The effectiveness of implementing *statcheck* in the peer review process to avoid statistical reporting errors

Michèle B. Nuijten[1] & Jelte M. Wicherts[1]

[1] Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University

Corresponding author: Michèle Nuijten, m.b.nuijten@tilburguniversity.edu.

Acknowledgements

Author Note

This report contains the results of the main, preregistered hypotheses tests. Additional exploratory analyses will follow in a more detailed version of this manuscript.

Abstract

We investigated whether statistical reporting inconsistencies could be avoided if journals implement the tool *statcheck* in the peer review process. In a preregistered study covering over 7000 articles, we compared the inconsistency rates between two journals that implemented *statcheck* in their peer review process (*Psychological Science* and *Journal of Experimental and Social Psychology*) with two matched control journals (*Journal of Experimental Psychology: General* and *Journal of Personality and Social Psychology*, respectively), before and after *statcheck* was implemented. Preregistered multilevel logistic regression analyses showed that the decrease in both inconsistencies and decision inconsistencies around $p = .05$ is considerably steeper in *statcheck* journals than in control journals, offering support for the notion that *statcheck* can be a useful tool for journals to avoid statistical reporting inconsistencies in published articles. We discuss limitations and implications of these findings.

Many conclusions in psychological research are based on the results of Null Hypothesis Significance Tests (NHST; Cumming et al., 2007; Nuijten et al., 2016). It is important that results of NHST are reported correctly: if we cannot rely on the reported numbers, we cannot rely on the overall robustness of the findings (Nuijten et al., 2018; Nuijten, 2021; Nosek et al., 2022). Unfortunately, in previous research we found a high prevalence of statistical reporting inconsistencies in published psychology articles (Nuijten et al., 2016). Specifically, ~50% of the articles with statistical results contained at least one p-value that did not match its accompanying test statistic and degrees of freedom. In ~12.5% of the articles with statistics we found at least one "decision inconsistency" (or "gross inconsistency"), in which the reported p-value was statistically significant, whereas the recomputed p-value was not, or vice versa (Nuijten et al., 2016).

In response to these high inconsistency rates, several journals started recommending or even requiring authors to scan their submitted manuscripts with *statcheck* (Nuijten & Epskamp, 2023). *Statcheck* is a free software tool that can automatically scan academic articles and detect inconsistencies in reported NHST results; effectively it acts as a "spellchecker" for statistics. Of course, the hope of the journals that implemented *statcheck* was that inconsistency rates would decrease, but this remains an empirical question. In this preregistered study, we empirically estimate the effectiveness of using *statcheck* in the peer review process in two major psychology journals that started using *statcheck* in 2016 (Psychological Science) and 2017 (Journal of Experimental Social Psychology). Specifically, we compare the trends in inconsistencies of these two journals against trends in two matched journals that did not have a specific policy to use *statcheck* to test the following hypothesis:

Articles published in journals that include statcheck in their peer review process show a steeper decline in statistical reporting inconsistencies and decision inconsistencies compared to matched journals that do not use statcheck.

The hypothesis and study protocol (rationale, methods, and analysis plan) were preregistered before data collection on recent articles on the Open Science Framework study registry on November 30rd, 2022 (https://osf.io/umwea). All departures from that protocol are explicitly acknowledged. We report all data exclusions and measures conducted during this study in this manuscript. The data and analysis scripts are available at https://osf.io/q84jn/. We cannot share the specific articles we scraped due to copyright restrictions, but our description of our sample should allow anyone with the relevant journal subscriptions to obtain the same sample.

Methods

Design

This is a retrospective observational study in which we compare statistical reporting inconsistencies in two sets of journals over time: journals that use *statcheck* in their peer review process and journals that do not.

