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

Michèle B. Nuijten & Jelte M. Wicherts

- Preregistration
  - Background
  - Hypothesis
  - Method
    - Design
    - Sample
    - Detecting statistical reporting inconsistencies with statcheck
    - Analyses
      - Confirmatory
      - Power analysis
        - Estimating the number of available articles
        - Estimating effect size
        - Power simulation
      - Exploratory analyses
  - Limitations
  - Additional information
- References

Preregistration

Background

In previous research we found a high prevalence of statistical reporting inconsistencies in published psychology papers. Specifically, ~50% of the papers 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 papers 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*. *Statcheck* is a free software tool we developed that can automatically scan academic papers and detect inconsistencies in reported NHST results (Nuijten & Epskamp, 2021).

Of course, the hope of the journals that implemented *statcheck* was that inconsistency rates would decrease. In this paper, we empirically estimate the effectiveness of using *statcheck* in the peer review process.

Hypothesis

We will 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*.*

Method

Design

This is a retrospective observational study in which we compare statistical reporting inconsistencies in two groups 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#STATCHK; last accessed 28/10/22)

We will include all articles published in PS from 2003 - the first year that articles in *html* format were available, which is more suitable for text mining - up to October 2022 (the latest complete issue), 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/guide-for-authors; last accessed 29/10/22)

We will include all articles published in JESP between 2003 and November 2022 (the latest complete issue), 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). Here, too, we excluded retractions, corrections, and editorials. JEPG and JPSP articles up until 2013 had already been downloaded in previous research (Nuijten et al., 2016).

We will use 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*

We will use the R package *statcheck*, version 1.4.0-beta.4 (Nuijten & Epskamp, 2021), to automatically detect statistical reporting inconsistencies in the included papers.

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 (ɑ = .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 it would otherwise expect) 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 .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.

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.

The validity of statcheck in classifying inconsistencies is high. The interrater reliability between manual coding and *statcheck* was .76 for inconsistencies and .89 for decision inconsistencies (Nuijten et al., 2016). 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. The overall accuracy of statcheck ranged from 96.2% to 99.9% (Nuijten, Van Assen, et al., 2017).

Analyses

We will closely follow the analysis plan of Study 2 in Nuijten et al. (2017), in which we studied the change in statistical reporting inconsistencies before and after implementation of a data sharing policy in two journal types.

Confirmatory

To test our hypothesis, we are interested in the following logistic multilevel models:

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

,

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

,

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 hypothesize that in both models the coefficient b_3 is smaller than zero, which would indicate a steeper decrease in (decision) inconsistencies in "statcheck" journals.

Power analysis

We have a fixed number of articles for PS we can obtain (from 2003 to now) and we intend to include the same time frame for the other three journals. Based on the number of articles available per cell (*Period* * *statcheckJournal*), we can estimate what the minimum effect size is that we could detect with .80 power.

Estimating the number of available articles

At the time of preregistration, we did not download all articles in the sampling frame yet. We did have access to the articles that were downloaded for previous research (Nuijten et al., 2016): PS 2003-2013, JEPG 2003-2013, JPSP 2003-2013. We did not yet have articles from JESP, and we had not yet selected which articles from JPSP were published in the subsection Attitudes and Social Cognition.

We estimated the number of available articles in each journal and year as follows:

- PS in period 0: the number of articles already downloaded for a previous project
- JEPG in period 0: the number of articles already downloaded for a previous project
- JPSP (subsection ASC) in period 0: we roughly estimated that a third of the articles each volume are in the subsection ASC, so we divided the number of articles already downloaded for a previous project by three
- JESP in period 0: we roughly estimated this by multiplying the number of articles in 2003 (61) by 10 years
- All journals in period 1: we estimated the number of articles published per year by dividing the number of articles in period 0 by 10. Period 1 spans 8.8 years (January 2014 to October 2022), so we multiplied the number of articles per year by 8.8.

Table 1 shows a summary of the number of estimated articles available per journal type and time period.

To estimate the number of articles in which *statcheck* will be able to detect statistics, we used the data from (Nuijten et al., 2016), in which statcheck was used to check a large number of statistics in eight major psychology journals, among which were PS, JEPG, and JPSP. *Statcheck* detected statistics in 72.9% of PS articles, 69.3% of JEPG articles, and 85.1% of JPSP articles. We averaged these percentages to estimate the percentage of JESP articles in which *statcheck* would detect statistics (75.8%). We then used these percentages to estimate the total number of articles *statcheck* will detect results in, in this study. The results are shown in Table 1.

Table 1.

Number of research articles that we can download and that statcheck will find statistics in from the journals using statcheck *and their matched control journals, before and after* statcheck *was implemented.*

		T1: Before (matched) journal implemented *statcheck*		**T2: After (matched) journal implemented *statcheck***	
		# downloads	**# articles with statistics**	**# downloads**	**# articles with statistics**
Journals using *statcheck*	**PS**	2319[a]	1691[b]	2041[b]	1488[b]
	JESP	610[b]	462[b]	537[b]	407[b]
	Total	**2929**	**2153**	**2578**	**1895**
Matched control journals	**JEPG**	618[a]	428[b]	544[b]	377[b]
	JPSP	526[b]	448[b]	463[b]	394[b]
	Total	**1144**	**876**	**1007**	**771**

[a] observed; [b] estimated

Estimating effect size

The next section is copied for a large part from the preregistration of Study 2 in Nuijten et al. (2017), because of the high similarity in procedure.

