The reporting of statistical results in sociology: a systematic review

Supervisor: Marcel van Assen

19 October 2022

In this article, several aspects of statistical reporting in sociology were studied: the number of sociology journals that require adherence to the APA statistical reporting guidelines, the prevalence of statistical reporting errors, the presence of publication bias, the presence of a 'bump' in just significant p-values, and the prevalence of marginal significance among results of articles. Data were automatically retrieved from the 2014-2016 volumes of five sociology journals using the R package statcheck (Epskamp & Nuijten, 2016). Furthermore, data were retrieved manually from the 2014-2016 volumes of three sociology journals previously studied in research on publication bias by Gerber & Malhotra (2008). We found that only 13 of 143 sociology journals (9.1%) requested authors to adhere to the APA statistical reporting guidelines. Using statcheck, we found a slightly higher prevalence of inconsistent results (13.7%) and a prevalence of grossly inconsistent results comparable to previous studies in psychology (Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2016). In our manually retrieved statistical results related to explicitly stated hypotheses, 3.5% of results were inconsistent and 0.5% grossly inconsistent. No convincing evidence of publication bias and a 'bump' in p-values were found. Marginal significance was rather prevalent, with 81.5% of results with p-values in the interval (.05-.10] labelled as (marginally) significant. Implications of these results are discussed.

Keywords statcheck, publication bias, statistical reporting errors, marginal significance, statistical reporting guidelines

Statistical results in scientific articles provide scientific and nonscientific communities with essential information about studied phenomena. It is therefore vital that they live up to certain quality standards. Firstly, they should provide sufficient information for reproduction; this will make it easier for readers to critically assess the reported results of a study (Simera et al., 2011). Secondly, statistical results should not contain errors, because errors can lead to incorrect statistical conclusions, placing readers at risk of being misinformed about the nature of studied phenomena. Finally, the reporting of statistical results in articles should be standardized at least within disciplines to enable authors to clearly communicate them and readers to critically evaluate them. In this systematic review, we will examine several aspects of the quality of reporting of statistical results in sociology, namely requested adherence to statistical reporting guidelines by journals, prevalence of statistical reporting errors, evidence of publication bias, evidence of a 'bump' in just significant p-values, and prevalence of p-values reported as marginally significant.

It has been suggested that the presence of statistical reporting guidelines in a discipline may lead to less reporting errors (Lang & Altman, 2013). Their absence, on the other hand, leads to authors providing insufficient information when reporting statistics, making critical assessment of results difficult (Simera et al., 2011). Statistical reporting guidelines also increase comparability and communicability of statistical results in a discipline by providing one standard of reproducible reporting. Thus, adopting discipline-wide statistical reporting guidelines may well lead to better statistical reporting quality in a discipline. In psychology, the American Psychological Association (APA) manual and its accompanying statistical reporting guidelines serve this function to some extent. At present, they have been adopted in more than 90 psychology journals (American Psychological Association, 2022b). Unfortunately, within sociology, no statistical reporting guidelines have been developed. Different sociology journals require authors to adhere to different style manuals, such as the APA, American Sociological Association (ASA), Chicago, Harvard, and Oxford style manuals. Of these manuals, only the APA manual contains statistical reporting

guidelines. We examined which journals request authors to follow the APA manual to assess to what extent sociology journals require authors to adhere to statistical reporting guidelines.

Statistical reporting errors, also called inconsistencies, occur when there is a discrepancy between the following parameters of a reported result: the test statistic, (if used) the degrees of freedom (*df*), and the *p*-value. Inconsistencies are undesirable because they reflect inaccuracies in reported results. Gross inconsistencies are inconsistencies that change statistical conclusions based on null hypothesis significance testing (NHST). This can cause audiences to inadvertently decide a true effect exists, or that it does not exist. An example of an inconsistency is '$t(50) = 1.88,\ p = .056$', since $t(50) = 1.88$ implies $p = .066$. An example of a gross inconsistency is '$t(50) = 1.99,\ p = .049$'. This suggests a statistically significant result, but $t(50) = 1.99$ implies $p = .052$, which means H_0 should not be rejected. To our knowledge, no research on the prevalence of statistical reporting errors in sociology has previously been conducted. However, research in psychology has found that 4.3%-12.8% of results are inconsistent (Bakker & Wicherts, 2011; Krawczyk, 2015; Nuijten et al., 2016; Veldkamp, Nuijten, Dominguez-Alvarez, Van Assen, & Wicherts, 2014; Vermeulen et al., 2015; Wicherts, Bakker, & Molenaar, 2011). Gross inconsistencies make up 0.8%-2.5% of reported results in psychology (Bakker & Wicherts, 2011; Nuijten et al., 2016; Veldkamp et al., 2014; Vermeulen et al., 2015), and occur relatively often in results which are reported as significant, but are in fact non-significant. Nuijten et al. (2016) found that due to gross inconsistencies, there were 2.2 percentage points less significant recalculated *p*-values than significant reported *p*-values, while Vermeulen et al. (2015) found that 76.9% of gross inconsistencies were results erroneously reported as significant. Similarly, research in psychology found that 38.7%-67.45% of *p*-values reported as $p = .05$ had non-significant counterparts (Hartgerink, Van Aert, Nuijten, Wicherts, & Van Assen, 2016; Leggett, Thomas, Loetscher, & Nicholls, 2013). This could point to authors using the questionable research practice (QRP) of incorrect *p*-value rounding to obtain (false)