Sample

Two journals that implemented statcheck in their peer review process are *Psychological Science* (PS) and the *Journal of Experimental and Social Psychology* (JESP). PS implemented statcheck in July 2016 (Nuijten, Borghuis, et al., 2017, p. 9). Their policy states:

"*StatCheck is an R program that is designed to detect inconsistencies between different components of inferential statistics (e.g., t value, df, and p). StatCheck is not designed to detect fraud, but rather to catch typographical errors (see https://mbnuijten.com/statcheck/ for more about StatCheck).* **Authors of accepted manuscripts must also provide a StatCheck report run on the accepted version of the manuscript that indicates a clean (i.e., error-free) result.** *A web app version of StatCheck can be accessed at http://statcheck.io/. If StatCheck does detect errors in the accepted*

version of the manuscript, authors should contact the action editor directly to determine the

best course of action."

(emphasis added;

https://www.psychologicalscience.org/publications/psychological_science/ps-submissions#S

TATCHK; last accessed 28/10/22)

We included all articles published in PS from 2003 - the first year that articles in html

format were available, which is more suitable for text mining than pdf files - up to October

2022 (the latest complete issue at the time of data collection), excluding retractions, errata,

and editorials. PS articles up until 2013 had already been downloaded in previous research

(Nuijten et al., 2016).

The second journal that implemented statcheck in their peer review process, JESP,

announced the new policy in August 2017 (*JESP Piloting the Use of Statcheck*, 2017). Their

policy states:

"**For all manuscripts that are deemed to fit within the Aims and Scope of the**

journal , the editorial team will be using statcheck as part of their initial triage of

manuscripts. *For any manuscripts found to have important discrepancies in reporting, we*

will ask authors to resolve these in the manuscript before they can be sent on for further

review. The pilot is intended to help editors and authors to work together to decrease the

number of errors in published articles in the journal.

Before submitting, authors are invited to run a HTML or PDF version of their

APA-formatted manuscript through statcheck prior to submitting their manuscripts,

via this link: http://statcheck.io/. This will be the same portal that the JESP Editorial Team will

be using."

(emphasis added;

https://www.elsevier.com/journals/journal-of-experimental-social-psychology/0022-1031/guid

e-for-authors; last accessed 29/10/22)

We included all articles published in JESP between 2003 and November 2022 (the latest complete issue at the time of data collection), excluding retractions, errata, and editorials.

Any decrease in statistical reporting inconsistencies in these journals after implementation of statcheck could in principle also be due to other factors (e.g., an increased awareness of reporting errors and/or statcheck in general). To control for this (at least in part), we also include two matched comparison journals that did not implement statcheck.

We looked for journals to match with PS and JESP that were similar in content and impact, and were likely to contain results reported in APA style to facilitate the automated detection of statistical results. As a matched control for PS, we chose the *Journal of Experimental Psychology: General* (JEPG). Both journals focus on general psychology and often publish experimental designs. As a matched control for JESP, we chose the *Journal of Personality and Social Psychology* (JPSP), but only articles from the subsection Attitudes and Social Cognition to ensure a closer match with the articles in JESP.

From both JEPG and JPSP, we included articles from the same years as PS and JESP: 2003 - October 2022 (in both cases, the latest complete issue at the time of data collection). Here, too, we excluded retractions, corrections, and editorials. Thirty-nine articles from JEPG did not seem to be available in html format, so these were excluded[1]. JEPG and JPSP articles up until 2013 had already been downloaded in previous research (Nuijten et al., 2016).

We used regular expressions to extract the submission date of each article to classify them as submitted before (T1) or after (T2) statcheck was implemented in the journal or its matched control journal (July 1st 2016 for PS and JEPG, and August 1st 2017 for JESP and JPSP).

Detecting Statistical Reporting Inconsistencies with *Statcheck*

---

[1] For a list of titles, see the metadata file at https://osf.io/q9xuf.

6

We used the R package *statcheck*, version 1.4.0-beta.7 (Nuijten & Epskamp, 2023), to automatically detect statistical reporting inconsistencies in the included articles.[2] *Statcheck*'s algorithm works in roughly four steps:

1. *Statcheck* first converts an article in pdf or html format to plain text.

