The validity of the tool “statcheck” in discovering statistical reporting inconsistencies

Nuijten, Michele B.; Assen, Marcel A. L. M. van; Hartgerink, Chris; Epskamp, Sacha; Wicherts, Jelte M.

DOI:
10.31234/osf.io/tcxaj

Publication date:
2017

Document Version
Early version, also known as pre-print

The Validity of the Tool "statcheck" in Discovering Statistical Reporting Inconsistencies

Michèle B. Nuijten[1*]

Marcel A. L. M. van Assen[1, 2]

Chris H. J. Hartgerink[1]

Sacha Epskamp[3]

&

Jelte M. Wicherts[1]

1. Department of Methodology & Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, the Netherlands.
2. Section Sociology, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, the Netherlands.
3. Department of Psychological Methods, University of Amsterdam, Amsterdam, the Netherlands

* Corresponding author. Department of Methodology & Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, the Netherlands. PO Box 90153, 5000 LE Tilburg. Email: m.b.nuijten@tilburguniversity.edu. Phone: 0031 13 466 2053.

The preparation of this article was supported by the ERC consolidator grant IMPROVE (grant no. 726361) from the European Research Council. We would like to thank Sofie Swaans for her assistance in coding the articles for statistical corrections.

Abstract

The R package "statcheck" (Epskamp & Nuijten, 2016) is a tool to extract statistical results from articles and check whether the reported *p*-value matches the accompanying test statistic and degrees of freedom. A previous study showed high interrater reliabilities (between .76 and .89) between statcheck and manual coding of inconsistencies (.76 - .89; Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2016). Here we present an additional, detailed study of the validity of statcheck. In Study 1, we calculated its sensitivity and specificity. We found that statcheck's sensitivity (true positive rate) and specificity (true negative rate) were high: between 85.3% and 100%, and between 96.0% and 100%, respectively, depending on the assumptions and settings. The overall accuracy of statcheck ranged from 96.2% to 99.9%. In Study 2, we investigated statcheck's ability to deal with statistical corrections for multiple testing or violations of assumptions in articles. We found that the prevalence of corrections for multiple testing or violations of assumptions in psychology was higher than we initially estimated in Nuijten et al. (2016). Although we found numerous reporting inconsistencies in results corrected for violations of the sphericity assumption, we demonstrate that inconsistencies associated with statistical corrections are not what is causing the high estimates of the prevalence of statistical reporting inconsistencies in psychology.

Keywords: statcheck, reporting errors, validity, sensitivity and specificity, statistical corrections, sphericity, Bonferroni

In psychological research, most conclusions are based on Null Hypothesis Significance Testing (NHST; see e.g., Cumming et al., 2007; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). Unfortunately, there is increasing evidence that reported NHST results are often inconsistent. In psychology, roughly half of all articles published in reputable journals contain at least one inconsistent result in which the reported *p*-value does not correspond with the accompanying test statistic and degrees of freedom. In roughly one in eight articles there is at least one grossly inconsistent result in which the reported *p*-value is statistically significant (i.e., $p < .05$) but the recomputed *p*-value based on the test statistic and degrees of freedom is not, or vice versa (Bakker & Wicherts, 2011; Bakker & Wicherts, 2014; Caperos & Pardo, 2013; Nuijten et al., 2016; Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014). These inconsistencies can lead to erroneous substantive conclusions and affect the reliability of meta-analyses, so it is important that these inconsistencies can be easily spotted, corrected, and hopefully prevented.

To facilitate the process of detecting and correcting statistical reporting inconsistencies, we developed the R package "statcheck" (Epskamp & Nuijten, 2016; http://statcheck.io). Statcheck is a free and open-source algorithm that extracts NHST results reported in APA style from articles and recalculates *p*-values based on the reported test statistic and degrees of freedom. If the reported *p*-value does not match the computed *p*-value, the result is flagged as an inconsistency. If the reported *p*-value is significant ($\alpha = .05$) and the computed *p*-value is not, or vice versa, the result is flagged as a gross inconsistency.

To ensure that statcheck is a valid tool for detecting statistical inconsistencies, we included a detailed validity study in Nuijten et al. (2016), in which we ran statcheck on a set of articles that had previously been manually coded for statistical reporting inconsistencies by Wicherts, Bakker, and Molenaar (2011). Using the manually coded results as the standard we found that the interrater reliability of statcheck was .76 for flagging inconsistencies and .89 for gross inconsistencies. We reported

in detail where any discrepancies came from and concluded that the validity of statcheck was sufficiently high to recommend its use for self-checks, peer review, and research on the prevalence of (gross) inconsistencies in large bodies of literature (for details see Appendix A in Nuijten et al., 2016).

Since the publication of the study in which statcheck was introduced and validated, statcheck has begun to be used in large-scale assessments (Baker, 2015, 2016; Hartgerink, 2016) and in the peer-review process of the journals *Psychological Science* and the *Journal of Experimental Social Psychology*. Since then, we have received additional questions about different aspects of statcheck's validity. First, we chose to express statcheck's validity in interrater reliability coefficients, but both in personal communications and in an anonymous review, researchers asked for more information about statcheck's false positive and false negative rate. Therefore, in Study 1 in this paper we present an analysis of statcheck's accuracy by calculating its sensitivity (true positive rate) and specificity (true negative rate; Altman & Bland, 1994). Second, Schmidt (2016) published a critique online in which he questioned the validity of statcheck in the presence of NHST results that are adjusted to correct for multiple testing or possible violations of assumptions. In Study 2 in this paper we estimate the prevalence of such statistical corrections in psychology in the large sample of articles used in Nuijten et al. (2016), and investigate whether the presence of such corrections is associated with statistical reporting inconsistencies and could have caused the high prevalence of these inconsistencies.

