# USING "STATCHECK" TO DETECT AND PREVENT STATISTICAL REPORTING INCONSISTENCIES

Michèle B. Nuijten
Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences,
Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands
m.b.nuijten@uvt.nl

*Roughly half of the articles in psychology contain at least one inconsistent NHST result in which the p-value does not match the test statistic and degrees of freedom. In one in eight articles, these inconsistencies may actually lead to a different statistical conclusion. To detect and correct these kind of inconsistencies, we developed the free tool "statcheck". Statcheck is an R package with accompanying web app (http://statcheck.io) that automatically extracts statistics from paper, recalculates the p-values based on the reported test statistic and degrees of freedom, and checks if the result is internally consistent. If we teach our students to incorporate statcheck in their workflow, they can avoid statistical misreporting in their own work and easily scrutinize the statistical validity of conclusions in the literature.*

## STATISTICAL REPORTING INCONSISTENCIES IN PSYCHOLOGY

Most conclusions in psychological research are based on null hypothesis significance testing (NHST; Cumming et al., 2007; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). Therefore, it is important that NHST results are reported correctly. Unfortunately, there is increasing evidence that this is often not the case. Roughly half of the published psychology papers contain at least one inconsistent NHST result in which the reported *p*-value does not match the accompanying test statistic and degrees of freedom (df). About one in eight psychology papers contains at least one *gross* inconsistency; in these cases the reported *p*-value is significant ($\alpha = .05$) but the recalculated *p*-value based on the test statistic and df is not, or vice versa (Bakker & Wicherts, 2011; Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2016).

### *Causes of Inconsistencies*

The main cause of this high prevalence of inconsistencies in psychology is unclear. It is plausible that the large majority of misreported statistics are caused by innocent mistakes. Most psychologists have not had extensive statistical training, and it can be confusing which exact numbers belong to which statistical result. For instance, in SPSS output, the two degrees of freedom of an F-test can be reported separately in different tables. Another example of a common slip-up is that researchers report correlations with the accompanying sample size rather than the degrees of freedom. Besides copying the wrong numbers to the paper, simple typos are also a very likely cause of many inconsistencies.

It is possible that inconsistencies are caused by intentional misrepresentation of the result. This becomes more likely in cases in which the reported *p*-value is significant, and the recalculated *p*-value is not. In many scientific fields, there is high pressure to find significant results which may lead to questionable research practices (QPRs; Bakker, van Dijk, & Wicherts, 2012; Simmons, Nelson, & Simonsohn, 2011). Common examples of such QRPs are the selective reporting of measures that showed the desired effect, or removing outliers based on their impact on the conclusion. In a study about the prevalence of QRPs, it was found that 22% of the surveyed psychologists admitted to having wrongly rounded down a *p*-value to make it appear significant (e.g., rounding $p = .054$ to $p < .05$; John, Loewenstein, & Prelec, 2012). Indeed, patterns in misreported NHST results point into the direction of a systematic bias towards significant results (Nuijten et al., 2016; Wicherts, Bakker, & Molenaar, 2011).

Systematic bias in misreported results does not necessarily have to be caused by intentionally wrongly rounding *p*-values. It is also possible that all inconsistent NHST results are caused by random error, but because there is publication bias and significant results have a higher chance of being published, only the *p*-values that are wrongly rounded *down* end up in the literature. Another alternative cause for the excess of *p*-values that seem too low, is a double standard in checking results. If a result is not in line with your hypothesis, you might be more inclined to double check all analyses

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2018 ISI/IASE]

and their output and spot copying errors or typos, than when you find a significant result that is already in line with your expectations.

### *Consequences of Inconsistencies*

Regardless of their cause, statistical reporting inconsistencies undermine the trustworthiness of the scientific conclusions. This becomes immediately clear in the case of gross inconsistencies, in which the statistical conclusion actually changes when the *p*-value is recalculated. However, inconsistencies that do not affect the statistical conclusion can still have serious consequences. For instance, NHST results are often used as the input in meta-analyses, and misreported results can affect the overall meta-analytic conclusion (Bakker & Wicherts, 2011). Furthermore, inconsistent results diminish the level of reproducibility of a paper (Nuijten, Bakker, Maassen, & Wicherts, in press): an independent research would not be able to reproduce the reported results based on the original raw data and analyses.