2. Next, *statcheck* uses "regular expressions" to search for specific patterns of letters, numbers, and symbols that signal a NHST result. Specifically, statcheck can detect results of t-tests, F-tests, Z-tests, χ2-tests, correlation tests, and Q-tests, as long as they are reported completely (i.e., test statistic, degrees of freedom if applicable, and p-value), in text (*statcheck* usually misses results reported in tables), and in APA style (American Psychological Association, 2019).

3. Third, *statcheck* uses the reported test statistic and degrees of freedom to recalculate the (two-tailed) p-value.

4. In the final step, *statcheck* compares the reported and recalculated p-value. If these two values do not match, *statcheck* labels the result as an **inconsistency**. In cases where the reported p-value is statistically significant (assuming α = .05) and the recalculated one is not, or vice versa, *statcheck* labels the result as a **decision inconsistency**.

Statcheck takes into account one-tailed p-values (that are half the size of what *statcheck* would expect by default) that are explicitly identified as such. Specifically, if the word "one-tailed", "one-sided", or "directional" appears anywhere in the article, *and* the result would be internally consistent if the p-value was one-sided, *statcheck* treats it as such and counts it as correct.

Statcheck also takes correct rounding of the test statistic into account. Consider for instance the result $t(48) = 1.43$, $p = .158$. Recalculation would give a *p*-value of .159, not

---

[2] Note that in the preregistration we indicated we would use version 1.4.0-beta.4. We improved statcheck's performance for a couple of articles that contained unusual variations of statistical reporting. The updates made sure that statcheck would not throw an error when scanning these articles, but the updates did not affect the overall detection or error-flagging rate. For details on the updates, see https://github.com/MicheleNuijten/statcheck/releases.

.158, seemingly an inconsistency. However, the true *t*-value could lie in the interval [1.425, 1.435), with *p*-values ranging from .158 to .161. To take this into account, statcheck will count any *p*-value within this range as consistent.

Finally, *statcheck* counts $p = .000$ and $p < .000$ as inconsistent. A *p*-value of exactly zero is mathematically impossible, so the APA manual advises to report very small *p*-values as $p < .001$.

We estimate that *statcheck* detects roughly 60% of all reported null hypothesis significance test results (Nuijten et al., 2016). The results it does not detect are usually reported in tables instead of full text or the results are not reported in APA style.

We systematically assessed the validity of earlier versions of *statcheck* in classifying inconsistencies in previous research and concluded it is high (Nuijten et al., 2016, 2017). The interrater reliability between manual coding and statcheck was .76 for inconsistencies and .89 for decision inconsistencies. Furthermore, *statcheck*'s sensitivity (true positive rate) and specificity (true negative rate) are high: between 85.3% and 100%, and between 96.0% and 100%, respectively, depending on the assumptions and settings (e.g., the one-tailed test detection as described above, increases specificity). The overall accuracy of statcheck ranged from 96.2% to 99.9% (Nuijten, Van Assen, et al., 2017). Solving some issues in the previous *statcheck* versions has likely made its accuracy even higher.

Since the implementation of the *statcheck* policy in PS (July 1st 2016) and JESP (August 1st 2017), the *statcheck* web app has depended on different versions of the *statcheck* R package: version 1.0.1 (released Feb 4 2015), version 1.2.2 (released Aug 18 2016), and version 1.3.0 (released May 4 2018). For the current study, we have used the latest version of statcheck: version 1.4.0-beta.7. The subsequent versions became increasingly accurate in detecting statistical results and classifying them as consistent or not. Some of the most notable updates that improved the detection rate of statistical results and/or the classification of inconsistencies are: improvements in detecting χ^2 tests and minus signs, improvements in the detection and recalculation of one-tailed tests, the addition of the

Q-test for heterogeneity in meta-analyses, and improvements in determining correct rounding of test statistics.[3]

Results

Descriptives

Table 1 describes the number of available articles per journal, categorized by journal type (*statcheck* or matched control). For our analyses, it is important that we are able to determine whether an article was submitted before or after *statcheck* was implemented in the (matched) journal's peer review process. We therefore report the number of articles we could download, and the number of articles from which we could extract the date submitted/received. We were able to extract submission dates from almost all articles in our sample (97.4% - 99.6%, depending on the journal).