The next step in the power analysis is to estimate the expected effect sizes (b_0, b_1, b_2, and b_3 in this case). We estimated b_0 (the prevalence of inconsistencies in the control journals before *statcheck* was implemented in the other journals) and b_2 (the prevalence of inconsistencies in the *statcheck* journals before *statcheck* was implemented) with a multilevel logistic regression based on the data obtained by running *statcheck* version 1.4.0-beta.4 on the subsample of articles we already had (PS, JEPG, and JPSP 2003-2013):

$Logit[inconsistency] = b_0 + b_2 statcheckJournal + \theta_i$.

The estimated b_0 was -2.53 (SE = .035), and b_2 was .039 (SE = .055). The variance of the intercepts (i.e., the variance of b_0) was τ^2 = 1.33.

The effect size b_1 indicates the change in inconsistencies from period 0 to period 1 in the control journals. At the time of preregistration, we did not have articles from period 1 yet, so as an approximation, we used the trend of inconsistencies over time in five APA journals from Nuijten et al. (2016): b_1 = -.11 (SE = .02).

For b_3, the effect of interest, we tried several values. First, we estimated two extreme values for the probability of an inconsistency in *statcheck* journals in period 1 (after implementation of *statcheck*): either the probability of an inconsistency is equal to that in *statcheck* journals in period 0 (no effect of implementing *statcheck* at all), or the inconsistency probability has dropped immensely in period 1 from ~10% - the general probability that a statistical result is inconsistent (Nuijten et al., 2016) - to 1% (large effect of implementing *statcheck*). These estimates correspond to logits of 0 and -2.00, respectively. We then selected eight equally spaced logits within this range, leaving us with ten different inconsistency logits.

For each of the ten values combined with the earlier estimates of b_0, b_1, b_2, and θ_i , we ran a simulation study to estimate power.

Power simulation

We performed a simulation study to estimate the highest possible power with the estimated number of articles available per journal type and time period (see Table 2). We simulated data sets based on the number of available articles, and the estimates of the effects as described in the sections above.

We looped over the 10 different values for b_3 to calculate power for each potential effect. For each b_3 value, we repeated the simulation 100 times. In each repetition, we did the following:

1. Simulate inconsistencies for articles from control journals before *statcheck* implementation
    a. For one article, draw a random effect *t* from a normal distribution with mean 0 and variance τ^2 = 1.33, as estimated above.
    b. Estimate the probability that a *p*-value is inconsistent based on the following equation:
    $Logit[inconsistency = 1] = b_0 + t = \; - 2.53 + t$
    $P(inconsistency) = e^{logit}/(1 + e^{logit})$
    c. Determine the number of *p*-values within the article by drawing from the frequency distribution of the number of *p*-values in a control journal article (N; median = 25)

   d. Randomly determine the number of inconsistent p-values within the article using the binomial distribution with parameters N and P
   e. Repeat for the total number of articles in control journals before *statcheck* implementation
2. Simulate inconsistencies for articles from control journals after *statcheck* implementation
   a. See 1a
   b. Estimate the probability that a *p*-value is inconsistent based on the following equation:
      $Logit[inconsistency] = b_0 + b_1 + t = -2.53 -.11 + t$
      $P(inconsistency) = e^{logit}/(1 + e^{logit})$
   c. See 1c
   d. See 1d
   e. Repeat for the total number of articles in control journals after *statcheck* implementation
3. Simulate inconsistencies for articles from *statcheck* journals before *statcheck* implementation
   a. See 1a
   b. Estimate the probability that a *p*-value is inconsistent based on the following equation:
      $Logit[inconsistency] = b_0 + b_2 + t = -2.53 +.04 + t$
      $P(inconsistency) = e^{logit}/(1 + e^{logit})$
   c. Determine the number of *p*-values within the article by drawing from the frequency distribution of the number of *p*-values in a *statcheck* journal article (N; median = 8)
   d. See 1d
   e. Repeat for the total number of articles in *statcheck* journals before *statcheck* implementation
4. Simulate inconsistencies for articles from *statcheck* journals after *statcheck* implementation
   a. See 1a
   b. Estimate the probability that a *p*-value is inconsistent based on the following equation:
      $Logit[inconsistency] = b_0 + b_1 + b_2 + b_3 + t = -2.53 -.11 +.04 + b_3 + t$
      $P(inconsistency) = e^{logit}/(1 + e^{logit})$
   c. See 3c
   d. See 1d
   e. Repeat for the total number of articles from *statcheck* journals after *statcheck* implementation
5. Combine the data for the four conditions in one data set
6. Estimate the following multilevel logistic model (the model of interest):
   $Logit[Inconsistency] = b_0 + b_1 Period_i + b_2 statcheckJournal_i + b_3 Period_i * statcheckJournal_i + \theta_i$
7. Save the *p*-value of coefficient b_3

After 100 repetitions, we determined how often the *p*-value was below .05, which renders the power that can be obtained for specific values of b_3 with the number of articles available. The results are displayed in Table 4. The total probability of an inconsistency for a given b_3 is calculated as follows:

$$Logit[inconsistency] = b_0 + b_1 + b_2 + b_3 + t = -2.53 -.11 +.04 + b_3 + t,$$

so that

$$P(inconsistency) = e^{logit}/(1 + e^{logit}).$$

Note that the probabilities seem smaller than expected based on the general error prevalence found in Nuijten et al. (2016). This 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.

