This is a post-print. Official reference: Nuijten, M.B. (2016). Preventing statistical errors in scientific journals. European Science Editing, 42, 1, 8-10.

Preventing Statistical Errors in Scientific Journals

Michèle B. Nuijten

Tilburg University

Abstract

There is evidence for a high prevalence of statistical reporting errors in psychology and other scientific fields. These errors display a systematic preference for statistically significant results, distorting the scientific literature. There are several possible causes for this systematic error prevalence, with publication bias as the most prominent one. Journal editors could play an important role in preventing statistical errors in the published literature. Concrete solutions entail encouraging sharing data and preregistration, and using the automated procedure "statcheck" to check manuscripts for errors.

Keywords: statistical errors, publication bias, statcheck, data sharing

Author Note

Correspondence concerning this article should be addressed to Michèle Nuijten, PO Box 90153, 5000 LE Tilburg, The Netherlands, M.B.Nuijten@tilburguniversity.edu.

The preparation of this article was supported by VIDI grant number 016-125-385 from the Netherlands Organization for Scientific Research (NWO).

In a recent study [1], we documented the prevalence of statistical reporting inconsistencies in more than 250,000 *p*-values from eight major psychology journals, using the new R package "statcheck" [2]. The program *statcheck*: converts PDF and HTML articles to plain text files; extracts results of null hypothesis significance tests that are reported exactly according to APA style [3]; recomputes the *p*-value based on its accompanying test statistic and degrees of freedom, and checks if the reported *p*-value matches the recomputed *p*-value, taking rounding of the reported test statistic into account. We found that in half of the papers at least one *p*-value was inconsistent with the test statistic and degrees of freedom. In most of these cases, the reported *p*-value was only marginally different from the recomputed *p*-value. However, we also found that one in eight papers (12,5%) contained gross inconsistencies that may have affected the statistical conclusions: in those cases the reported *p*-value was significant, but the recomputed *p*-value was not, or vice versa. We found a higher prevalence of gross inconsistencies in *p*-values reported as significant, than *p*-values reported as non-significant, implying a systematic bias towards statistically significant findings.

This high prevalence of statistical errors in psychology papers is alarming, and there is evidence that this problem is not unique for psychology. Similar inconsistency rates have been found in, for instance, the medical sciences in general [4] and psychiatry in particular [5]. Even though small reporting errors might be inconsequential, wrongly reporting a *p*-value of .37 as .36 will probably not have serious effects, the apparent focus on significant results is worrying and can have far-reaching consequences. It may have added to the excess of (false) positive findings in science [6] [7].There are several explanations for this high error prevalence. Firstly, most of the inconsistencies could have been caused by mere sloppiness. Especially in psychology this is easy to imagine, since a single psychology paper on average already contains about ten statistical tests

[1]. In the tangle of statistical output, it is imaginable that a *p*-value (or test statistic or degree of freedom) is copied incorrectly. Matters probably become worse because many researchers are not in the habit of double checking their own or their co-authors' analyses who sometimes do not even have access to the raw data in the first place; [8]. However, sloppiness alone does not explain the apparent systematic preference for significant findings.

A possible explanation for the excess of *p*-values wrongly reported as significant is publication bias: significant results have a higher probability to be published than non-significant results [9-11]. It is imaginable that researchers just as often wrongly report a significant *p*-value as a non-significant *p*-value. However, because of publication bias, only the gross inconsistencies that wrongly present a *p*-value as significant are published, resulting in a systematic bias in favour of significant findings. Conversely, it is also possible that researchers *suspect* that their findings will not be published if they do not find a significant effect, and because of this, they more often wrongly round down a non-significant *p*-value to obtain a significant finding, than vice versa. This would be in line with the finding of John, Loewenstein (12), who found that 22% of a sample of over 2000 psychologists admitted to knowingly having rounded down a *p*-value to obtain significance, which would lead to an excess of false positive findings. Of course it could also just be the case that researchers unknowingly maintain double standards concerning the checking of their results: they would inspect their results with more scrutiny when the result is unexpectedly nonsignificant, but not when it is significant.

I believe journal editors can play an important role in preventing, detecting, and/or correcting statistical errors in scientific literature. There are several concrete steps that could be taken to actively improve the state of the published literature.

