussion in this debate correctly points out.

Instead, we understand the replication/reproduction crisis as a symptom for a larger, systematic problem in the social sciences. This problem speaks to all aspects of what we are as a profession. It speaks to: how we teach empirical research practices, how we engage with changing practices in data sciences and how we question our own past and present behavior as scientists. But most importantly it suggests that no matter how we engage with p-values in the future as a profession – even if we correctly interpret them and teach them from a mathematical perspective –, proposals for change need to take into

[5] http://www.sciencemag.org/news/2012/11/final-report-stapel-affair-points-bigger-problems-social-psychology

[6] Notice, however, that there is an ongoing debate about the research design and inferences done in the replication studies questioning the meaningfulness of its findings (Anderson et al. 2016; Gilbert et al. 2016).

account how much increasing competition on the academic job market, publication pressures and review practices will affect proposed reforms to engage with *p*. Having initiated the present debate, the authors hope that it helps colleagues to understand the issues involved with *p* while highlighting potential approaches how to address the issue with *p*.

ACKNOWLEDGMENTS

We are thankful for the comments by two reviewers and the editor which helped us to significantly improve our introduction and the entire debate.

References

Aarts, A. A., J. E. Anderson, C. J. Anderson, P. R. Attridge, A. Attwood, J. Axt, M. Babel et al. (2015). Estimating the reproducibility of psychological science. *Science* 349(6251): aac4716.

Anderson, C. J., Š. Bahník, M. Barnett-Cowan, F. A. Bosco, J. Chandler, C. R. Chartier, F. Cheung et al. (2016). Response to Comment on "Estimating the reproducibility of psychological science". *Science* 351(6277): 1037.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen et al. (2018). Redefine statistical significance. *Nature Human Behaviour* 2(1): 6–10.

Berkson, J. (1942). Tests of Significance Considered as Evidence. *Journal of the American Statistical Association* 37(219): 325–335.

Broockman, D., J. Kalla and P. M. Aronow (2015). Irregularities in LaCour (2014) Timeline of Disclosure. Online: https://stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf [accessed: 07.09.2019].

Chang, A. and P. Li (2015). Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'. Online: https://www.federalreserve.gov/econresdata/feds/2015/files/2015083pap.pdf [accessed: 07.09.2019].

Esarey, J. (2017). Lowering the Threshold of Statistical Significance to $p < 0.005$ to Encourage Enriched Theories of Politics. *The Political Methodologist* 24(2): 13–19.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh; London: Oliver And Boyd.

Gelman, A. (2018). Ethics in statistical practice and communication. *Significance* 138(83): 40–43.

Gelman, A. and E. Loken (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis. Online: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf [accessed 07.09.2019].

Gilbert, D. T., G. King, S. Pettigrew and T. D. Wilson (2016). Comment on "Estimating the reproducibility of psychological science". *Science* 351(6277): 1037.

Gill, J. (2018). Comments from the new Editor. *Political Analysis* 26(1): 1–2.

Hug, S. (2019). Just Say No to $p < \times (\forall \times 2 (0,1])$, *s and Other Evil Things. *Swiss Political Science Review* 25(3).

Lakens, D., J. Grange, F. Adolfi, C. Albers, F. Anvari, M. Apps, S. Argamon et al. (2018). Justify your alpha. *Nature Human Behaviour* 2(3): 168–171.

McShane, B. B., D. Gal, A. Gelman, C. Robert and J. L. Tackett (2019). Abandon statistical significance. *The American Statistician* 73(sup1): 235–245.

Nuzzo, R. (2014). Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature* 506(7487): 150–152.

Shikano, S. (2019). Hypothesis Testing in the Bayesian Framework. *Swiss Political Science Review* 25 (3).

Simmons, J. P., L. D. Nelson and U. Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11): 1359–1366.

Steenbergen, M. R. (2019). What Is in a (Non-)Significant Finding? Moving Beyond False Dichotomies. *Swiss Political Science Review* 25(3): 1–17.

Trafimow, D. and M. Marks (2015). Editorial. *Basic and Applied Social Psychology* 37(1): 1–2.

Troeger, V. (2019). To P or not to P? The Usefulness of P-values in Quantitative Political Science Research. *Swiss Political Science Review* 25(3).

Wasserstein, R. L. and N. A. Lazar (2016). The ASA's Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician* 70(2): 129–133.

Daniel Bischof is Ambizione Grant Holder at the University of Zurich. His research focuses mostly on comparative politics and political economy and is published in the *American Journal of Political Science, British Journal of Political Science* and *European Journal of Political Research* amongst others. E-mail: bischof@ipz.uzh.ch

Mariken A.C.G. van der Velden is an assistant professor of political communication and VENI laureate (2019) at the department of Communication Science at the *Vrije Universiteit* Amsterdam, NL.

