

Implementing statcheck during peer review is related to a steep decline in statistical reporting inconsistencies

Michèle B. Nuijten¹ & Jelte M. Wicherts¹

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

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

Abstract

We investigated whether statistical reporting inconsistencies could be avoided if journals implement the tool *statcheck* in the peer review process. In a preregistered pretest-posttest quasi-experiment covering over 7000 articles and over 147,000 extracted statistics, we compared the prevalence of reported *p*-values that were inconsistent with their degrees of freedom and test statistic in two journals that implemented *statcheck* in their peer review process (*Psychological Science* and *Journal of Experimental and Social Psychology*) and 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 preliminary 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.

Keywords: statistical reporting inconsistencies, peer review, *statcheck*, meta-science

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.

Methods

Disclosures

The hypothesis and study protocol (rationale, methods, and analysis plan, including a power analysis) were preregistered before data collection on recent articles¹ 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 how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. 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. This study was approved by the Ethics Review Board of the Tilburg School of Social and Behavioral Sciences on November 18th, 2022 (approval number: TSB_RP773).

Design

In this study, 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. This makes our design a pretest-posttest quasi-experiment.

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**

¹ The articles published up until 2013 were already downloaded and scraped with an older version of *statcheck* in Nuijten et al. (2016).

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://web.archive.org/web/20221006044320/http://www.psychologicalscience.org/publications/psychological_science/ps-submissions#STATCHK; snapshot at 14/10/2022)

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://web.archive.org/web/20221205222546/https://www.elsevier.com/journals/journal-of-experimental-social-psychology/0022-1031/guide-for-authors>; snapshot at 05/12/2022)

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 different volumes and publication years of JEPG did not seem to be available in html format, so these were excluded.² Given the large overall sample size, we deem it unlikely that these missing cases would influence our conclusions. 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

² For a list of titles, volume numbers, and publication years, see <https://osf.io/mukxc>.

matched control journal (July 1st 2016 for PS and JEPG, and August 1st 2017 for JESP and JPSP).

Detecting Statistical Reporting Inconsistencies with *Statcheck*

We used the R package *statcheck*, version 1.4.1-beta.2 (Nuijten & Epskamp, 2023), to automatically detect statistical reporting inconsistencies in the included articles.³

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, χ^2 -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 $\alpha = .05$) 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

³ Note that in the preregistration we indicated we would use an earlier *statcheck* version, version 1.4.0-beta.4. The new version recognizes a specific type of spacing that JESP articles used in 2019 articles. We also 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>.

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 .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 consider 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 several issues in the previous *statcheck* versions⁴ has likely further improved its accuracy.

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 1.4.1-beta.2. The subsequent versions became

⁴ See <https://github.com/MicheleNuijten/statcheck/releases>.

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.⁵

Results

Descriptives

Table 1 describes the number of available articles per journal, categorized by journal type (*statcheck* or matched control). For our analyses, we had 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.⁶ We were able to extract submission dates from almost all articles in our sample (97.4% - 99.6%, depending on the journal).

⁵ Detailed information about all updates can be found here: <https://github.com/MicheleNuijten/statcheck/blob/master/NEWS.md>

⁶ We manually verified the extracted dates in a random sample of 50 articles and after improving our code to handle different date formats, we reached 100% accuracy.

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 <i>statcheck</i>	PS	3,851	3,752	97.4
	JESP	2,557	2,534	99.1
	Total	6,408	6,286	98.1
Matched control journals	JEPG	1,826	1,819	99.6
	JPSP	716	709	99.0
	Total	2,542	2,528	99.4
Total		8,950	8,814	98.5

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. In total, we extracted 147,784 NHST results from 7,314 articles. Of these results, 10,160 were inconsistent (6.9%) and 1,226 were a decision inconsistency (0.8%). 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.

STATCHECK LINKED TO DECLINE IN STATISTICAL REPORTING ERRORS

We found that the mean inconsistency percentage decreased more steeply in the *statcheck* journals ($8.8 - 4.3 = 4.5$ 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. We note that the distributions of these percentages are highly skewed: in most articles, only a small percentage of reported results is inconsistent, but there are a few articles in which up to 100% of reported results are inconsistent. This skewness can be seen in Figure 1A and 1B,, where in panel B the y-axis is truncated to zoom in on the bulk of the observed percentages. To more clearly illustrate the difference in inconsistency percentages between journals, 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 (which in part also explains the skewness of the distributions), it is hard to directly interpret these percentages. The same holds for Figures 1 and 2: these average percentages are mainly useful to get a general image of the observed patterns. We appropriately test differences in the next section.