significance (Hartgerink et al., 2016; John, Loewenstein, & Prelec, 2012). We studied the prevalence of statistical reporting errors in a selection of APA[1] and non-APA sociology journals in two ways: manually, and with the R package statcheck (Epskamp & Nuijten, 2016), which automatically checks the consistency of APA-reported results.

Publication bias occurs when statistically non-significant results are published relatively less often than significant ones (Dickersin, 1990). It has been found in various scientific fields (e.g., Dickersin, 1990; Easterbrook, Gopalan, Berlin, & Matthews, 1991; Fanelli, 2010, 2011; Franco, Malhotra, & Simonovits, 2014, 2016; Kühberger, Fritz, & Scherndl, 2014; Lakens, 2015a in his reanalysis of De Winter & Dodou, 2015). Publication bias is caused by the selection of articles for publication based on significant results (Maxwell, 1981; Song et al., 2010) combined with the dependence of scientists' career success on publishing articles with significant results (Dickersin, 1990; Lawrence, 2003; Song et al., 2010). This mechanism increases the relative amount of published significant results by reducing the amount of published non-significant results (Hartgerink et al., 2016). Publication bias is problematic because an overrepresentation of significant results limits the scientific community's ability to nuance or correct previous findings (Knight, 2003). Thereby, it might inhibit scientific progress. Potential causes of a high prevalence of significant results are QRPs that lead to *p*-hacking, like rounding down *p*-values such that they become significant or conducting multiple statistical analyses and only reporting the lowest obtained *p*-value (Hartgerink et al., 2016; John et al., 2012; Ulrich & Miller, 2015). In various scientific fields, publication bias has been indicated by a low prevalence of just nonsignificant *p*-values relative to just significant ones (De Winter & Dodou, 2015; Ginsel, Aggarwal, Xuan, & Harris, 2015). In psychology, Lakens (2015b) found evidence of publication bias in his reanalysis of Masicampo & Lalande (2012), and Kühberger et al. (2014) found three times more just significant results than just nonsignificant ones in 531 psychology articles using caliper

[1]In this article, when discussing sociology journals, the term 'APA journal' does not refer to a journal which belongs to the APA organization. Instead, it refers to a journal that requires its authors to adhere to the APA manual.

tests. In sociology, Gerber & Malhotra (2008) found evidence of publication bias among results corresponding to hypotheses from the 2003-2005 volumes of American Sociological Review (*ASR*), American Journal of Sociology (*AJS*) and Sociological Quarterly (*SQ*). Like Kühberger et al. (2014), they compared numbers of just significant z-values to numbers of just non-significant z-values using caliper tests and found that the number of just significant z-values was 2.4 to 4 times higher than that of just nonsignificant z-values. Following De Winter & Dodou (2015) and Masicampo & Lalande (2012), we studied the p-value intervals (.04-.05] and (.05-.06]. Additionally, we studied the p-value intervals (.03-.05] and (.05-.07], since larger intervals provide higher power. We studied p-values rather than z-values because in this way, we could include results of a wider range of statistical analyses.

A 'bump' in p-values occurs when there are more p-values in a just statistically significant p-value interval than in the adjacent lower p-value interval. According to Hartgerink et al. (2016), it is evidence of specific QRPs that can lead to left-skewedness in the distribution of significant p-values, for instance: exclusion of outliers after having conducted analyses (Bakker & Wicherts, 2014), or using a different sample of data if the previous one(s) did not provide significant results (also called data peeking, see Armitage, McPherson, & Rowe (1969)). Across a variety of disciplines, Head, Holman, Lanfear, Khan, & Jennions (2015) found indications of a 'bump' in the p-value intervals (.045-.05) and (.025-.05). However, Hartgerink (2017) found no evidence of a 'bump' when reanalyzing this study's data with different binwidths. In psychology, some studies focusing on the p-value interval (.04-.05] also claimed to have found an overrepresentation of just significant p-values (Hartgerink et al., 2016; Leggett et al., 2013; Masicampo & Lalande, 2012). However, a reanalysis by Lakens (2015b) of Masicampo & Lalande (2012) showed that the overrepresentation of just significant p-values they found for one specific binwidth was likely coincidental. Furthermore, Simonsohn, Nelson, & Simmons (2014) and Hartgerink et al. (2016) argued that data peeking does not result in a 'bump' if true effect sizes are medium (Cohen's $d = 0.5$) or larger. Although this implies that the absence of a 'bump' is no evidence of absence

of QRPs, the presence of a 'bump' can only be explained by QRPs such as those mentioned above. Following Hartgerink et al. (2016), we studied the presence of a 'bump' using the p-value intervals (.04-.05] versus (.03-.04] and (.03-.05] versus (.01-.03]. Larger intervals were again used because they may provide higher testing power, although power may also decrease because p-values near .01 will be more prevalent than p-values near .05 in case of true nonzero effects (Hartgerink et al., 2016).