Study 1: Sensitivity & Specificity

In Nuijten et al. (2016) we determined statcheck's validity by means of the interrater reliability between manual coding and statcheck's results. Another common way to determine an instrument's accuracy is to calculate its sensitivity and specificity (Altman & Bland, 1994). In calculating sensitivity and specificity we use the following terminology (Baratloo, Hosseini, Negida, & El Ashal, 2015):

True Positive (TP): the number of results correctly flagged as a (gross) inconsistency

False Positive (FP): the number of results incorrectly flagged as a (gross) inconsistency

True Negative (TN): the number of results correctly *not* flagged as a (gross) inconsistency

False Negative (FN): the number of results incorrectly *not* flagged as a (gross) inconsistency

Sensitivity refers to the "true positive rate": the proportion of "true" (gross) inconsistencies that were also flagged by statcheck as such:

$$sensitivity = \frac{TP}{TP+FN}. \quad (1)$$

Specificity refers to the "true negative rate": the proportion of results that are "truly" not (grossly) inconsistent, and statcheck correctly did not flag them as (gross) inconsistencies:

$$specificity = \frac{TN}{TN+FP}. \quad (2)$$

Together, sensitivity and specificity say something about statcheck's accuracy: the ability to correctly differentiate between consistent and (grossly) inconsistent results, or more mathematically:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \quad (3)$$

Ideally, accuracy should be 100%, which would mean there are no false positives or false negatives, but for an automated algorithm such as statcheck this is not tenable. Statcheck will probably be less accurate than a manual check, but we do want to minimize false positives and false negatives in flagging (gross) inconsistencies. To calculate statcheck's accuracy, sensitivity, and specificity, we used the same reference data set that we used to calculate the interrater reliabilities in Nuijten et al. (2016), namely data from the manual checks of Wicherts et al. (2011).

Reference Sample

As a reference set, we used the same sample as Wicherts et al. (2011), who manually coded the internal consistency of NHST results from 49 articles from the *Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP:LMC)* and the *Journal of Personality and Social Psychology (JPSP)*. These authors included results from *t*, *F*, or χ^2-tests that were reported completely (test statistic, degrees of freedom, and *p*-value) in the Results section of an article. From this set, they only selected results with *p*-values smaller than .05. This resulted in a total set of 1,148 NHST results.

Procedure

We ran different versions of statcheck over the articles included in Wicherts et al. (2011). One article was excluded from the set, because it was retracted due to misconduct. Our final sample consisted of 48 articles and 1,120 NHST results. In all runs, we ran statcheck both with and without automated one-tailed test detection (OneTailedTxt = TRUE). With this option, statcheck considers a result consistent if (1) the reported *p*-value would be consistent if it belonged to a one-tailed test (specifically: if the reported *p* times two equals the computed *p*), and (2) if anywhere in the full text of the article statcheck found the word "one-tailed", "one-sided", or "directional". All other statcheck options were set to their default settings (see section 3.5 in the statcheck manual at http://rpubs.com/michelenuijten/statcheckmanual).

We compared the sensitivity and specificity of the three different versions of statcheck that are published on CRAN: statcheck 1.0.0 (Epskamp & Nuijten, 2014), statcheck 1.0.1 (Epskamp & Nuijten, 2015), and statcheck 1.2.2 (Epskamp & Nuijten, 2016). There were no major changes in the core code of statcheck 1.0.1 as compared to statcheck 1.0.0, but there were some relevant changes in version 1.2.2. In statcheck 1.2.2 there was a bug fix to ensure that statcheck does not misread *t*, *F*, or *r* statistics with a subscript as chi-square tests. We also adapted the code so that statcheck would still recognize a degree of freedom reported as the lower case letter L ("l") as the number "1". Furthermore, if statcheck detects

a correlation that is reported as > 1, it neither calculates a *p*-value nor determines whether the result is inconsistent. The main reason is that when statcheck found a correlation larger than one the risk was too high that it had mistakenly identified a different test as a correlation, leading to a falsely flagged inconsistency. Finally, in version 1.2.2 we fixed a bug in the way statcheck flagged inconsistencies in inexactly reported test statistics (e.g., $t(38) < 1.00$, p = …). The full history and specific code of all changes to statcheck can be found on GitHub at https://github.com/michelenuijten/statcheck.

From the statistical results that statcheck detected, we selected results from *t*, *F*, or χ^2-tests that had a *p*-value smaller than .05 to match the inclusion criteria of Wicherts et al. (2011). Wicherts et al. only included results reported in the Results section, but statcheck cannot distinguish in which section of an article a result was reported and hence also extracted results from different sections. Conversely, Wicherts et al. included results from tables, but statcheck only extracts complete, complete NHST results reported in APA style, which are usually not reported in tables.

Next, we compared the results from Wicherts et al. (2011) with the results from statcheck. We first checked what percentage of manually extracted results were also detected by statcheck. We then selected the NHST results that were extracted both by Wicherts et al. (2011) and statcheck, and continued to investigate if the manual classifications of inconsistencies and gross inconsistencies matched those of statcheck.

The reference data from Wicherts et al. (2011) and the full R scripts to clean and select the data and calculate sensitivity and specificity are available from https://osf.io/753qd/. The articles on which the data are based and on which statcheck was run are published on the private web page https://osf.io/ske8z/, and can be shared upon request.

Results

The accuracy, sensitivity, and specificity were almost exactly the same for the three versions of statcheck. The reason for this is that all updates to the code were made to solve specific problems that only occurred in a single instance the set of reference articles: one result was reported as "F(l, 76) = 23.95, p <.001", and only statcheck version 1.2.2 recognized the first degree of freedom as a 1. This means that the versions 1.0.0 and 1.0.1 of statcheck detected 684 NHST results of the 1,120 results (61.1%) that were included in Wicherts et al. (2011), and version 1.2.2 detected 685 NHST results (61.2%).[1] To calculate statcheck's sensitivity and specificity in detecting (gross) inconsistencies, we only focused on the NHST results that were detected both by Wicherts et al. (2011) and statcheck. The results of the sensitivity and specificity analysis for all three statcheck versions are displayed in Table 1.