Table 1.

Number of downloaded articles and number of articles where we could automatically extract information about the date the article was submitted/received.

		# downloads	# articles with date submitted/received	
			N	%
Journals using *statcheck*	**PS**	3851	3752	97.4
	JESP	2557	2534	99.1
	Total	**6408**	**6286**	**98.1**
Matched control journals	**JEPG**	1826	1819	99.6
	JPSP	716	709	99.0
	Total	**2542**	**2528**	**99.4**

---

[3] Detailed information about all updates can be found here:
https://github.com/MicheleNuijten/statcheck/blob/master/NEWS.md

Table 2 shows the most important descriptive statistics, split up by journal type (*statcheck* or control) and period: before (T1) and after (T2) statcheck was implemented. To get a general sense of the inconsistency rates in the different journal types and periods, we can look at the mean percentage of (decision) inconsistencies in articles with NHST results. This statistic is calculated as follows: say that *statcheck* detects 10 NHST results in a given article, of which 2 are inconsistent. The percentage of inconsistencies in this article is then 20%. We calculated this percentage for each article with NHST results and averaged these percentages per journal type and period. More detailed descriptive results, split up per journal, can be found in Table 3.

We found that the mean inconsistency percentage decreased more steeply in the *statcheck* journals (8.9 - 4.4 = 4.5 percentage points) than in the control journals (7.4 - 6.4 = 1.0 percentage point). We see a similar pattern in the decision inconsistencies: 1.2 - 0.3 = 0.9 percentage points in the *statcheck* journals and 1.0 - 0.8 = 0.2 percentage points in the control journals. The same pattern can be seen in Figure 1A and 1B, that show the distributions of these percentages in violin plots, where in panel B the y-axis is truncated to zoom in on the bulk of the observed percentages. Similarly, Figure 2 shows the change in the mean inconsistency percentages per type of journal and period, and split up for the journal pairs separately.

Note that these descriptives are mainly reported for illustrative purposes: because the number of NHST results differs greatly per article, it is hard to directly interpret these percentages. The same holds for Figures 1 and 2: it is useful to look at these average percentages to get a general image of the observed patterns. We appropriately test differences in the next section.

Table 2.

Descriptive statistics split up for journals using statcheck *and their matched control journals, before (T1) and after (T2)* statcheck *was implemented.*
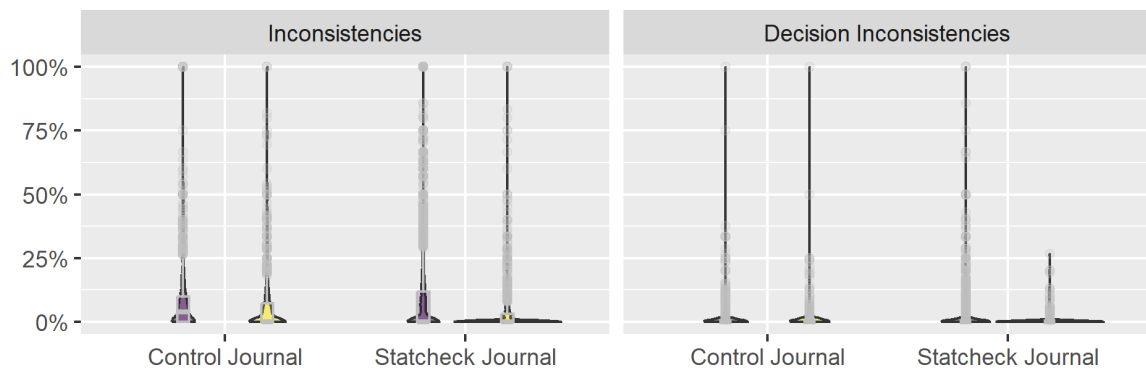
	Time period	# articles[a]	# articles with NHST results (%)	Median # of NHST results per article with NHST results	Mean % of inconsistencies per article with NHST results	Mean % of decision inconsistencies per article with NHST results
Journals using *statcheck*	**T1**	4941	4012 (81.2)	11	8.9	1.2
	T2	1345	876 (65.1)[b]	17	4.4	0.3
Matched control journals	**T1**	1613	1445 (89.6)	24	7.4	1.0
	T2	915	763 (83.4)	19	6.4	0.8