In short, statistical reporting inconsistencies are widespread in psychology and can have serious consequences for the interpretation and trustworthiness of scientific findings. Therefore, it is important that we find a way to detect, correct, and prevent inconsistencies in the literature.

## SPOT INCONSISTENCIES WITH STATCHECK

Checking papers for statistical inconsistencies is time consuming and error prone work, so we developed the R package “statcheck” (Epskamp & Nuijten, 2014) and the accompanying web app at http://statcheck.io to automatically extract statistical results from papers and recompute *p*-values. The algorithm behind statcheck roughly encompasses four steps:

*Step 1*: Convert an article from a PDF or HTML file to a plain text file. In the web app it is also possible to upload a Word file.

*Step 2:* Use regular expressions to look for APA reported NHST results of *t*-test, *F*-tests, $\chi^2$-tests, *z*-tests, and correlations. Statcheck identifies NHST results by looking for specific combinations of letters, numbers and mathematical symbols such as parentheses or equal signs. This means that any deviation from APA style will not be picked up by statcheck.

*Step 3*: Use the reported degrees of freedom and test statistic to recompute the *p*-value. By default, statcheck assumes two-tailed testing.

*Step 4*: Compare the reported and recomputed *p*-value. If they do not match, statcheck flags this as an inconsistency. If the reported *p*-value is significant and the recomputed *p*-value is not, or vice versa, statcheck flags this as a gross inconsistency. Statcheck assumes a default α of .05. A reported $p = .05$ is considered significant, because most researchers interpret it this way (Nuijten et al., 2016).

There are a couple of things that statcheck takes into account. First, statcheck takes into account correct rounding of the test statistic. For instance, a reported *t*-value of 2.35 could correspond to an actual value of 2.345 to 2.354 with a range of *p*-values that can slightly deviate from the recomputed *p*-value. Statcheck will not count cases like this as inconsistencies. Second, statcheck has an option to take one-sided testing into account: if somewhere in the article the words “one-tailed”, “one-sided”, or “directional” are mentioned, and the result would have been correct if it was one-sided, it is counted as a correctly reported one-sided test.

There are two main things to note when using statcheck. First, statcheck implicitly assumes that any inconsistencies are caused by a wrong *p*-value, but it is also possible that the degrees of freedom, the test statistic, or a combination of these three elements caused the inconsistency. All that statcheck flags is that the three components of the result do not match. Second, statcheck can flag results as inconsistent if one or more components have been adjusted to correct for multiple testing, violations of assumptions. For instance, if a researcher multiplies her *p*-values by the number of tests in the paper (a common interpretation of the Bonferroni correction), the adjusted *p*-value does not match the rest of the result anymore and statcheck will flag it as an inconsistency.

More details about what statcheck can and cannot do, how to install and use it, and how to interpret the results can be found in the manual at http://rpubs.com/michelenuijten/statcheckmanual.

### *Accuracy of statcheck*

It is important that statcheck is accurate in categorizing results as consistent, inconsistent, or grossly inconsistent. To investigate statcheck's accuracy, we ran statcheck on a set of articles that were also manually coded for inconsistencies, and compared the results (Nuijten, Van Assen, Hartgerink, Epskamp, & Wicherts, 2017). The first main finding was that statcheck detects roughly 60% of the reported NHST results. The results that it did not find were not reported completely, not reported in APA style, or reported in tables. The second main finding was that statcheck's sensitivity (true positive rate) and specificity (true negative rate) were high: between 85.3% and 100%, and between 96.0% and 100%, respectively, depending on the assumptions and settings. The overall accuracy of statcheck ranged from 96.2% to 99.9%.