[1] We excluded one article with 28 NHST results reported in APA style because it was retracted due to misconduct; this article was included by Wicherts et al. (2011). Note that in the validity study of Nuijten et al. (2016) we reported that statcheck detected 775 NHST results, as that study also included results that were detected by statcheck but not included in the manual check (mainly results reported in a section other than the Results section).

Table 1

Results of the sensitivity and specificity analysis of statcheck 1.0.0 (Epskamp & Nuijten, 2014), statcheck 1.0.1 (Epskamp & Nuijten, 2015), and statcheck 1.2.2 (Epskamp & Nuijten, 2016), with and without one-tailed test detection. The statcheck version was indicated (e.g., "v. 1.0.0") if results differed between versions. The reference set consisted of manually coded data by Wicherts et al. (2011). TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative.

	statcheck (default)					**statcheck (with automated one-tailed test detection)**				
	TP	**FP**	**TN**	**FN**		**TP**	**FP**	**TN**	**FN**	
Inconsistencies	34	26	624 v. 1.0.0 - 1.0.1 625 v. 1.2.2	0		29	19	631 v. 1.0.0 - 1.0.1 632 v. 1.2.2	5	
Sensitivity					100%					85.3%
Specificity					96.0%					97.1%
Accuracy					96.2%					96.5%
Inconsistencies (strict)*	52	8	624 v. 1.0.0 - 1.0.1 625 v. 1.2.2	0		47	5	631 v. 1.0.0 - 1.0.1 632 v. 1.2.2	5	
Sensitivity					100%					90.4%
Specificity					98.7%					99.8%
Accuracy					98.8%					99.1%
Gross Inconsistencies	8	6	670 v. 1.0.0 - 1.0.1 671 v. 1.2.2	0		7	0	676 v. 1.0.0 - 1.0.1 677 v. 1.2.2	1	
Sensitivity					100%					87.5%
Specificity					99.1%					100%
Accuracy					99.1%					99.9%

* Here we consider the 7 results reported as "p = .000" and the 11 cases in which a Huynh-Feldt correction was applied, but the uncorrected degrees of freedom were reported, as true inconsistencies.

True inconsistencies. The diagnostic accuracy of statcheck depended on whether statcheck was run with or without its automated one-tailed test detection. In default mode (without one-tailed test detection) statcheck's sensitivity was 100%; all 34 true inconsistencies in the selected set of NHST result were correctly flagged by statcheck as such. In other words, in default mode there were no false negatives when flagging inconsistencies. When statcheck was run with one-tailed test detection, however, sensitivity decreased to 85.3%. In this case, statcheck failed to flag 5 of the 34 true inconsistencies as such. In these cases, statcheck was too lenient in counting results as one-tailed tests: if the words "one-tailed", "one-sided", or "directional" were mentioned anywhere in the article, *all* results that would be consistent if they were one-tailed were counted as correct. To be able to decide for individual results whether or not it is one-tailed, we would need an algorithm that can interpret text substantively, which does not currently fit within the scope of the statcheck project.

The specificity of statcheck also depended on whether the one-tailed test detection was used, but here we see the opposite pattern: specificity was higher with one-tailed test detection. When statcheck was run with default options (without one-tailed test detection) its specificity was 96.0%: depending on the version, either 624 of the 650 truly consistent results (version 1.0.0 and 1.0.1) or 625 of the 651 truly consistent results (version 1.2.2) were correctly flagged as consistent by statcheck. This means that 26 results were "false positives": results that statcheck flagged as an inconsistency, whereas they were counted as correct in the manual check. Eight of these false positives were one-tailed tests that statcheck did not recognize. Indeed, when we ran statcheck with one-tailed test detection, the specificity increased to 97.1%: now statcheck correctly flagged 631 of the 650 consistent results in version 1.0.0 and 1.0.1, and 632 of the 651 results in version 1.2.2. Note that one one-tailed result was still wrongly flagged as an inconsistency. In this case, the one-tailed *p*-value was probably based on an unrounded test statistic, whereas the (correctly) rounded test statistic was reported. In general, statcheck takes into account correct rounding of the test statistic, but when the one-tailed test

detection is used, statcheck uses the exact reported test statistic to calculate whether a one-tailed *p*-value would be consistent. In future versions of statcheck, this feature will be adapted to take into account one-tailed *p*-values based on correctly rounded test statistics.

The sensitivity and specificity of statcheck can be combined to reflect its accuracy: the ability to correctly differentiate between consistent and (grossly) inconsistent results (see Equation 3). In default mode, without one-tailed test detection, the accuracy was 96.2%. If one-tailed test detection was switched on, statcheck's overall performance slightly increased to an accuracy of 96.5%.

Eight of the 26 false positives could be explained by the use of one-tailed tests. The remaining 18 false positives were caused by two main things. First, statcheck counted results reported as $p = .000$ as inconsistent, because this is not in line with APA reporting practices. Cases like this should, according to the APA, be reported as $p < .001$. Wicherts et al. (2011) did not automatically count this as incorrect. There were seven such cases in this data set. Second, there were eleven results in which a Huynh-Feldt correction was applied to correct for a violation of the assumption of sphericity. A Huynh-Feldt correction adjusts the result by multiplying the degrees of freedom by a factor "ε". However, in all eleven cases that we detected, the unadjusted degrees of freedom were reported along with the adjusted *p*-value, which rendered the result internally inconsistent. In the manual check, this was still counted as consistent, because it was traceable how the correction had been applied. We consider and discuss such corrections in more detail below.