We also examined the prevalence of results reported as marginally significant in sociology. The reporting of marginally significant results occurs when authors argue that statistically non-significant results ($p > .05$) provide evidence of nonzero true effects, although one can argue they have low evidential value (Benjamin et al., 2017; Ohlsson Collentine, Van Assen, & Hartgerink, 2019; Pritschet, Powell, & Horne, 2016). Thus, assigning marginal significance may result in (unwarranted) false positives. Since this can lead to audiences assuming a true effect exists while evidence for it is slight, marginally significant p-values can be considered undesirable. P-values reported as marginally significant can mainly be found in the interval (.05-.10]; according to Pritschet et al. (2016), 92.6% of p-values reported as marginally significant in psychology were found here. Ohlsson Collentine et al. (2019) found that almost 40% of p-values in the (.05-.10] interval retrieved from the text of 44,200 articles of 70 psychology journals were reported as marginally significant. They also found that almost 20% of articles containing p-values had at least one p-value in the (.05-.10] interval that was reported as marginally significant. As for studies on assignment of marginal significance in sociology, Leahey (2005) found that in 10% of articles from two unnamed top sociology journals from 1995-2000, a significance level of low evidential value of $p < .10$ was used. In this article, we followed Ohlsson Collentine et al. (2019) by studying the prevalence of marginal significance in the p-value interval (.05-.10] at the results and article levels.

We studied statistical reporting errors, publication bias, a 'bump' in p-values, and p-values reported as marginally significant among results of explicitly stated hypotheses

(hypotheses referred to in the article's text as hypotheses to be tested) and results not related to explicitly stated hypotheses. We formally tested whether there were differences in the prevalence of three of these topics - statistical reporting errors, publication bias, and assignment of marginal significance - between results of explicitly stated hypotheses and other results. One would hope that at least reported results related to hypotheses would be without inaccuracies through careful checking by authors before submission and by reviewers and editors before accepting an article. On the other hand, publication bias has been assumed to primarily operate on results related to hypotheses (see Gerber & Malhotra, 2006). This may also be hypothesized for statistical reporting errors and marginal significance. Statistical reporting errors may be more prevalent among results related to hypotheses if publication bias and its accompanying pressure to publish positive results lead to QRPs such as (accidentally or intentionally) rounding down p-values incorrectly or adding extra zeroes (e.g., turning $p = 0.13$ into $p = 0.013$) for variables key to testing hypotheses. Assigning marginal significance may also be more prevalent in results related to hypotheses, since it allows authors to try to convince readers that there is reason to assume a proposed effect central to the article's hypotheses is a true effect, even though it is non-significant. Therefore, we expected the prevalence of (gross) inconsistencies, publication bias, and marginal significance to be higher among results corresponding to explicitly stated hypotheses. More specifically, we hypothesized the following:

Elise

Originally, H3a and H3b were one hypothesis, but I thought it might be clearer to split them into two based on bin-width.

H1: The prevalence of statistical reporting inconsistencies is higher among results of explicitly stated hypotheses than among results not related to explicitly stated hypotheses.

H2: The prevalence of gross statistical reporting inconsistencies is higher among results of explicitly stated hypotheses than among results not related to explicitly stated hypotheses.

H3a: The discrepancy between the amounts of p-values in the intervals (.04-.05] and (.05-.06] is larger among results of explicitly stated hypotheses than among results not related to explicitly stated hypotheses.

H3b: The discrepancy between the amounts of p-values in the intervals (.03-.05] and (.05-.07] is larger among results of explicitly stated hypotheses than among results not related to explicitly stated hypotheses.

H4: The prevalence of p-values in the interval (.05-.10] reported as marginally significant is higher among results of explicitly stated hypotheses than among results not related to explicitly stated hypotheses.

We did not construct similar hypotheses for a possible 'bump' in *p*-values. Due to sample sizes generally being larger in sociology than in psychology, statistical power has been suggested to be higher in sociology (Cohen, 1992; Sedlmeier & Gigerenzer, 1989). Assuming the same distribution of examined true effects in both fields, higher statistical power implies lower *p*-values on average in sociology. Therefore, we expected neither a 'bump' in *p*-values in sociology in general, nor a difference in the presence or size of a 'bump' between results related to hypotheses and results not related to hypotheses.