STATCHECK LINKED TO DECLINE IN STATISTICAL REPORTING ERRORS

Table 2

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

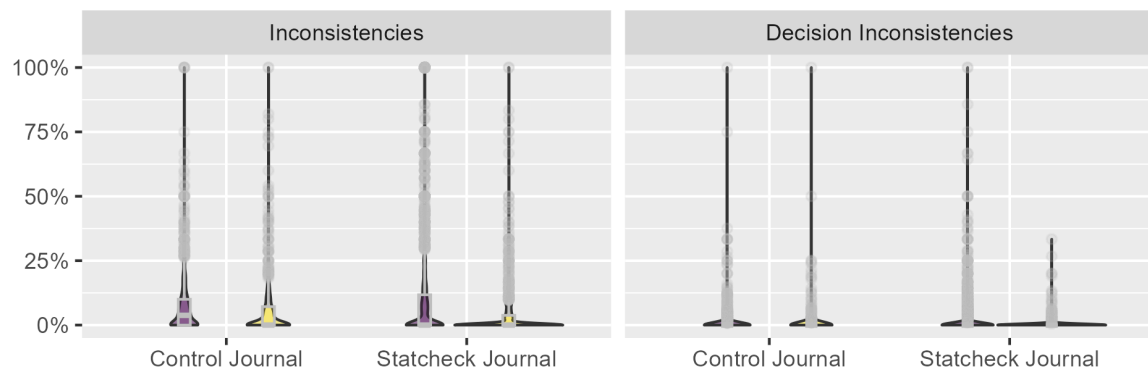
Time period		# articles ^a	# articles with NHST results (%)		Total # of extracted NHST results	Median # of NHST results per article with NHST results	Total # of inconsistencies	Total # of decision inconsistencies	Mean % of inconsistencies per article with NHST results	Mean % of decision inconsistencies per article with NHST results
Journals using statcheck	T1	4,950	4,053	(81.9) ^b	59,099	11	4,959	631	8.8	1.2
	T2	1,336	1,053	(78.8) ^b	24,983	18	971	94	4.3	0.3
Matched control journals	T1	1,613	1,445	(89.6)	43,210	24	3,046	379	7.4	1.0
	T2	915	763	(83.4)	20,492	19	1,184	122	6.4	0.8
Total		8,814	7,314	(83.0)	147,784	14	10,160	1,226	7.7	1.0

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

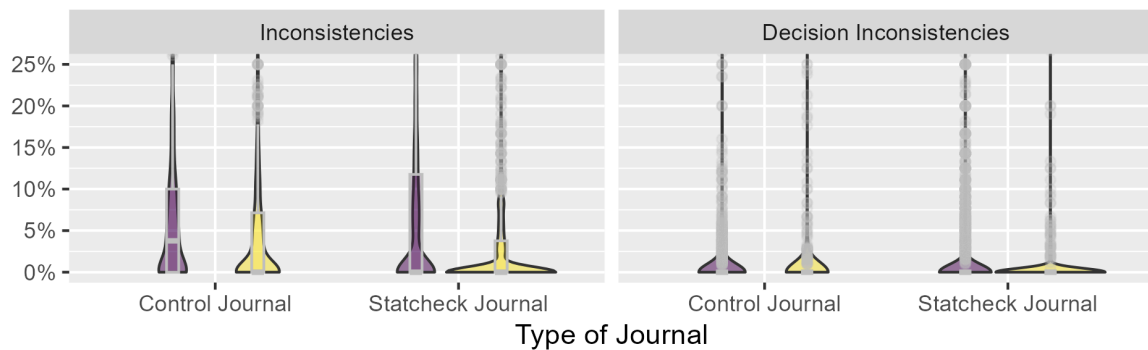
^b Note that in a previous preprint of this paper (<https://doi.org/10.31234/osf.io/bxau9>; Version 1) we found fewer articles with NHST results in journals using *statcheck* (4012 in T1 and 876 in T2). This was mainly caused by the use of a specific style of spaces in statistical results in 2019 JESP articles that *statcheck* version 1.4.0-beta.7, which we used at that time, did not recognize. In the current *statcheck* version 1.4.1-beta.2, this issue was fixed, hence the higher detection rate of NHST results in *statcheck* journals.