True inconsistencies (strict). Our results showed that most of the "false positives" (i.e., results marked by statcheck as inconsistent, but counted as consistent in the manual check) resulted from conscious choices in programming statcheck. We deliberately chose to consider $p = .000$ as inconsistent, because a *p*-value can never be exactly zero. The APA prescribes that such results should be reported as $p < .001$ (American Psychological Association, 2010). The same line of reasoning applies to the Huynh-

Feldt corrections. If these corrections were correctly reported (i.e., the adjusted degrees of freedom together with the adjusted *p*-value), they would have been consistent. If we retain both of these stricter criteria for flagging inconsistencies, statcheck's accuracy in detecting inconsistencies increases (see Table 1). With these stricter criteria, statcheck's sensitivity remains at 100% when it is run in default mode, or increases from 85.3% to 90.4% when one-tailed test detection is used. Similarly, the specificity increases from 96.0% to 98.7% (default) and from 97.1% to 99.8% (one-tailed test detection). These increases in sensitivity and specificity are also reflected in the overall accuracy, which increases from 96.2% to 98.8% (default) and from 96.5% to 99.1% (one-tailed test detection). Retaining these stricter criteria for flagging inconsistencies had no bearing on the sensitivity and specificity in detecting gross inconsistencies.

True gross inconsistencies. Similar to its performance when flagging inconsistencies, statcheck's sensitivity in detecting true gross inconsistencies depended on whether one-tailed test detection was used. In default mode, without one-tailed test detection, statcheck's sensitivity was 100%: all 8 true gross inconsistencies were correctly flagged as such. However, when one-tailed test detection was used, the sensitivity dropped to 87.5%: statcheck correctly identified 7 out of the 8 gross inconsistencies. The one missed gross inconsistency was due to the automatic one-tailed test detection being too lenient. The article that contained this specific gross inconsistency mentioned "directional" in the full text, which caused statcheck to count the result as a one-tailed test, but the manual check revealed that "directional" did not refer to the statistical analyses.

The specificity of statcheck in detecting results that were truly not grossly inconsistent also depended on the one-tailed test detection. In default mode, without one-tailed test detection, statcheck's specificity was 99.1%: 670 of the 676 results that were truly not gross inconsistencies were correctly identified as such. There were 6 results that statcheck wrongly flagged as a gross inconsistency, because statcheck did not recognize that these were one-tailed tests. Indeed, when we ran statcheck

with its one-tailed test detection the specificity increased to 100%. In other words, with one-tailed test detection, there were no false positives in detecting gross inconsistencies.

The sensitivity and specificity in detecting gross inconsistencies combined led to an accuracy of 99.1% if statcheck was run in default mode without one-tailed test detection, and increased to 99.9% when one-tailed test detection was used.

Conclusion

The analysis of statcheck's diagnostic accuracy showed low false positive and false negative rates in flagging inconsistencies and gross inconsistencies. The sensitivity (flagging true [gross] inconsistencies) ranged from 85.3% to 100%, and the specificity (flagging results that are truly not [grossly] inconsistent) ranged from 96.0% to 100%. Combined, these results indicate that the accuracy ranged from 96.2% to 99.9%. We considered these results evidence that the validity of statcheck is high.

The exact sensitivity and specificity depended on several conditions. Firstly, we found that statcheck's sensitivity and specificity depended on whether its automated one-tailed test detection was used. The sensitivity of statcheck was highest without one-tailed test detection, whereas the specificity was highest when statcheck was run *with* one-tailed test detection. Users can take this into account when they decide whether or not to use the one-tailed test detection; if they find it most important to avoid false positives and not falsely flag correct results as inconsistent, they should use one-tailed test detection. Conversely, when they find it most important to avoid false negatives and they want to flag every result that could potentially be inconsistent, they should use statcheck without one-tailed test detection. This is a standard trade-off with any diagnostic instrument.

Other factors that influenced statcheck's exact specificity and sensitivity depended on conscious programming choices. Specifically, the large majority of "false positives" in statcheck's flagged inconsistencies was due to either the conscious choice of considering $p = .000$ as inconsistent, or

(incorrectly) reported adjustments of statistical results. These adjustments and, and the inconsistencies related to them, are the focus of the second part of this paper.

Generalizability Sensitivity & Specificity

A clear limitation of this additional validity study is that we used a single manually coded sample as a reference set. However, we do not believe that statcheck's diagnostic accuracy varies considerably across articles, journals, or disciplines. This belief is strengthened by the fact that we keep finding very similar inconsistency rates across different samples (see, e.g., the different estimates in the three studies in Nuijten et al. (in press), or the summary of different studies about the prevalence of inconsistencies in Table 2 in Nuijten et al. (2016)).

We also see no reason to expect large differences in sensitivity and specificity between the different versions of statcheck. The adaptations to the code across different versions were mainly aimed at improving statcheck's detection rate of NHST results in peculiar and rather infrequent cases. Hence, statcheck's sensitivity and specificity will remain the same or likely only slightly increase over versions. As an extra check, we ran statcheck 1.0.0, 1.0.1, and 1.2.2 on a different sample of articles to see if there were any changes in the detected prevalence of (gross) inconsistencies. For this check we used a set of 137 psychology meta-analyses that we had collected for a different study (Nuijten, van Assen, Augusteijn, Crompvoets, & Wicherts, in preparation). We found that even though the detection rate of statcheck increased in version 1.2.2 (226 NHST results as opposed to 215 results in the previous two versions), the numbers of detected inconsistencies (25) and gross inconsistencies (4) were the same. Since the total number of detected results increased, the percentage of results that were inconsistent or grossly inconsistent decreased slightly (from 11.6% to 11.1%, and from 1.9% to 1.8%, respectively). In short, even though the sensitivity and specificity were calculated based on only one reference set of

articles, we argue that these results can be generalized to different versions of statcheck and different sets of articles.