Method

Data sources

For our study on statistical reporting guidelines, we consulted Clarivate Analytics' Web of Science (2016) to create a data set of sociology journals called 'SRG' ('Statistical Reporting Guidelines'). For each journal in 'SRG', we verified whether it requested adherence to the APA manual - and thus, to its statistical reporting guidelines - or not.

To study statistical reporting errors, publication bias, the 'bump' in p-values, and marginal significance, we collected data from articles of several journals. Since statcheck only recalculates APA-reported results, we collected articles from two sociology journals from Clarivate Analytics' Web of Science (2016) that require APA statistical reporting: Cornell Hospitality Quarterly (*CHQ*) and Journal of Marriage and Family (*JMF*). Of sociology journals requiring APA statistical reporting, these were the ones with the highest impact factors from which statcheck could extract results (*CHQ* ranked first with 2.657, *JMF* third with 2.238)[2]. We examined all 310 articles from the 2014-2016 volumes of these journals. To compare the prevalence of statistical reporting errors in APA journals and non-APA journals, we also examined results from the 322 articles of the 2014-2016 volumes of three non-APA journals from Clarivate Analytics' Web of Science (2016): *ASR*, *AJS*, and *SQ*. Gerber & Malhotra (2008) used three volumes from these journals in their study on publication bias in sociology, which we wanted to conceptually replicate. APA-reported results retrieved by statcheck were put into a data set called 'APA', and all p-values retrieved by statcheck (APA-reported or not) were put into a data set called 'AllP', implying that 'APA' is a subset of 'AllP'. Finally, we created a data set called 'Hyp', which contains all manually retrieved p-values and statistical results related to explicitly stated hypotheses from *ASR*, *AJS*, and *SQ*. Thus, some APA-reported results are also included in 'Hyp', as

[2]Initially, we had collected articles from *CHQ* and Work and Occupations (*WOX*), which had the second highest impact factor (2.355). However, for an unknown reason, no results could be extracted by statcheck from neither the HTML nor PDF versions of *WOX* articles.

'Hyp' contains all statistical results related to explicitly stated hypotheses.

Data collection[3]

For each sociology journal in Clarivate Analytics' Web of Science (2016), we verified if it explicitly required authors to adhere to the APA, ASA, Chicago and/or Harvard style guide and/or another external style guide. We also examined if journals explicitly required authors to follow their own journal's style guide, and if they allowed authors to follow different style guides. This information was put into data set 'SRG.'[4] There was explicitly required adherence to the own journal's guidelines if one of the following expressions was found on the journal's website: 1) 'House style (guide) X' or 'Journal style (guide) X', where X represents the journal's name, or 2) 'X (format) requirements' or 'X (format) requirements', where again X represents the journal's name. If some form of style guidelines was available, but there was no explicitly named style guide, a journal was put into the category 'Other'.

Before extracting statistical data with statcheck, we converted all relevant articles to HTML format. Statcheck namely converts HTML or PDF files to plain text before extracting statistics, and conversion from HTML format is accompanied by less errors (Nuijten et al., 2016). We then applied statcheck's 'checkHTMLdir' function to a folder with HTML files to automatically obtain APA-reported results and recalculated p-values. Data set 'APA' contains information retrieved by statcheck on all aspects of APA-reported results from all five journals: test statistics (t, z, F, χ^2, and r), df, and p-values reported using '=', '<', '>', or 'non-significant'. If p-values were reported as non-significant, statcheck assigned them the label 'NA'. 'APA' also contains p-values recalculated by statcheck, and information from statcheck on whether reported results are (grossly) inconsistent with their

[3]R code used to complete the data sets on results-related topics is available at https://github.com/elisecj94/thesis/tree/development/Code

[4]'SRG' is available at https://github.com/elisecj94/thesis/tree/development/Data/SRG

recalculated counterparts.[5] If a reported result seemed inconsistent (and this could not be due to correct rounding), statcheck applied a one-sided test to it. If this led to a consistent reported result, statcheck kept the one-sided test. Otherwise, it kept the two-sided test (Nuijten et al., 2017). We also manually put the part of the article's text from which we concluded that a result was (not) related to an explicitly stated hypothesis in a separate column. Our definition of explicitly stated hypotheses followed that of Gerber & Malhotra (2008), i.e., hypotheses were considered explicitly stated if they were bolded, italicized, or indented, or if they were listed using one of the following terminologies: 'Hypothesis 1', 'H1', 'H_1', or 'the first hypothesis'.

Data set 'APA' was used to test our hypotheses on statistical reporting errors (H1 and H2). Of 524 retrieved statistical results, we removed 19 (3.6%) because they did not refer to APA-reported results. In total, 505 statistical results from 76 articles were used in descriptive analyses and hypothesis testing (see Table 1 and Table 2).