[a] Total number of downloaded articles from which we could extract the date submitted/received. For details, see Table 1.

[b] Note that this percentage is considerably lower than in the other three cells. This is mainly caused by the use of a specific style of spaces in statistical results in 2019 JESP articles that statcheck does not recognize. In this year, statcheck detected NHST results in only 2.4% of the articles. Without JESP 2019 articles, statcheck would detect NHST results in 71.4% of the articles with a date. As a sensitivity analysis, we also computed the mean percentage of (decision) inconsistencies per article with NHST results, and the only change was that without 2019 JESP articles the mean % of inconsistencies in *statcheck* journals in T2 decreased from 4.4% to 4.3%.
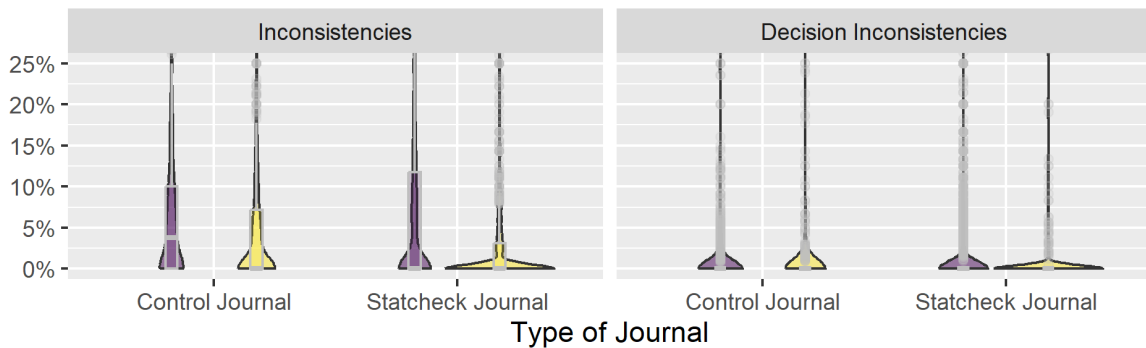
Figure 1. Distribution of the percentages of inconsistencies and decision inconsistencies per article with NHST results, per type of journal and period. Panel A shows the full distributions, in panel B, the y-axis is truncated at 25% to improve readability. The number of articles with NHST results from the control journals before and after statcheck implementation were N = 1445 and N = 763, respectively, and for the *statcheck* journals they were N = 4012 and N = 876, respectively.

Figure 2. Illustrative representation of the mean % of inconsistencies (top row) and decision inconsistencies (bottom row) per article with NHST results, before and after *statcheck* implementation. Solid lines represent *statcheck* journals, dashed lines represent matched control journals. The first column depicts mean percentages for the journals combined, the second and third column show the results for the journal pairs separately. Preregistered multilevel logistic regressions showed that the interaction between journal type and period was statistically significant when predicting inconsistencies and decision inconsistencies. Exploratory follow-up analyses of the journal pairs separately only found a significant interaction when predicting inconsistencies in PS/JEPG.