Study 2: Accounting for Corrections for Multiple Testing, Post Hoc Testing, or Possible Violations of Assumptions

A possible cause for a detected inconsistent result is the use of statistical corrections for multiple testing, post hoc testing, or possible violations of assumptions. Take for example the Bonferroni correction for multiple testing. This correction is used to control the Type I error rate by dividing the level of significance (α) by the number of hypotheses tested. However, we often see cases in which instead of dividing α, researchers multiply the *p*-values by the number of tests. This then results in an internally inconsistent statistical result: the original test statistic and degrees of freedom no longer correspond to the reported (multiplied) *p*-value. Such cases will be flagged by statcheck as an inconsistency.

In Nuijten et al. (2016) we intended to give a rough estimate of the prevalence of corrected *p*-values to illustrate that these were an unlikely cause of the many inconsistent *p*-values we found. We used the "Search" function in Windows Explorer to search the entire folder of downloaded articles for "Bonferroni" and "Huynh-Feldt", and reported the following results:

"[...] when we automatically searched our sample of 30,717 articles, we found that only 96 articles reported the string "Bonferroni" (0.3 %) and nine articles reported the string "Huynh-Feldt" or "Huynh Feldt" (0.03 %). We conclude from this that corrections for multiple testing are rarely used and will not significantly distort conclusions in our study." (Nuijten et al., 2016, p. 1207)

On the post-publication peer review forum “PubPeer” and the e-print service “arXiv”, Schmidt (2016) expressed his concern that we underestimated the prevalence of corrections for two reasons. First, we only searched for "Bonferroni" and "Huynh-Feldt", but did not include several other types of corrections.

Second, estimates based on Schmidt's own library and our validation sample (Wicherts et al., 2011) resulted in a higher prevalence of corrections than the one we had found in our full search. Therefore, we decided to re-estimate the prevalence of corrected p-values in our sample of the literature. We also examined whether corrections were associated with inconsistently reporting statistical results.

Re-estimating Prevalence of Correction-Related Strings

We re-estimated the prevalence of articles that might contain statistical results that were adjusted by one of the corrections mentioned by Schmidt: Bonferroni, Scheffé, Tukey, Greenhouse-Geisser, and Huynh-Feldt. To this end, we ran a shell script with full text searches on the full database of 30,717 articles used in Nuijten et al. (2016). The full script is available at https://osf.io/v9msf/. The script counts all articles that contain the string "Bonferroni", "Tukey", "Scheff", "Greenhouse", or "Huynh". Using this method, the results showed a much higher prevalence of strings of text that could point to corrected statistical results than we had found in our original estimate, confirming Schmidt's (2016) suspicion that we had initially missed many of these corrected results (see Table 2). We speculate that something went wrong in the Windows Explorer Search function in Nuijten et al. (2016), but we were not able to determine this with certainty. However, more important than finding the cause of this discrepancy is investigating whether these results of statistical corrections are associated with reporting inconsistencies and hence might influence our original conclusion that statistical corrections are an unlikely cause for the high prevalence of statistical reporting inconsistencies in the psychological literature.

Table 2

New estimates of the number and percentage of articles that mentioned any of the listed types of corrections compared to the estimates we mentioned in Nuijten et al. (2016), based on the total number of 30,717 downloaded articles.

Correction Type	**# Articles found with shell script**	**% Articles estimated with shell script**	**% Originally estimated with Windows Explorer**
Bonferroni	2,744	8.93	0.30
Tukey	1,691	5.51	NA*
Scheffé	667	2.71	NA
Greenhouse-Geisser	898	2.92	NA
Huynh-Feldt	234	0.76	0.03
Articles with one or multiple corrections	5,513	17.9	NA

* NA = Not Available, i.e., these were not examined in Nuijten et al. (2016).

Percentage of Inconsistencies in Articles without Adjusted Statistics

One way to determine whether the presence of statistical corrections and adjustments has influenced our estimate of the prevalence of (gross) inconsistencies in the psychological literature is to remove all articles from the analysis that show any sign of containing such a correction or adjustment. If a large part of the detected inconsistencies in Nuijten et al. (2016) were due to statistical adjustments, one would expect that a subset of articles without any corrections would show a lower prevalence of inconsistencies and gross inconsistencies.

For this analysis, we first needed to identify which of the 16,695 articles that we analyzed in our paper contained evidence for adjusted statistics. We used the same shell script as above to determine

which articles mentioned any of the keywords "Bonferroni", "Tukey", "Scheff", “Greenhouse”, or “Huynh” (the full script can be found at https://osf.io/v9msf/). This resulted in a list of 6,234 article titles, of which 5,513 were unique (some of the articles contained multiple keywords and appeared in the list two or more times). We found that 2,396 of the 16,695 (14.4%) articles in which statcheck detected APA reported NHST results contained at least one of the keywords that possibly indicated the presence of statistical corrections.[2] To extract the NHST results and detect inconsistencies, statcheck version 1.0.1 was used with automated one-tailed test detection (Nuijten et al., 2016).

We removed all articles with any evidence for the presence of statistical adjustments from the sample and re-estimated the general prevalence of inconsistencies and gross inconsistencies (see Table 3). The results showed that removing articles with possible corrections led to a slightly *higher* prevalence of inconsistencies and gross inconsistencies than was found in Nuijten et al. (2016). This suggests that the original estimates of the high prevalence of inconsistencies in the psychological literature are not driven by the presence of tests corrected for multiple testing, post hoc testing, or violations of assumptions. A possible explanation for this unexpected increase in the detection rate of inconsistencies here is that researchers who apply statistical corrections might be more diligent when it comes to their statistics, which might also decrease the probability that they report one or more results inconsistently.

The full R code of this analysis can be found at https://osf.io/t7b6m/. The raw data of Nuijten et al. (2016) with article identifiers, from which we selected a sample of articles to manually code, is published on a private page at OSF (https://osf.io/sa87e/); due to ethical restrictions these data are only

[2] For 20 articles in the original sample we were unable to automatically check if these titles also occurred in the list of articles with possible corrections, because problems in automatically reading in the file names due to special symbols in the article title. Because we were not sure if these articles contained evidence for statistical corrections, we ran our analyses with and without these articles. Removing these articles on top of the articles that did contain evidence for corrections did not change the estimates of inconsistency prevalence.

available upon request. The articles that were scanned in Nuijten et al. (2016) are also published on a private page (https://github.com/MicheleNuijten/sampleStatcheck) and are available upon request.