Table 1: Overview of information provided by 'AllP', 'APA', and 'Hyp'.

	'AllP'	'APA'	'Hyp'
Journals	all	all	*ASR*/*AJS*/*SQ*
Part(s) of article from which info was retrieved	text	text	text/table/figure
Statistical results related to hypotheses?	partly	partly	yes
Total number of articles	471	80	91
Number of articles used	314	76	91
Total number of statistical results	7,280	524	4,929
Number of valid statistical results	2,960	505	4,929

[5]'APA' is available at https://github.com/elisecj94/thesis/tree/development/Data/APA

Table 2: Overview of numbers of statistical results and accompanying articles used in analyses of results-related topics for 'AllP', 'APA', and 'Hyp'.

	'AllP'	'APA'	'Hyp'
Statistical reporting errors			
Descriptive information	-	505 (76)	404 (19)
Testing hypotheses (gross) inconsistencies (H1 & H2)	-	505 (76)	-
Publication bias			
Descriptive information			
(.04-.05] - (.05-.06]	73 (50)	-	14 (7)
(.03-.05] - (.05-.07]	127 (71)	-	26 (11)
Testing hypotheses publication bias			
H3a: (.04-.05] - (.05-.06]	73 (50)	-	-
H3b: (.03-.05] - (.05-.07]	127 (71)	-	-
Bump in *p*-values			
Descriptive information			
(.03-.04] - (.04-.05]	64 (40)	-	14 (7)
(.01-.03] - (.03-.05]	184 (80)	-	37 (12)
Marginal significance			
Descriptive information	199 (107)	-	130 (30)
Testing hypothesis marginal significance (H4)	199 (107)	-	-

Note. Numbers of articles from which results were used in analyses are shown between parentheses.

The third data set, 'AllP', consists of all reported *p*-values retrieved by statcheck from all five journals[6]. We manually added information on whether reported *p*-values were related to an explicitly stated hypothesis as we did for 'APA'. Of 7,280 results retrieved by

[6]'AllP' is available at https://github.com/elisecj94/thesis/tree/development/Data/AllP

statcheck from 471 articles, we removed 4,320 (59.3%) because they did not refer to reported p-values[7]. After this, 'AllP' contained 2,960 reported p-values from 314 articles (see Table 1). From these data, descriptive information on publication bias, the 'bump' in p-values, and marginal significance was obtained. Furthermore, 'AllP' was used to test H3a, H3b, and H4, and to study the prevalence of assignment of marginal significance (see Table 2). To determine if marginal significance was assigned to a reported p-value, we looked up p-values in the (.05-.10] interval in the text of articles. Then, following Ohlsson Collentine et al. (2019), we decided that a p-value from 'AllP' was assigned marginal significance by authors if the expressions 'margin*' or 'approach*' were mentioned in relation to its significance. The text used to conclude that a p-value was (not) assigned marginal significance was stored manually in a separate column of 'AllP'. Finally, we obtained descriptive statistics on the number of articles with at least one p-value in the interval (.05-.10] to which marginal significance was assigned.

A fourth data set, 'Hyp', was created to conceptually replicate the study of Gerber & Malhotra (2008) on publication bias by manually retrieving results from articles[8]. Manual retrieval allows one to retrieve information from tables, figures, and text, whereas statcheck can only retrieve information from text. We only collected data from articles that met our inclusion criteria. Like Gerber & Malhotra (2008), we only studied articles that explicitly stated one or more hypotheses before their results were presented. Of the 322 articles from *ASR*, *AJS* and *SQ*, 99 (30.7%) met this criterion. Furthermore, articles had to contain at least one 'required statistic', i.e., at least one p-value or reproducible result related to an explicitly stated hypothesis. This was the case for 91 articles (28.3%) (see Figure 1 for an overview of the selection process)[9]. Following Gerber & Malhotra (2008), 'Hyp' contains

Elise

However, I mention several differences between our approach and that of Gerber and Malhotra below. Should I drop the term 'replicate' here, since it is not strictly a replication?

Marcel

Maak er dan maar 'conceptually' van

[7]The removed results corresponded to, for instance, duplicates of p-values that had already been extracted, and p-values that were not related to a specific result but were mentioned in the article to indicate which significance levels were used. There was also one case in which 'Ns' was extracted while meaning 'numbers' instead of 'non-significant'.

[8]'Hyp' is available at https://github.com/elisecj94/thesis/tree/development/Data/Hyp

[9]The data set used to select articles for 'Hyp' is available at https://github.com/elisecj94/thesis/tree/development/Data/Hyp

all statistics from all models that were essential to testing explicitly stated hypotheses. Since results corresponding to control variables are not the focal point in hypothesis testing, they were not included in the data set. Whether individual statistics were essential for hypothesis testing or not was determined by reading the article's hypotheses and the part(s) of the article's text in which the results of hypothesis testing were discussed. Furthermore, following Gerber & Malhotra (2008), information from appendices was also included, but information from supplements was not, since only appendices are part of articles as published.