Distribution of % (decision) inconsistencies per article with NHST results

A.



B. (truncated y-axis)





Time Period  Before statcheck implementation  After statcheck implementation

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 $N = 1,445$ and $N = 763$, respectively, and for the *statcheck* journals they were $N = 4,053$ and $N = 1,053$, respectively.

STATCHECK LINKED TO DECLINE IN STATISTICAL REPORTING ERRORS

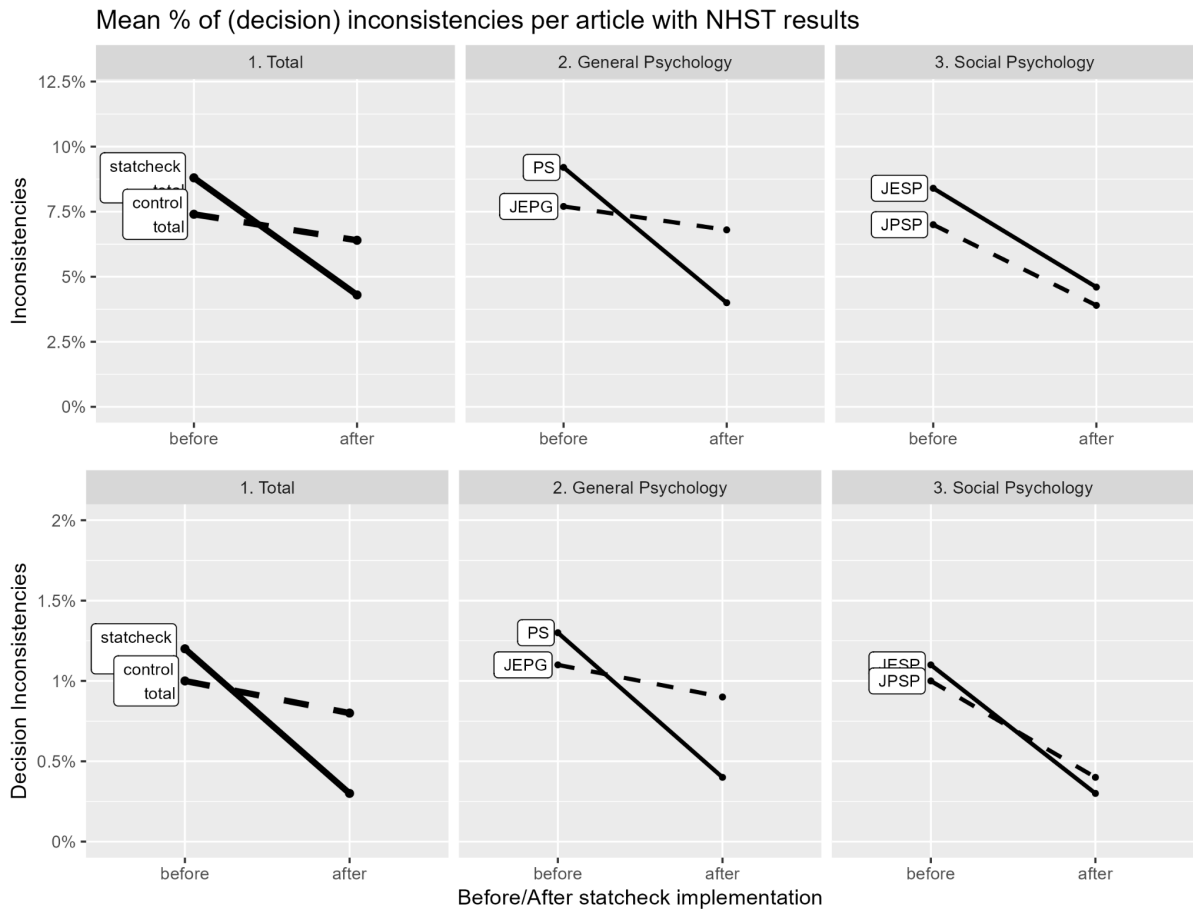


Figure 2. Illustrative representation of the mean % of inconsistencies (top row) and decision inconsistencies (bottom row) in articles with NHST results, before and after *statcheck* implementation. Solid lines represent *statcheck* journals, dashed lines represent matched control journals. The first 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.