Table 3

Estimates of the prevalence of (gross) inconsistencies in the full sample of Nuijten et al. (2016) compared to the sample without articles that showed evidence for containing one or more statistical corrections or adjustments.

	All articles (data from Nuijten et al., 2016)	Articles without evidence for corrections
% articles with at least one inconsistency	49.6%	49.8%
% articles with at least one gross inconsistency	12.9%	14.4%
average % of p-values that are inconsistent per article	10.6%	11.1%
average % of p-values that are grossly inconsistent per article	1.6%	1.9%

Percentage of Inconsistencies that are Associated with a Statistical Correction

Schmidt (2016) argued that it is misleading if statistics are flagged as inconsistencies if they were affected by statistical corrections, because the use of statistical corrections is usually good practice and statcheck would "punish" that by flagging them as inconsistencies. However, we disagree that flagging inconsistently reported corrected statistics as such is misleading. Each of the five examples of statistical corrections that Schmidt mentioned can and should be reported in an internally consistent way.

First, a Bonferroni correction for multiple testing consists of dividing the level of significance, α, by the number of tests to adjust the level of significance. For instance, if you run 6 different tests, and you want to retain an overall α of .05, the Bonferroni corrected α for each of the tests is $\alpha = .05/6 = .00833$. However, instead of correcting α, researchers often adjust the *p*-values themselves by multiplying each *p*-value by the number of tests. In fact, this is also how SPSS carries out the Bonferroni post hoc test with the POSTHOC Bonferroni command in UNIANOVA. However, if one reports the original test statistic and degrees of freedom, but a multiplied *p*-value, the result is no longer consistent. An additional reason not to multiply the *p*-value is that with this procedure it is possible to obtain *p*-values larger than one, which are meaningless by definition.

Furthermore, Schmidt (2016) mentions the Greenhouse-Geisser correction and Huynh-Feldt correction to adjust for violations of the assumption of sphericity. In both procedures, the degrees of freedom of an *F*-test are multiplied by a factor ε that lies between 0 and 1, which increases the *p*-value of the observed *F*-statistic. Sometimes, as Schmidt also illustrates, researchers report the original, uncorrected degrees of freedom with the corrected test statistic and *p*-value. In these cases, the original degrees of freedom may be reported so that others may deduce the sample size on which the test was based, but making the statistical results inconsistent. It is recommended to report the corrected statistical result, as well as the value of ε (see, e.g., Field, 2009, p. 481).

We initially did not expect problems with statistical results of the Tukey and Scheffé post hoc tests that Schmidt mentioned. The Tukey test has its own statistic and *p*-value, and as such is not a correction of another statistical result. The Scheffé test compares the original *F*-statistic with a recalculated critical value of the *F*-test (i.e., $(K\text{-}1)\times F_{CV(K\text{-}1,N\text{-}K)}$, with *K* and *N*-*K* denoting the degrees of freedom of the *F*-test, and F_{CV} denoting the critical value of the test). Since the Scheffé test does not yield an exact *p*-value but a comparison with a significance level, e.g. $p < .05$, statcheck will not detect

these results because they are not reported in line with APA guidelines. However, as we did not know how researchers report these results, we also examined reported results of these statistical tests.

In short, we contend that all of the five corrections Schmidt mentioned (or any other, to our knowledge) can be reported in an internally consistent and informative way. However, we agree with Schmidt that some of these corrections might be reported incorrectly in research articles. To see how statistical corrections are usually reported and how often a statistical correction led to a flagged inconsistency, we manually coded a subsample of the articles investigated in Nuijten et al. (2016).

Method. We selected all articles in which we found a keyword that could indicate the use of statistical corrections (see the shell script at https://osf.io/v9msf/). For each of the five corrections (Bonferroni, Scheffé, Tukey, Greenhouse-Geisser, and Huynh-Feldt), we randomly selected 100 articles that contained the specific keyword and had at least one NHST result that statcheck was able to verify. There were only 39 articles that contained both the keyword "Huynh" and had statcheck output, so we included all of those. This procedure led to a sample of 439 articles. For the current analysis, we were only interested in cases where a statistical correction could have led to an inconsistency, so from the 439 articles we then only selected those articles in which statcheck flagged at least one inconsistency. This resulted in a final sample of 229 articles (see Table 4).

Table 4

The number of randomly selected articles that contained both a keyword indicating a statistical correction and statcheck output, and the number of these articles that also contained a flagged inconsistency.

Correction type	**# Selected articles selected with statcheck output**	**# Selected articles with statcheck output that contained at least one inconsistency**
Bonferroni	100	50
Tukey	100	51
Scheffé	100	50
Greenhouse-Geisser	100	65
Huynh-Feldt	39	25
Total	439	229

We then manually coded the inconsistencies in the selected articles, using three coders (MN, MvA, and a research assistant). As we could not make a specific protocol beforehand that anticipated all possible errors and contingencies, we decided to code results by discussion. That is, while coding different results independently in the same office, we discussed those instances where the coder was unsure about how to code the result. Each type of correction required a slightly different approach in coding, but we retained the following general approach for each article:

1. Use the Search function to find sentences that mentioned the correction that was the selection criterion for the article, to determine whether any results might be associated with this correction.

   Example: "Note that all ANOVAs reported in this article use the Greenhouse-Geisser correction for violations of sphericity."

2. Use the Search function to find all results that statcheck flagged as an inconsistency, to determine whether these results are associated with the correction that was the selection criterion for the article.

   Example: "Greenhouse–Geisser F(1.77,1098.32) = 2.34, p = 0.06."