Logistic regression coefficients to indicate the effect of implementing statcheck *in the peer review process of two journals with corresponding power values and total probabilities of an inconsistency based on 100 runs.*

	No effect *statcheck*				←→				Massive effect *statcheck*	
b_3 (logit)	0.00	-0.22	-0.44	-0.67	-0.89	-1.11	-1.33	-1.56	-1.77	-2.00
Power	.05	.64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Total probability of inconsistency	.069	.056	.045	.037	.030	.024	.019	.015	.012	.010

From these results we conclude that we have a power of .80 if the true regression coefficient of interest lies between -.22 and -.44 (see also Figure 1). This is equivalent to a reduction in inconsistencies of 1 - 2 percentage points (.069 - .056 = .013 and .069 - .045 = .024). We expect a much larger reduction in the probability of an inconsistency, since the intervention is so straightforward. However, if the population effect size is smaller (in absolute sense) than -.22, we might not have enough power to detect it.

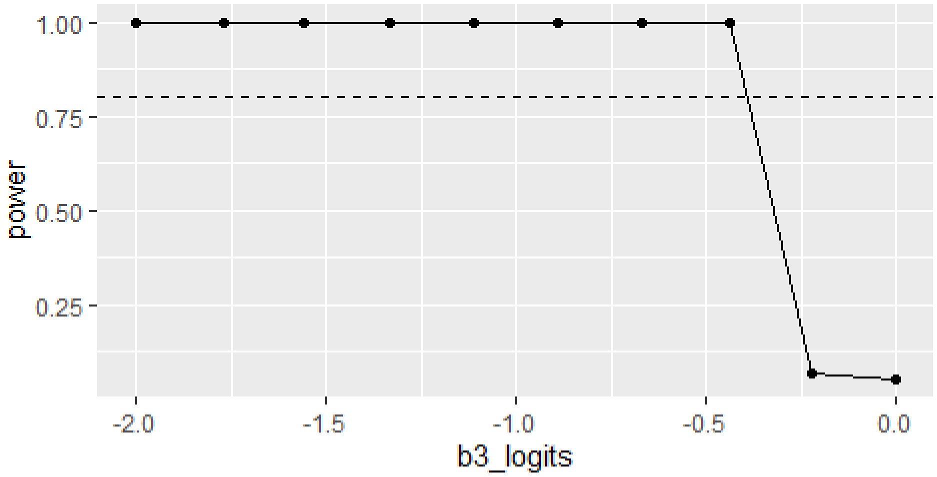

Figure 1. *Estimated power for different values of logistic regression coefficient b_3. The horizontal dotted line indicates power at the nominal .80 level.*

Exploratory analyses

We will also look at the following exploratory questions:

- Did the inconsistency rate at the control journals change (decline) over time?
- How did trends in inconsistencies change compared to the results in Nuijten et al., 2016?
- Do the results change if we look at the journal pairs separately?

We do not rule out that we might include additional exploratory analyses in the final paper, but any exploratory analyses will be clearly identified as such.

Limitations

The proposed design has some limitations. First, causal conclusions are not warranted because of the inherent limitations of an observational design. By including the matched control journals we take into account some potential confounding variables, but not all. For example, we cannot rule out selection effects: researchers who are more diligent in their reporting and/or are already familiar with statcheck, might be more likely to submit to the journals that use statcheck in peer review, because it aligns with their own values.

Another limitation is that we do not know to what extent the editorial teams actually implemented their *statcheck* policy. If editors did not follow up on flagged statistical inconsistencies in a manuscript, or failed to run *statcheck* at all, the observed effect would decrease.

Additional information

The ethics review board of the Tilburg School of Social and Behavioral Sciences at Tilburg University approved this research design (nr. TSB_RP773).

References

American Psychological Association. (2019). *Publication manual of the American Psychological Association* (7th ed.).

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

Nuijten, M. B., Borghuis, J., Veldkamp, C. L., Dominguez-Alvarez, L., Van Assen, M. A., & Wicherts, J. M. (2017). Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra: Psychology*, *3*(1). https://doi.org/10.1525/collabra.102

Nuijten, M. B., & Epskamp, S. (2021). *statcheck: Extract statistics from articles and recompute p-values* (1.4.0-beta.4) [R]. https://github.com/MicheleNuijten/statcheck/releases/tag/v1.4.0-beta.4

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