STATCHECK LINKED TO DECLINE IN STATISTICAL REPORTING ERRORS

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 (%)	Total # of extracted NHST results	Median # of NHST results per article with NHST results	Total # of inconsistencies	Total # of decision inconsistencies	Mean % of inconsistencies per article with NHST results	Mean % of decision inconsistencies per article with NHST results
Journals using <i>statcheck</i>	PS	T1	2,943	2,215	(75.3)	24,281	9	2,077	295	9.2	1.3
		T2	809	591	(73.1)	10,632	14	392	50	4.0	0.4
	JESP	T1	2,007	1,838	(91.6)	34,818	15	2,882	336	8.4	1.1
		T2	527	462	(87.7)	14,351	26	579	44	4.6	0.3
	Total	T1	4,950	4,053	(81.9)	59,099	11	4,959	631	8.8	1.2
		T2	1,336	1,053	(78.8)	24,983	18	971	94	4.3	0.3
Matched control journals	JEPG	T1	1,024	879	(85.8)	21,684	19	1,536	171	7.7	1.1
		T2	795	653	(82.1)	15,980	18	1,020	108	6.8	0.9
	JPSP	T1	589	566	(96.1)	21,526	34	1,510	208	7.0	1.0
		T2	120	110	(91.7)	4,512	38.5	164	14	3.9	0.4
	Total	T1	1,613	1,445	(89.6)	43,210	24	3,046	379	7.4	1.0
		T2	915	763	(83.4)	20,492	19	1,184	122	6.4	0.8
Total			8,814	7,314	(83.0)	147,784	14	10,160	1,226	7.7	1.0

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

Confirmatory Analyses

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

$$\text{Logit}[\text{Inconsistency}] = b_0 + b_1 \text{Period}_i + b_2 \text{statcheckJournal}_i + b_3 \text{Period}_i * \text{statcheckJournal}_i + \theta_i,$$

$$\text{Logit}[\text{DecisionInconsistency}] = b_0 + b_1 \text{Period}_i + b_2 \text{statcheckJournal}_i + b_3 \text{Period}_i * \text{statcheckJournal}_i + \theta_i,$$

,

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. In a preregistered power analysis, we estimated that we have a power of .80 if the true b_3 lies between -.22 and -.44. This is equivalent to a reduction in inconsistencies of 1 - 2 percentage points. We expect a much larger reduction in the probability of an inconsistency, since the intervention is so straightforward. Furthermore, we were able to include more articles in our final sample than we had anticipated, increasing power. As a conservative estimate, we conclude that if the population effect size is smaller (in absolute sense) than -.22, we might not have enough power to detect it. For details, see our preregistration at <https://osf.io/hqt49>.

When predicting the inconsistencies, we found a significant interaction effect of *Period * statcheckJournal* in the predicted direction, $b_3 = -0.71$, $SE = 0.11$, 95% CI = [-0.92; -0.51], $Z = -6.73$, $p < .001$. This corresponds to an odds of $\exp(-.71) = 0.49$, 95% CI = [0.40; 0.60] and a probability of $.49/(1+.49) = .33$, 95% CI = [.28; .38]. Also when predicting

decision inconsistencies, we found a significant interaction effect of *Period* *

statcheckJournal in the predicted direction, $b_3 = -0.82$, $SE = 0.35$, 95% CI = [-1.51; -0.14], $Z = -2.37$, $p = .018$. This corresponds to an odds of 0.44, 95% CI = [0.22; 0.87] and a probability of .31, 95% CI = [.18; .47]. See Table 4 for detailed results.

As a simplified illustration of the coefficients in Table 4, consider an imaginary article with 100 NHST results. This article would show a decrease from 4 to 3 inconsistencies in a control journal, and a decrease from 5 to 2 inconsistencies in a *statcheck* journal. In this same imaginary article, the number of decision inconsistencies would decrease from .03 to .02 in a control journal, and from .04 to .01 in a *statcheck* journal. Note, however, that these inconsistency rates are lower than the ones reported in our descriptive statistics. This discrepancy is due to the estimation of the b-coefficients, which takes into account the random intercept, resulting in a lower probability of an inconsistency than observed directly in the data.