A possible solution to the problem of statistical reporting errors is to promote data sharing. In previous research it has been found that if researchers were unwilling to share data of a certain paper, there was a higher probability that the paper contained reporting errors, often concerning statistical significance [13]. This finding could illustrate that authors are aware of the inconsistencies in their paper and refuse to share their data out of fear to be exposed. An alternative explanation for this finding is that researchers who manage their data with more rigor both make fewer mistakes and archive their data better, which makes data sharing easier. In both cases the prevalence of reporting errors might decrease when journal editors would encourage data sharing.

Besides the possibility that authors themselves may become more precise in reporting their results if they have to share their data, encouraging data sharing has more benefits. If authors would submit their data and analysis scripts alongside their manuscript, it would allow for so-called analytic review [14]. In analytic review, peer reviewers or statistical experts verify if the reported analyses and results are in line with the provided data and syntax. Not only will this encourage authors to manage their data more carefully in order for a third party to understand it, statistical errors that were overlooked at first have a higher probability of being detected before publication.

Editors could decide to make data sharing mandatory, taking into account certain exceptions concerning privacy etc. (see e.g. the policy of PLoS One). Another option is to simply reward authors who share data. For instance, the journal Psychological Science awards badges to papers that are accompanied by open data and also awards badges for open materials and preregistered studies. Although at first sight these badges might seem trivial, they can be considered a quality seal and have inspired many researchers to share their data.

Of course, researchers could still conceal deliberate rounding errors towards significance by manipulating the raw data before submitting them. However, falsifying research data like this

is explicit scientific fraud. Data from self-reports show that scientific fraud is much more uncommon than questionable research practices such as wrongly rounding a *p*-value[12], so it seems implausible that encouraging data sharing will result in researchers hiding rounding errors by manipulating the raw data. In any case, there will always remain ways to commit fraud in science, but encouraging data sharing will definitely make it harder.

Another way to avoid reporting errors and to facilitate analytic review, is for editors of journals that adhere to APA reporting style to make use of statcheck [2]. As described above, statcheck is a package for the statistical software R[15] that can automatically scan articles, extract statistical results reported in APA style, and recompute *p*-values. Editors could make it standard practice to use statcheck to automatically scan papers upon submission to check for statistical reporting inconsistencies. This takes almost no time; on average, statcheck can scan approximately 250 papers per minute. Since many journals already have an automatic plagiarism check, it is a small step of adding a check for reporting inconsistencies. Results that are flagged as problematic can then be corrected before publication. R and statcheck are both open source and freely available. For more information about statcheck and an extensive analysis of its validity, see our paper [1]. For instructions on how to install statcheck, see http://mbnuijten.com/statcheck.

The excess of results wrongly presented as significant is probably caused by publication bias. A promising way for editors to try to avoid publication bias is to encourage preregistration. Preregistration can take many forms, but in general the idea is that researchers write a detailed research (and analysis) plan *before* collecting the data. This research plan is then "registered" somewhere online (e.g., in a repository for clinical trials such as https://www.clinicaltrialsregister.eu), or even submitted to a journal. In the latter case, the research plan is peer reviewed, and if the plan meets the standards of the journal, the researchers can receive

an "in principle acceptance", no matter what the results will be – given that they will adhere to the research plan (see e.g. the guidelines for registered reports in the journals Cortex, Comprehensive Results in Social Psychology, and Perspectives on Psychological Science). This way, the decision to publish a paper cannot be influenced by whether the results were significant or not, avoiding the selective publishing of *p*-values wrongly rounded down as compared to the ones wrongly rounded up. On top of that, it takes away an incentive for researchers to deliberately report a non-significant *p*-value as significant.

Besides side-stepping publication bias and avoiding systematic reporting errors, preregistration also solves the problem of HARKing: Hypothesizing After the Results are Known [16]. When researchers are HARKing, they first explore the data to find interesting patterns, and then present these findings as having been predicted from the start. If a researcher performs a lot of exploratory tests, he or she is bound to find at least one significant result purely by chance. Reporting only the tests that were significant leads to an excess of false positive findings. However, if the research plan and hypotheses are registered beforehand, there is a clear distinction between confirmatory and exploratory tests in the paper, which allows for a more reliable interpretation of the results [17].