Confirmatory Analyses

Following our preregistration, we tested our main hypotheses using two multilevel logistic models:

$$Logit[Inconsistency] = b_0 + b_1 Period_i + b_2 statcheckJournal_i + b_3 Period_i * statcheckJournal_i + \theta_i,$$

$$Logit[DecisionInconsistency] = b_0 + b_1 Period_i + b_2 statcheckJournal_i + b_3 Period_i * statcheckJournal_i + \theta_i$$

,

where subscript *i* indicates article, *statcheckJournal* indicates if the journal implemented *statcheck* in peer review [no = 0 [JEPG and JPSP] and yes = 1 [PS and JESP]), *Period* is the period in which an article is submitted (before [0] or after [1] statcheck was implemented in peer review), and θ_i is a random effect on the intercept b_0. We include a random intercept because the statistical results are nested within the article, which means there can be dependency in the inconsistencies within the same article.

We hypothesized that in both models the coefficient b_3 is smaller than zero, which would indicate a steeper decrease in (decision) inconsistencies in "statcheck" journals. We maintained an α of .05 and used two-tailed tests in line with our preregistration.

When predicting the inconsistencies, we found a significant interaction effect of *Period * statcheckJournal* in the predicted direction, b_3 = -0.73, SE = 0.11, 95% CI = [-0.95; -0.51], Z = -6.57, *p* < .001. Also when predicting decision inconsistencies, we found a significant interaction effect of *Period * statcheckJournal* in the hypothesized direction, b_3 = -0.77, SE = 0.36, 95% CI = [-1.47; -0.07], Z = -2.15, *p* = .031.

These findings indicate that the prevalence of inconsistencies and decision inconsistencies decreased more steeply in *statcheck* journals than in control journals. This finding is in line with the notion that implementing *statcheck* in the peer review process could decrease the prevalence of reporting inconsistencies.

Table 3.

Descriptive statistics split up for journals using statcheck *and their matched control journals, before (T1) and after (T2) statcheck was implemented.*

		Time period	# articles[a]	# articles with NHST results (%)		Median # of NHST results per article with NHST results	Mean % of inconsistencies per article with NHST results	Mean % of decision inconsistencies per article with NHST results
Journals using *statcheck*	PS	T1	2943	2215	(75.3)	9	9.2	1.3
		T2	809	591	(73.1)	14	4.0	0.4
	JESP	T1	1998	1797	(89.9)	15	8.5	1.1
		T2	536	285	(53.2)	26	5.3	0.2
	Total	T1	4941	4012	(81.2)	11	8.9	1.2
		T2	1345	876	(65.1)	17	4.4	0.3
Matched control journals	JEPG	T1	1024	879	(85.8)	19	7.7	1.1
		T2	795	653	(82.1)	18	6.8	0.9
	JPSP	T1	589	566	(96.1)	34	7.0	1.0
		T2	120	110	(91.7)	38.5	3.9	0.4
	Total	T1	1613	1445	(89.6)	24	7.4	1.0
		T2	915	763	(83.4)	19	6.4	0.8

[a] Total number of downloaded articles from which we could extract the date submitted/received. For details, see Table 1.

Discussion

In this retrospective observational study, we compared the prevalence of statistical reporting inconsistencies in two journals that implemented *statcheck* in their peer review process and two matched control journals, before and after the *statcheck* implementation. We found a steeper decline of both inconsistencies (b_3 = -0.73, 95% CI = [-0.95; -0.51]) and decision inconsistencies (b_3 = -0.77, 95% CI = [-1.47; -0.07]) in the *statcheck* journals than in the control journals, which is in line with the notion that implementing *statcheck* in the peer review process can be effective in avoiding reporting errors in published articles.

An important limitation is that our study was observational. We did not randomly assign journals or manuscripts to be checked by *statcheck* or not, which means there can be selection effects and other potential confounding factors that could explain the observed effect. For example, it is possible that we see a steeper decline in reporting errors in the *statcheck* journals because the implementation of *statcheck* signaled a commitment to improved reporting practices, which inspired the more conscientious authors to submit to those journals. Or conversely, authors with a tendency towards less diligent reporting may have been deterred from submitting to *statcheck* journals due to the increased likelihood of errors being detected by *statcheck*. A related limitation is that we picked two comparison journals based on relatively subjective criteria based on similarity in subfields and impact. Alternative comparison journals may also have been suitable, and could potentially have shown other results. It would be interesting in future work to assess declines in reporting errors across a wider range of journals.