Table 4

Results of estimating the multilevel logistic regression models that predict whether a statistical result is an inconsistency (Model 1) or decision inconsistency (Model 2) based on period and type of journal that the article was published in, with random intercepts for articles.

Effect	Model 1: Predicting Inconsistencies					Model 2: Predicting Decision Inconsistencies				
	<i>Estimate</i>	<i>SE</i>	<i>Z</i>	<i>p</i>	95% CI	<i>Estimate</i>	<i>SE</i>	<i>Z</i>	<i>p</i>	95% CI
Fixed effects										
Intercept	-3.120	0.046	-68.368	<.001	[-3.209; -3.030]	-8.010	0.242	-33.165	<.001	[-8.484; -7.537]
Period	-0.275	0.080	-3.458	.001	[-0.431; -0.119]	-0.418	0.252	-1.662	0.097	[-0.911; 0.075]
statcheckJournal	0.125	0.053	2.347	.019	[0.021; 0.230]	0.229	0.162	1.415	0.157	[-0.088; 0.545]
Period * statcheckJournal	-0.713	0.106	-6.728	<.001	[-0.921; -0.505]	0.824	0.348	-2.366	0.018	[-1.506; -0.141]
Random effects										
Std. deviation of the intercept	1.289					3.840				

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, but note that causality cannot be established and alternative explanations are possible, due to the nature of the design. We list potential alternative explanations for these results in the Discussion.

Exploratory Analyses

Bayesian Analysis

Next to conducting a frequentist hypothesis test, we also computed Bayesian hypothesis tests. We computed approximated adjusted fractional Bayes factors (Gu et al., 2018) using the default implementation in the R package BFpack (Mulder et al., 2021; Version 1.0.0). The approximated adjusted fractional Bayes factor uses a minimal fraction of the available data to train a non-informative normally distributed prior and approximate the marginal likelihood of the tested hypotheses. We compared two models with each other: $Period * statcheckJournal < 0$ (which would indicate that inconsistencies decreased stronger in *statcheck* journals than in non-*statcheck* journals) vs. its unconstrained complement.

Given our data, we found that the model predicting inconsistencies where $Period * statcheckJournal < 0$ was 1.16^{e11} times more likely than a model where $Period * statcheckJournal$ was not < 0 ($BF_{10} = 1.16^{e11}$, posterior probability = 1). When predicting decision inconsistencies, we found that the model where $Period * statcheckJournal < 0$ was 110.2 times more likely than its complement ($BF_{10} = 110.2$, posterior probability = .991). These Bayesian hypothesis tests indicated very strong and strong evidence, respectively (Kass & Raftery, 1995), in favor of our hypothesis that both inconsistencies and decision inconsistencies decreased more steeply in *statcheck* journals after *statcheck* implementation than in non-*statcheck* journals.

Inconsistency Rates in Journal Pairs Separately

Exploratively, we also looked at the inconsistency rates in the journal pairs separately. Table 3 and Figure 2 show the mean percentages of inconsistencies and decision inconsistencies in an article per journal and time. We fitted the multilevel logistic models explained above for the two journal pairs separately. When predicting the inconsistencies, we did find a statistically significant interaction term for the pair PS/JEPG ($b_3 = -0.92$, 95% CI = [-1.20; -0.64], $Z = -6.42$, $p < .001$), but not for JESP/JPSP ($b_3 = -0.12$, 95% CI = [-0.51; 0.28], $Z = -0.58$, $p = 0.561$). When predicting decision inconsistencies, we did not find statistically significant effects for either journal pair: PS/JEPG ($b_3 = -0.80$, 95% CI = [-1.84; 0.24], $Z = -1.51$, $p = 0.130$) and JESP/JPSP ($b_3 = -0.37$, 95% CI = [-1.38; 0.63], $Z = -0.73$, $p = 0.466$). These exploratory results could indicate that the overall significant interaction term of Period*statcheckJournal in the confirmatory results may be driven by a strong effect of implementing *statcheck* in PS, specifically. Note, however, that we did not formally test this three-way interaction to assess if there is a significant difference in interaction effects between the pairs PS/JEPG and JESP/JPSP. Such an analysis would likely be underpowered and because of its post-hoc nature, any resulting p-values would be hard to interpret.