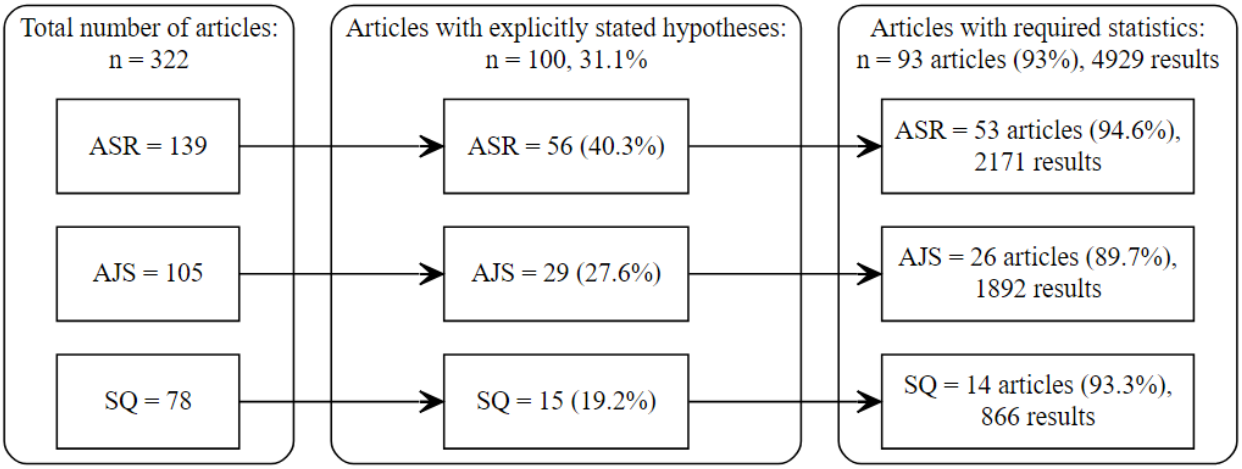

Figure 1: Flowchart describing the process of selecting articles from which results were retrieved for 'Hyp'.

There are some differences between our study and that of Gerber & Malhotra (2008). Gerber & Malhotra (2008) studied publication bias using caliper tests on z-values and t-values (converted to z-values) within 5%, 10%, 15% or 20% of $z = 1.64$ (one-sided testing) and $z = 1.96$ (two-sided testing). If z-values or t-values were unavailable, regression coefficients and standard errors were used to calculate z-values. We used only exactly reported (not recalculated) p-values in the intervals (.04-.06] and (.03-.07] to study publication bias instead.[10] One reason for this is that it was often unknown what kind of

[10]It should be noted that these p-value intervals largely overlap with the 5% and 10% caliper tests of Gerber & Malhotra (2008). E.g., for two-sided tests, the 5% caliper had an equivalent p-value interval of (.040-.063), while the 10% caliper had an equivalent p-value interval of (.031-.077].

distribution an analysis was based on. Furthermore, this allowed us to also include *p*-values based on *F*-values, *r*-values, and χ^2-values. We did not mix reported and recalculated *p*-values in our analyses because Krawczyk (2015) and Hartgerink et al. (2016) found that there can be differences in reported and recalculated *p*-value distributions around $p = .05$, and Hartgerink et al. (2016) have argued that these two types of *p*-values should therefore not be mixed in studies on reporting practices. Finally, Gerber & Malhotra (2008) excluded articles with more than 38 relevant coefficients because their inclusion could have a disproportionate impact on analyses. We did not do so, since we wanted to include all *p*-values relevant for studying publication bias. If one or more articles would influence the results disproportionately, we would do extra analyses without these articles.

Elise

Which I did not do. I could still do this, but I don't know if that's what we want. Maybe discuss in next meeting.

We organized all available aspects of a result of an explicitly stated hypothesis - *p*-value, regression coefficient (or odds ratio, proportional hazard, etc.), *z*-value, *t*-value, *F*-value, *r*-value, χ^2-value, standard error, sample size, *df*, phrasing of the hypothesis a result belonged to as retrieved from the article, and, if applicable, text from the article in which a result was mentioned - as we did for 'APA'. In total, 'Hyp' contained 4,929 results (see Table 1). It was used to study all our results-related phenomena of interest (see Table 2). Where possible, we checked whether statistical results were (grossly) inconsistent by recalculating their *p*-values using the following functions from the R stats package (R Core Team, 2013): pt() for *t*-values and *r*-values[11], pnorm() for *z*-values, pf() for *F*-values, and pchisq() for χ^2-values. For detailed information on how this was done, see Table 3 and Table 4. We also manually added information to 'Hyp' on assignment of marginal significance to in-text *p*-values in the (.05-.10] interval as we did for 'AllP'. For *p*-values in captions of tables and figures, we considered significance levels of $p < .10$ (indicated by, e.g., an asterisk) to be assignment of marginal significance. Finally, we studied the percentage of articles in

[11] In order to use the pt() function for results with *r*-values, *r*-values first had to be converted to *t*-values with the following formula:

$$t = \frac{r\sqrt{df}}{\sqrt{1 - r^2}}$$

where $df = n$ - 2.