We note that the descriptive statistics showed that both *statcheck* journals still published articles with statistical reporting inconsistencies after *statcheck* was implemented (see Table 3). On the face of it, this seems surprising: if both journals require a "clean" *statcheck* report for publication, what could be explanations for the remaining statistical inconsistencies? Three potential explanations are that 1) the remaining inconsistencies would not have been picked up by *statcheck* version 1.3.0 (the latest published version at the time of writing) or older versions, as compared to version 1.4.0-beta.7 that we have used

in this study, 2) the remaining inconsistencies are intentional and can be explained by, e.g., the use of statistical corrections, use of one-tailed tests that are not explicitly identified in the way *statcheck* expects, or *p*-values that were deliberately reported as *p* = .000, or 3) some remaining errors slipping through the review process. To better understand the remaining inconsistencies, we plan to analyze the full text of a subsample of articles in more detail in future work. We will report our findings and additional exploratory analyses in the forthcoming, more detailed manuscript.

Even when taking the limitations into account, we would still tentatively recommend journals that adhere to APA reporting guidelines to consider using *statcheck* in their peer review process. Human peer reviewers seem to often overlook statistical reporting inconsistencies (judging from the high prevalence of inconsistencies in the literature and recent experimental work; Augusteijn et al., 2023), so including automated tools like *statcheck* in the peer review process might be a viable solution.

Even though we did not assess the direct effect of the use of *statcheck* during peer review, we argue that any mechanism that lowers the currently high prevalence of reporting errors in the literature is preferred. Given its ease of use and provided that *statcheck* is used carefully and in a collaborative manner by reviewers, authors, and editors, we see few potential downsides in implementing statcheck more widely to improve the quality of reporting of statistical results. We do caution editors not to rely solely on the results of *statcheck* (or any automated tool, for that matter) when deciding to accept or reject an article: software can be useful in reducing workload or human errors, but is not free of its own pitfalls. That said, we think *statcheck* can be a quick and easy way to help journals avoid statistical reporting inconsistencies in their articles and increase their overall quality and robustness.

References

Augusteijn, H. E. M., Wicherts, J., Sijtsma, K., & Assen, M. A. L. M. van. (2023). *Quality assessment of scientific manuscripts in peer review and education*. OSF Preprints.

https://doi.org/10.31219/osf.io/7dc6a

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N., & Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, *18*(3), 230–232.

JESP piloting the use of statcheck. (2017, August). https://www.journals.elsevier.com/journal-of-experimental-social-psychology/news/journals.elsevier.com/journal-of-experimental-social-psychology/news/jesp-piloting-the-use-of-statcheck

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Nuijten, M. B. (2021). Assessing and improving robustness of psychological research findings in four steps. In W. O'Donohue, A. Masuda, & S. O. Lilienfeld (Eds.), *Clinical psychology and questionable research practices*. Springer. https://psyarxiv.com/a4bu2

Nuijten, M. B., Bakker, M., Maassen, E., & Wicherts, J. M. (2018). Verify original results through reanalysis before replicating. *Behavioral and Brain Sciences*, *41*, e143. https://doi.org/10.1017/S0140525X18000791

Nuijten, M. B., & Epskamp, S. (2023). *statcheck: Extract statistics from articles and recompute p-values* (1.4.0-beta.7) [R]. https://github.com/MicheleNuijten/statcheck/releases/tag/v1.4.0-beta.7

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. https://doi.org/10.3758/s13428-015-0664-2

Nuijten, M. B., Van Assen, M. A. L. M., Hartgerink, C. H. J., Epskamp, S., & Wicherts, J.

(2017). *The validity of the tool "statcheck" in discovering statistical reporting inconsistencies*. PsyArXiv. https://doi.org/10.31234/osf.io/tcxaj