Trends in the prevalence of reporting inconsistencies over time

Following Nuijten et al. (2016), we considered trends over time in the prevalence of (decision) inconsistencies. Figure 3 shows the percentage of articles with NHST results that contained at least one inconsistency (pink line) or decision inconsistency (blue line) over time, split up per journal. We did not do a regression analysis to estimate the overall trends, because in the *statcheck* journals we would not expect a linear decrease, but a difference in trends before and after *statcheck* implementation. Visual inspection of the trends shows a clear drop in the prevalence of articles with inconsistencies in PS after *statcheck* implementation: before, roughly 40% of articles contained at least one inconsistency, but after *statcheck* implementation, this halved to about 20% of articles. Other patterns in this graph do not jump out as much, but overall the results seem to be in line with Nuijten et al. (2016), who found an overall decrease in statistical reporting inconsistencies and decision

inconsistencies over time. After implementation of *statcheck* during review at PS, the prevalence of reporting inconsistencies visibly diminished further.

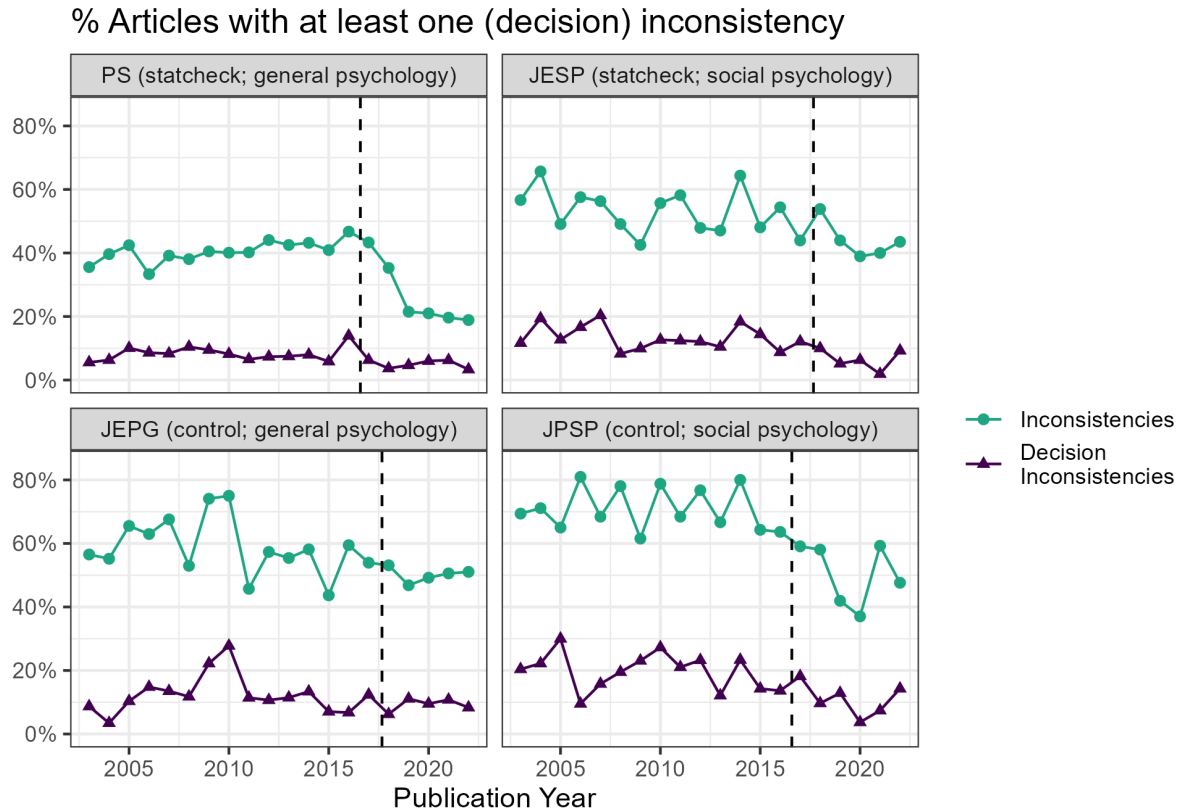


Figure 3. Percentage of articles with NHST results that contained at least one inconsistency (pink line) or decision inconsistency (blue line) over time, split up per journal. The vertical dotted line indicates the moment that *statcheck* was implemented in a journal or its counterpart.