To conclude, there is evidence for a high prevalence of statistical reporting inconsistencies in the scientific literature. Even though many of these inconsistencies are minor errors that are probably due to mere sloppiness, there is also a high prevalence of gross inconsistencies that may have affected the statistical conclusion, mainly in favour of statistical significance. Even though we can only speculate why there are more results wrongly presented as significant (deliberately rounding down, publication bias, less rigorous checks of findings in line with expectations, etc.) it

remains a worrying finding, reflecting a systematic preference for "success" and leading to an excess of false positive findings in the literature.

There are several concrete steps that journal editors can take in order to avoid or reduce the number of reporting errors. For instance, editors could encourage data sharing and preregistration, or use the program statcheck to automatically check for inconsistencies during the review process. Besides decreasing the prevalence of reporting errors, these measures also reduce publication bias, HARKing, and other questionable research practices.

Statistical reporting errors are not the only problem we are currently facing in science but at least it seems like one that is relatively easy to solve. I believe journal editors can play an important role in achieving change in the system, in order to slowly but steadily decrease statistical errors and improve scientific practice.

References

1. Nuijten MB, Hartgerink CHJ, Van Assen MALM, Epskamp S, Wicherts JM. The prevalence of statistical reporting errors in psychology (1985-2013). Behavior Research Methods. 2015. doi: 10.3758/s13428-015-0664-2

2. Epskamp S, Nuijten MB. statcheck: Extract statistics from articles and recompute p values. R package version 1.0.1. http://CRAN.R-project.org/package=statcheck2015.

3. American Psychological Association. Publication Manual of the American Psychological Association. Sixth Edition. Washington, DC: American Psychological Association; 2010.

4. Garcia-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. BMC Medical Research Methodology. 2004;4:13. doi: 10.1186/1471-2288-4-13

5. Berle D, Starcevic V. Inconsistencies between reported test statistics and p-values in two psychiatry journals. International Journal of Methods in Psychiatric Research. 2007;16(4):202-7. doi: 10.1002/mpr.225

6. Francis G. The Frequency of Excess Success for Articles in Psychological Science. Psychonomic Bulletin & Review. 2014;21:1180-7. doi: 10.3758/s13423-014-0601-x

7. Fanelli D. "Positive" Results Increase Down the Hierarchy of the Sciences. PLoS One. 2010;5(3):e10068. doi: 10.1371/journal.pone.0010068

8. Veldkamp CLS, Nuijten MB, Dominguez-Alvarez L, van Assen MALM, Wicherts JM. Statistical reporting errors and collaboration on statistical analyses in psychological science. Plos One. 2014;9(12):e114876. doi: 10.1371/journal.pone.0114876

9. Greenwald AG. Consequences of prejudice against the null hypothesis. Psychological Bulletin. 1975;82:1-20.

10. Sterling TD. Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance--Or Vice Versa. Journal of the American Statistical Association. 1959;54(285):30-4.

11. Sterling TD, Rosenbaum WL, Weinkam JJ. Publication decisions revisited - The effect of the outcome of statistical tests on the decision to publish and vice-versa. American Statistician. 1995;49(1):108-12.

12. John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth-telling. Psychological Science. 2012;23:524-32. doi: 10.1177/0956797611430953

13. Wicherts JM, Bakker M, Molenaar D. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. PLoS One. 2011;6(11):e26828. doi: 10.1371/journal.pone.0026828

14. Sakaluk J, Williams A, Biernat M. Analytic Review as a Solution to the Misreporting of Statistical Results in Psychological Science. Perspectives on Psychological Science. 2014;9(6):652-60. doi: 10.1177/1745691614549257

15. R Core Team. R: A Language and Environment for Statistical Computing. http://www.R-project.org/2014.

16. Kerr NL. HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review. 1998;2:196-217.

17. Wagenmakers EJ, Wetzels R, Borsboom D, Maas HLJvd, Kievit RA. An agenda for purely confirmatory research. Perspectives on Psychological Science. 2012;7:632-8. doi: 10.1177/1745691612463078