'Hyp' containing marginally significant results in the p-value interval (.05-.10]. Note that 'AllP', 'APA', and 'Hyp' overlap. For instance, an in-text APA-reported result related to an explicitly stated hypothesis is included in all three data sets.

Elise

Final two phrases may be repetitive, since this can also be understood from earlier writings?

Table 3: Inconsistencies and gross inconsistencies as determined for different types of reported p-values in 'Hyp'.

Type of Rep$P$	Inconsistent if...	Grossly inconsistent if...
ns	Cal$P \leq .05$	Cal$P \leq .05$
$<$	Cal$P \geq$ Rep$P$	Cal$P > .05$ & Rep$P \leq .05$
$\geq$	Cal$P <$ Rep$P$	Cal$P < .05$ & Rep$P \geq .05$
$=$	Cal$P \neq$ Rep$P$ not due to rounding*	Cal$P \neq$ Rep$P$ not due to rounding, Cal$P \leq .05$ & Rep$P > .05$ or vice versa

Note. Cal$P$ = recalculated p-value, Rep$P$ = reported p-value, ns = non-significant.

* See Table 4 for methods used to determine whether a difference between recalculated and reported p-values could be due to correct rounding or not.

Table 4: Ways of determining whether discrepancies between exactly reported p-values and their recalculated counterparts from 'Hyp' could be due to correct rounding or indicate an inconsistency.

b & SE

Only used for recalculation if a result was explicitly based on the z-distribution or t-distribution. Take, e.g., $b = 3.11$, SE $= 2.11$, $p = 0.07$, from a z-distribution:

- Correct Cal$P$ stem from $b_{\text{lb}} \leq b < b_{\text{ub}}$ (e.g., $3.105 \leq b < 3.115$) and $\text{SE}_{\text{lb}} \leq \text{SE} < \text{SE}_{\text{ub}}$ (e.g., $2.105 \leq \text{SE} < 2.115$).
- Calculate $t/z_{\text{ub}} = \frac{b_{\text{ub}}}{SE_{\text{lb}}}$ and $t/z_{\text{lb}} = \frac{b_{\text{lb}}}{SE_{\text{ub}}}$, the largest and smallest t/z consistent with b and SE (e.g., $z_{\text{ub}} = \frac{3.115}{2.105} = 1.47981$ and $z_{\text{lb}} = \frac{3.105}{2.115} = 1.468085$).
- Calculate $t/z_{\text{lb}} = \text{Cal}P_{\text{ub}}$ and $t/z_{\text{ub}} = \text{Cal}P_{\text{lb}}$, i.e., boundaries of correctly

Table 4: Ways of determining whether discrepancies between exactly reported p-values and their recalculated counterparts from 'Hyp' could be due to correct rounding or indicate an inconsistency. *(continued)*

	rounded Rep$P$. For this, the R stats package pt() function (for t) or the pnorm() function (for z) is used. - Round CalP_{lb} and CalP_{ub} to the same number of decimals as Rep$P$ with R base round() function. In our example, Cal$P_{\text{lb}} \approx 0.07$ and Cal$P_{\text{ub}} \approx 0.07$. - If Cal$P_{\text{lb}} \leq$ Rep$P \leq$ CalP_{ub}, Rep$P$ is considered correct. This is the case in our example.
test statistics	Functions of the R stats package used to recalculate p-values: pt() for t and r, pnorm() for z, pf() for F, and chisq() for χ^2. All functions, except pnorm(), require *df*. We use the example of $t(61) = 3.11$, $p = 0.0001$: - Correct Cal$P$ stem from $t_{\text{lb}} \leq t < t_{\text{ub}}$ (e.g., $3.105 \leq t < 3.115$). - Calculate the p-values consistent with the highest and lowest t-values possible under correct rounding with pt() function. - Round CalP_{lb} and CalP_{ub} to the same number of decimals as Rep$P$ with R base round() function. In our example, Cal$P_{\text{lb}} \approx 0.001$ and Cal$P_{\text{ub}} \approx 0.001$. - If Cal$P_{\text{lb}} \leq$ Rep$P \leq$ CalP_{ub}, Rep$P$ is considered correct. This is not the case in our example, since Rep$P <$ Cal$P_{\text{lb}} <$ CalP_{ub}.

Note. Rep$P$ = reported p-value, Cal$P$ = recalculated p-value, CalP_{lb} = lower bound recalculated p-value, CalP_{ub} = upper bound recalculated p-value, b_{lb} = lower bound b b_{ub} = upper bound b, SE$_{\text{lb}}$ = lower bound SE, SE$_{\text{ub}}$ = upper bound SE.