Looking at the Remaining Inconsistencies in Detail

Notwithstanding the introduction of *statcheck* during peer review, journals continued to publish articles with statistical reporting inconsistencies (see Table 3). On the face of it, this seems surprising: if both journals require a “clean” *statcheck* report for publication, how could statistical inconsistencies remain? Four 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

1.4.1-beta.2 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 $p = .000$, 3) *statcheck* incorrectly flagged a correct result as an inconsistency, or 4) some remaining errors slipping through the review process, for example if an editorial team did not routinely (or correctly) follow through on their *statcheck* policy.

To better understand the remaining inconsistencies, we analyzed the full text of a subsample of articles in more detail. Specifically, we randomly selected ten articles that had at least one inconsistency from each of the two *statcheck* journals, published after *statcheck* was implemented and categorized likely reasons for remaining errors. We summarize the results in Table 5.

Table 5

Classification of likely reasons for remaining inconsistencies in a random subsample of twenty articles from the two statcheck journals, after statcheck was implemented.

Likely reason for remaining inconsistencies	# Inconsistencies		
	PS	JESP	Total
1. Version updates	0	0	0
<ul style="list-style-type: none"> None of the inconsistencies from the sampled papers were due to differences in <i>statcheck</i> versions 			
2. Statistical corrections	10	54	64
<ul style="list-style-type: none"> Corrections for multiple testing (incl. Bonferroni adjusted <i>p</i>-values) Greenhouse-Geisser adjusted <i>dfs</i> Unidentified one-tailed tests 			
3. Error <i>statcheck</i>	0	14	14
<ul style="list-style-type: none"> All from the same paper: last two digits of the extracted <i>p</i>-value were accidentally extracted from next row in table 			
4. Statistical reporting error slipped through review process	11	24	35
<ul style="list-style-type: none"> Rounding errors Mix-up of "$p < \dots$" and "$p = \dots$" Typos 			
Total	21	92	113

In total, these twenty papers contained 737 NHST results, of which 113 were an inconsistency (15.3%) and 8 were a decision inconsistency (1.1%). The majority of remaining inconsistencies seemed to be due to statistical corrections to either the p -value or the degrees of freedom (64 cases; 57%). In previous work, we have argued that such corrections can and should be reported in a way that does not render the full result inconsistent (Nuijten et al., 2017), but this is not a convention yet. Next, we classified 35 inconsistencies as actual statistical errors that seemed to have slipped through the review process. These included cases where a p -value appeared to be rounded incorrectly, where a “<” sign was reported, while a “=” would have been correct, or where digits were switched. Lastly, in 14 cases, all from the same paper, *statcheck* made an error in extracting the results from the paper: it added two additional digits from the next row of a table to the reported p -value, rendering it inconsistent. This illustrates that despite *statcheck*’s high performance in classifying inconsistencies (Nuijten et al., 2017), it is not perfect, and its results should not be acted upon blindly. We noted that none of the remaining inconsistencies could be attributed to *statcheck* version updates.

Discussion

In this pretest-posttest quasi-experiment, 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 ($b_3 = -0.71$, 95% CI = [-0.92; -0.51]) and decision inconsistencies ($b_3 = -0.82$, 95% CI = [-1.51; -0.14]) 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. Exploratory Bayesian hypothesis tests provided strong evidence in line with this notion.

In additional exploratory analyses, we observed a decrease in the prevalence of articles with (decision) inconsistencies across all included years in all but one journal (JEPG). This may indicate that without any interventions the prevalence of statistical

reporting inconsistencies may decrease somewhat over time, but our confirmatory results do give an initial, potential indication that certain interventions may reduce reporting errors a lot faster.

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*, which might be a good outcome for the journal at hand, but might also mean that reporting errors remain appearing elsewhere.