3. If, based on the text, no inconsistency is associated with the correction, classify this as 0.

   Example: "F(3, 357) = 5.44, p<.001 (all ps still reliable when the Geisser–Greenhouse adjustment was applied)."

4. If, based on the text, an inconsistency is associated with the correction, and the cause of the inconsistency seems to be the use of the correction, classify this as 1.

   Example: "A main effect of rank [F(5, 155) = 3.57, p = 0.006, epsilon = 0.89] was observed"

   Additional signs that a result is associated with any of the corrections are:

   a. The reported *p*-value is higher than the *p*-value computed by statcheck
   b. Scheffé, Greenhouse-Geisser, and Huynh-Feldt corrections only apply to *F*-tests
   c. Tukey tests/corrections only apply to *t*-tests or *F*-tests where $df_1 = 1$

5. If, based on the text, an inconsistency is associated with the correction, but the result is still inconsistent when this correction is taken into account, classify this as 2.

   Example: "F(6,114) = 2.67, p = 0.057, ε = 0.45." If we multiply the degrees of freedom (6 and 114) by epsilon (.45), the result would be F(2.7, 51.3) = 2.67, and the accompanying p-value should be .063, which does not correspond to the reported p-value of .057.

   Example: "Greenhouse–Geisser F(1.77,1098.32) = 2.34, p = 0.06." This result is reported with the corrected degrees of freedom and should be internally consistent. However, based on the reported degrees of freedom and test statistic, the p-value should be .103.

Note that following this approach, we only looked at the type of correction that was the selection criterion for the article. If an article belonged in the category "Greenhouse-Geisser" but a result was

affected by a Bonferroni correction, this was coded as a 0. Furthermore, it was sometimes hard to distinguish between category 0 and 2; there were cases in which it was unclear whether a result was not associated with a correction, or whether the result was associated with a correction but simply wrongly reported. For instance, there were cases in which the text stated that "all *p*-values were corrected for multiple testing with a Bonferroni procedure", but then some reported *p*-values were *lower* than the recomputed ones, which is impossible if the original *p*-value was multiplied to correct for multiple testing. In this case, it is unclear whether this result was in fact not Bonferroni corrected at all (category 0), or whether the impossibly low *p*-value was the result of a typo (category 2). Therefore, in our analyses, we only focused on category 1: results in which the inconsistency was clearly associated with the statistical correction. The coded data and R scripts to analyze them are available at https://osf.io/gf9zx/.

Results. An overview of the results is shown in Table 5. In total, the 229 selected articles contained 5,606 APA reported NHST results that were extracted by statcheck, of which 798 results were inconsistently reported (14.2%). From these 798 inconsistencies, 97 (12.2%) were associated with one of the five investigated statistical corrections.

If we zoom in on the specific types of corrections, it turns out that Tukey or Scheffé corrections never led to a reporting inconsistency being flagged by statcheck. As stated earlier, Scheffé's method does not yield an exact *p*-value, which means that any test result based on Scheffé's method will not be reported in the APA style that statcheck can detect. Typically, results of Scheffé tests are reported along these lines: "Scheffé multiple-comparisons tests revealed significant differences ($ps < .05$)". It is therefore not surprising that none of the inconsistencies flagged by statcheck were associated with Scheffé's test. Similarly, we also did not expect to find many cases in which Tukey's test seemed to have affected the consistency of a result. Tukey's test has its own test statistic (q), which statcheck cannot detect. Furthermore, it turned out that the results of Tukey's test were often reported in-text, e.g.,

"Tukey's HSD test was used to specify the nature of the differences between conditions ($p < .05$, for all differences reported)", or in tables in which the significance of the results was indicated by stars. In neither of these cases could statcheck have detected the results, let alone flag an inconsistency.

In the articles that contained the keyword "Bonferroni", 17 of the 184 inconsistencies (9.2%) were caused by researchers multiplying the *p*-values instead of dividing α. The percentage of inconsistencies associated with a correction was higher in articles that showed evidence for a correction for the violation of the assumption of sphericity; in all articles that mentioned Huynh-Feldt, 14 of the 73 inconsistencies were caused by reporting the uncorrected degrees of freedom (19.2%), and in the articles that mentioned Greenhouse-Geisser, 66 of 198 inconsistencies (33.3%) were caused by reporting the uncorrected degrees of freedom.

Table 5

The total number of APA reported NHST results extracted by statcheck, the total number of those NHST results that were inconsistently reported, and the number of those inconsistencies that were caused by the use of a statistical correction. The results are split up per type of correction.

Correction type	**# APA reported NHST results in selected articles**	**# inconsistent results**	**# inconsistencies associated with correction**
Bonferroni	1,108	184	17 (9.2%)
Tukey	1,185	208	0 (0.0%)
Scheffé	898	135	0 (0.0%)
Greenhouse-Geisser	1,646	198	66 (33.3%)
Huynh-Feldt	769	73	14 (19.2%)
Total	5,606	798	97 (12.2%)

The results of this analysis of articles using statistical corrections showed that the vast majority of inconsistencies were not associated with these corrections. Test results based on Tukey's test or Scheffé's test were never reported in such a way that statcheck could detect them, which meant that these corrections never led to a reporting inconsistency being flagged by statcheck. When a Bonferroni correction was used, less than one in ten inconsistencies was actually caused by a multiplied *p*-value. Corrections for violations of sphericity, in contrast, led to more inconsistencies. Here, the uncorrected degrees of freedom were often reported alongside the corrected *p*-value (e.g., "Hereafter, when violations of sphericity occurred, we report Huynh-Feldt corrected *p*-values; for clarity, unadjusted degrees of freedom are reported.").

General Discussion

In this paper we investigated statcheck's diagnostic accuracy by calculating its sensitivity and specificity, and examined whether statistical corrections could have caused the high prevalence of statistical reporting inconsistencies as found in Nuijten et al. (2016). The results of Study 1 showed that all current versions of statcheck have high sensitivity and specificity. The majority of "false positives" in flagged inconsistencies were caused by the deliberate choice to always count "*p* = .000" as incorrect (not applied in the manual checking by Wicherts et al., 2011), and by results that had been subject to a statistical correction and therefore inconsistently reported.