## HOW CAN STATCHECK BE USED

There are three main things that statcheck is particularly useful for: self-checks, peer review, and research. First, researchers can use statcheck to scan their paper for accidental typos or other mistakes in the statistics before submitting the manuscript to a journal. This way, statistical reporting inconsistencies are prevented at a very early stage from ending up in the literature. Second, statcheck can be used in peer review, to prevent statistical misreporting at the gate of the journal. Two flagship journals in psychology are already incorporating statcheck in their review process: *Psychological Science* and the *Journal of Experimental Social Psychology*. Additional journals such as *Advances in Methods and Practices in Psychological Science* are advising authors to use statcheck on their manuscript before submitting. A third way statcheck can be useful, is for research. Examples of research questions that can be answered through the use of statcheck concern the prevalence of reporting inconsistencies in large bodies of literature, the general discrepancies between reported and recomputed *p*-values, and general questions about the number of reported statistics, the type of tests that are used most often, and sample sizes as deduced from reported degrees of freedom.

All these potential uses of statcheck can also be translated to an educational setting. As long as students report their NHST results in APA style, they can scan any paper they write with statcheck to avoid handing in a paper with misreported statistics. Next, the teacher who has to grade their papers can use statcheck to quickly check if there are any misreported statistics in the paper. Finally, since statcheck is relatively straightforward to use, students can scan scientific articles to check them for any sign of statistical errors. This can be useful simply to check if the analyses and conclusions of a specific paper seem to hold, but students can also run statcheck on large sets of articles to analyze these for inconsistencies or other statistical characteristics.

## CONCLUSION

The prevalence of statistical reporting inconsistencies in published psychology papers is high. To detect these inconsistencies automatically, we developed the free tool statcheck. Statcheck automatically extracts APA reported NHST results and recalculates the *p*-value based on the reported test statistic and degrees of freedom. Statcheck can be a useful tool to prevent statistical inconsistencies to end up in the literature, if authors and peer reviewers or editors scan articles before publication. Statcheck can also be used for research purposes; for instance, to document the prevalence of inconsistencies in large bodies of literature or draw conclusions about the popularity of different types of statistical tests. Statcheck can also be useful in an educational setting. Students can quickly and easily check their own papers for accidental mistakes in the statistics, before handing them in. Likewise, teachers can quickly check students' submitted papers for statistical errors. Statcheck could also provide a useful tool for students to investigate published statistical results.

## REFERENCES

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554. doi:10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*(3), 666-678. doi:10.3758/s13428-011-0089-5

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., . . . Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological science, 18*(3), 230-232. doi:10.1111/j.1467-9280.2007.01881.x

Epskamp, S., & Nuijten, M. B. (2014). statcheck: Extract statistics from articles and recompute p values. R package version 1.0.0. http://CRAN.R-project.org/package=statcheck.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological science, 23*, 524-532. doi:10.1177/0956797611430953

Nuijten, M. B., Bakker, M., Maassen, E., & Wicherts, J. M. (in press). Verify original results through reanalysis before replicating: a commentary on "Making Replication Mainstream" by Rolf A. Zwaan, Alexander Etz, Richard E. Lucas, & M. Brent Donnellan. *Behavioral and Brain Sciences*.

Nuijten, M. B., Hartgerink, C. H. J., Van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods, 48*(4), 1205-1226. doi:10.3758/s13428-015-0664-2

Nuijten, M. B., Van Assen, M. A. L. M., Hartgerink, C. H. J., Epskamp, S., & Wicherts, J. M. (2017). *The Validity of the Tool "statcheck" in Discovering Statistical Reporting Inconsistencies*. Preprint retrieved from https://psyarxiv.com/tcxaj.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22*, 1359 –1366. doi:10.1177/0956797611417632

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. doi:10.2307/2282137

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited - The effect of the outcome of statistical tests on the decision to publish and vice-versa. *American Statistician, 49*(1), 108-112. doi:10.2307/2684823

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