The implementation of *statcheck* in PS and JESP has also not occurred in a vacuum. In both cases, *statcheck* was implemented relatively shortly after a new editor-in-chief was installed. New editors often install new policies, which could affect the type of articles submitted to these journals and the way that submissions are handled. Indeed, around the same time that *statcheck* was introduced, both PS and JESP published editorials emphasizing the importance of open practices and replication (Giner-Sorolla, 2016; Lindsay, 2015). We cannot rule out that this emphasis on responsible research practices has affected the prevalence of reporting inconsistencies. However, the control journal JPSP also published a similar editorial promoting open practices around the same time (Cooper, 2016), and there the observed decline in reporting inconsistencies is less stable than in PS and JESP (see Figure 3).

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. Furthermore, several other journals besides PS and JESP have recently started to use

statcheck.⁷ In time, when a sufficient number of articles has been published in these journals after *statcheck*'s implementation, it would be interesting to assess declines in reporting inconsistencies across a wider range of journals.

Another limitation is that we only looked at reporting inconsistencies that could be picked up by *statcheck*. That means that inconsistencies in non-APA reported results, errors in data entry, or other types of statistical problems have not been detected. Since our initial goal was to have an indication whether implementing *statcheck* could be effective in decreasing "*statcheck*-type inconsistencies", this limitation has no direct bearing on our conclusion. However, for a deeper understanding of the different types of problems in statistical reporting, we would need richer data and much more checking of the many ways in which analyses and reporting of results can go awry.

A final, related limitation is that *statcheck* itself is not flawless. Based on exploratory, in-depth analyses of a subsample of articles where *statcheck* flagged inconsistencies, we found that *statcheck* wrongly extracted statistics from a table in one article, resulting in falsely flagged inconsistencies. Additionally, in this subsample, *statcheck* flagged quite some cases that turned out to be results that were corrected for multiple testing or violations of assumptions. Whether or not this is problematic is debatable, but we previously argued that such corrections can and should be reported in a way that does not render the full result inconsistent (Nuijten et al., 2017), to allow a reproducible report of applied corrections.

Even when taking the limitations of this study and of *statcheck* 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

⁷ We have no systematic overview, but examples are the journal *Child Development* that requires authors to run *statcheck* on their manuscript (<https://web.archive.org/web/20240325055348/https://www.srctd.org/research/journals/child-development/child-development-submission-guidelines#Revised>) and the journal *Consumer Research* that runs *statcheck* on submissions themselves (<https://web.archive.org/web/20230606191958/https://consumerresearcher.com/jcrs-data-policy-in-practice>).

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 and peer reviewers 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.

Author Contributions

Conceptualization: MN, JW

Methodology: MN, JW

Software: MN

Investigation: MN

Formal Analysis: MN

Writing - Original Draft: MN

Writing - Review & Editing: MN, JW

Conflicts of Interest

MN is the main developer of the tool *statcheck*. Both MN and JW have several publications where *statcheck* was used as the main method of data collection.

Acknowledgements

We thank Afra Kiliç and Tsz Keung Wong for their assistance in downloading the articles.

Funding

This work was funded by an ERC Consolidator grant (IMPROVE project, grant number 726361) and an NWO Veni Grant (grant number 11507).

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>
- Cooper, M. L. (2016). Editorial: Journal of Personality and Social Psychology. *Journal of Personality and Social Psychology*, 110(3), 431–434. <https://doi.org/10.1037/pspp0000033>
- 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.
- Giner-Sorolla, R. (2016). Approaching a fair deal for significance and other concerns. *Journal of Experimental Social Psychology*, 65, 1–6. <https://doi.org/10.1016/j.jesp.2016.01.010>
- Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, 71(2), 229–261. <https://doi.org/10.1111/bmsp.12110>
- 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>
- Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, 26(12), 1827–1832. <https://doi.org/10.1177/0956797615616374>
- Mulder, J., Lissa, C. van, Williams, D. R., Gu, X., Olsson-Collentine, A., Boeing-Messing, F., & Fox, J.-P. (2021). *BFpack: Flexible Bayes factor testing of scientific expectations* (1.0.0) [Computer software]. <https://cran.r-project.org/web/packages/BFpack/>
- 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* (1.4.1-beta.2) [R]. <https://github.com/MicheleNuijten/statcheck/releases/tag/v1.4.1-beta.2>
- 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>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (pp. xxi, 623). Houghton, Mifflin and Company.