The fact that statistical corrections can lead to inconsistently reported results has been presented as an argument against the use of statcheck (Schmidt, 2016).[3] However, we argue that there

[3] Schmidt's (2016) critique even led to an official statement from the DGPs (the German Psychological Society) in which they argue against the use of statcheck. In our reply we maintained our position that even though statcheck is not 100% accurate, its validity is high enough to recommend its use. Their letter, our reply, and a summary of the discussion can be found on the following Retraction Watch post:

is no reason to report the result of a corrected test in a manner that creates an inconsistency between the test statistic, degrees of freedom, and the *p*-value. In Study 2, we found no reporting inconsistencies associated with Scheffé and Tukey tests, and only some inconsistencies associated with the Bonferroni correction (9.2% of inconsistencies). We did find numerous inconsistencies associated with corrections for violations of the sphericity assumption (Greenhouse-Geisser and Huynh-Feldt; 33.3% and 19.2% of the inconsistencies, respectively). We therefore conclude that Schmidt (2016) was correct to raise the issue that some statistical corrections may be detected as reporting inconsistencies, as some of these corrections may not be consistently reported. As the APA manual does not discuss reporting of statistical corrections, we sent a message via APA's feedback form recommending that a future edition of the APA Publication Manual should incorporate specific examples of how to report these corrections in articles (e.g., "Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(5) = 11.41$, $p = .044$), therefore degrees of freedom were corrected using Huynh-Feldt estimates of sphericity ($\varepsilon = 0.67$). The results show that X was significantly affected by Y, $F(2, 13.98) = 3.79$, $p = .048$, $\omega^2 = .24$."; adapted from Field, 2009; p. 482).

Any reporting inconsistencies associated with these tests and corrections could not explain the high prevalence of reporting inconsistencies in psychology as reported in Nuijten et al. (2016), for two reasons. First, because we found that the use of these corrections is infrequent, and second because a subset of articles that showed no evidence for any of these corrections had a lower prevalence of (gross) inconsistencies than the full set of articles. Furthermore, the estimates of the prevalence of (gross) inconsistencies in Nuijten et al. (2016) are very similar to the estimates reported in other studies in which manual procedures were used (see e.g., Bakker & Wicherts, 2011; Bakker & Wicherts, 2014; Caperos & Pardo, 2013).

---

http://retractionwatch.com/2016/10/25/psychological-society-wants-end-to-posting-error-finding-algorithm-results-publicly/.

Based on the results of this validity study and the earlier validity study in Nuijten et al. (2016), and also on the convergence of estimates from the present study with studies that were based on manual checking (e.g., Bakker & Wicherts, 2011), we consider that statcheck shows a high level of diagnostic accuracy. Therefore, we recommend the use of statcheck for checking one's own work, for use in peer review, and as a tool to estimate the general prevalence of reporting inconsistencies across a large sample of articles. We stress that statcheck is an algorithm that will, like any automated procedure exposed to real-world data, sometimes lead to false positives or false negatives. These limitations should be taken into account, preferably by manually double-checking inconsistencies detected by statcheck.

References

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 1: sensitivity and specificity. *British Medical Journal, 308*, 1552. doi:10.1136/bmj.308.6943.1552

American Psychological Association. (2010). *Publication Manual of the American Psychological Association. Sixth Edition*. Washington, DC: American Psychological Association.

Baker, M. (2015). Smart software spots statistical errors in psychology papers: One in eight articles contain data-reporting mistakes that affect their conclusions. *Nature News*. doi:http://www.nature.com/news/smart-software-spots-statistical-errors-in-psychology-papers-1.18657

Baker, M. (2016). Stat-checking software stirs up psychology. *Nature, 540*, 151–152. doi:0.1038/540151a

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

Bakker, M., & Wicherts, J. M. (2014). Outlier removal and the relation with reporting errors and quality of research. *PLoS One, 9*(7), e103360. doi:10.1371/journal.pone.0103360

Baratloo, A., Hosseini, M., Negida, A., & El Ashal, G. (2015). Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. *Emergency, 3*(2), 48-49.

Caperos, J. M., & Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema, 25*(3), 408-414. doi:10.7334/psicothema2012.207

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

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

Epskamp, S., & Nuijten, M. B. (2015). statcheck: Extract statistics from articles and recompute p values. R package version 1.0.1. http://CRAN.R-project.org/package=statcheck.

Epskamp, S., & Nuijten, M. B. (2016). statcheck: Extract statistics from articles and recompute p values. R package version 1.2.2. http://CRAN.R-project.org/package=statcheck.

Field, A. (2009). *Discovering statistics using SPSS*: Sage publications.

Hartgerink, C. H. (2016). 688,112 statistical results: Content mining psychology articles for statistical test results. *Data, 1*(3), 14.

Nuijten, M. B., Borghuis, J., Veldkamp, C. L. S., Dominguez-Alvarez, L., Van Assen, M. A. L. M., & Wicherts, J. M. (in press). Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra: Psychology*.

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

Nuijten, M. B., van Assen, M. A. L. M., Augusteijn, H. E. M., Crompvoets, E. A. V., & Wicherts, J. M. (in preparation). *Predicting Overestimated Effects in Intelligence Research*. https://osf.io/4gmrn/.

Schmidt, T. (2016). *Sources of false positives and false negatives in the STATCHECK algorithm: Reply to Nuijten et al. (2016)*. Retrieved from https://arxiv.org/abs/1610.01010.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - Or vice versa. *Journal of the American Statistical Association, 54*, 30-34. doi:10.2307/2282137

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

Veldkamp, C. L. S., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A. L. M., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS One, 9*(12), e114876. doi:10.1371/journal.pone.0114876

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