Statistical analyses

In our descriptive analyses (which consist of frequencies and percentages), we reported how many journals from 'SRG' require authors to adhere to the APA statistical reporting guidelines. For (gross) inconsistencies, descriptive statistics were based on reproducible results from 'APA' and 'Hyp'. We followed Vermeulen et al. (2015) and Nuijten et al. (2016) by studying the direction of gross inconsistencies: do errors make non-significant results significant, or vice versa? For publication bias and the 'bump' in p-values, descriptive results were based on exactly reported p-values from 'AllP' and 'Hyp'. Inexactly reported p-values were excluded for these topics because, as Hartgerink et al. (2016) has shown for a 'bump' in p-values, they can cause to 'spikes' in certain p-values (e.g., including $p < .05$ will likely lead to a spike at $p = .05$). Data sets 'AllP' and 'Hyp' also provided descriptive statistics at the results and article level for marginal significance. For all topics but statistical reporting guidelines, descriptive results were split by explicitly stated hypothesis (yes/no), journal (*ASR*, *AJS*, *SQ*, and, for results from 'APA' and 'AllP', *CHQ*, and *JMF*), and year (2014-2016).

Nuijten et al. (2017) have argued that the prevalence of (gross) inconsistencies can be studied in three ways. Firstly, one can calculate the percentage of inconsistencies and gross inconsistencies for each article and take the average of these percentages over all articles. Secondly, one can calculate the overall percentage of (gross) inconsistencies by dividing the amount of (gross) inconsistencies by the total number of reported results obtained. This is what we have done in our descriptive analyses. Finally, Nuijten et al. (2017) wrote that one can use multilevel logistic regression models to estimate the probability that a reported result is inconsistent, while controlling for the nesting of results within articles. Although this method is in theory statistically sound, simulation analyses revealed that it performs poorly; because both the number of results per article and the probability of a gross inconsistency are too low, it is accompanied by a too low Type I error, a lack of statistical power, and clearly inaccurate effect size estimates (Nuijten et al., 2017). Therefore, following Wicherts

et al. (2011) and Nuijten et al. (2016), we tested our hypotheses on statistical reporting errors (H1 and H2) using logistic regressions.

We also conducted logistic regressions to test our hypotheses on publication bias (H3a, H3b) with exactly reported p-values from 'AllP' as the dependent variable. Since statcheck interprets results with $p = .05$ as being statistically significant (Epskamp & Nuijten, 2016), $p = .05$ was included in the interval of just significant p-values for the logistic regressions. To test our hypothesis on p-values reported as marginally significant (H4), we conducted logistic regressions with exactly reported p-values in the interval (.05-.10] from 'AllP' as the dependent variable. All logistic regression analyses contained a binary predictor indicating whether a result was related to an explicitly stated hypothesis or not. We chose not to include other potentially relevant control variables, such as journal and year of publication, because some analyses had too little data for including multiple predictors .

Marcel

Dit snap ik niet Elise!! Maar dan heb je toch geen probleem met nullen, zonder covariaten? Dus je moet dan tekst in comment sowieso nog aanpassen... Enfin, ik ga weer verder in de tekst.

Results

In this section, we start by presenting our results regarding statistical reporting guidelines. Next, results on statistical reporting errors, publication bias, the 'bump' in p-values, and marginal significance are discussed. For each results-related topic, we first present automatically retrieved descriptive results and (if applicable) results of hypothesis testing. Then, we discuss descriptive statistics of results related to explicitly stated hypotheses from 'Hyp'. Results on specific years and journals that were of little theoretical interest or were based on too little data are not discussed in the text but can be found in the corresponding tables. Full tables of the results of logistic regressions used to test our hypotheses can be found in the supplement.

Elise

That's true, when it comes to separation. However, we still have a low EPV for gross inconsistencies (8 events in total). We could, of course, mention the EPV issue in the discussion section.

Statistical reporting guidelines

Of the 143 sociology journals in 'SRG', one journal (*Society*) did not seem to have any guidelines authors are explicitly required or allowed to follow when preparing their manuscripts. Four journals (2.8%) explicitly required authors to follow guidelines established by the journal itself, and 102 (71.3%) required authors to adhere to (reference) guidelines established by external organizations. Only 13 journals (9.1%) requested authors to adhere to the APA manual, and thereby, to the APA statistical reporting guidelines. See Table 5 for an overview of the numbers of sociology journals requesting/allowing adherence to different statistical reporting guidelines.

Marcel

Voor discussie over guidelines: het is niet helemaal duidelijk nog of andere guidelines dan APA misschien OOK wel heel goed zijn (voor reproducibility en checking). Dat zouden we nog in een zin in de discussie kunnen/moeten toevoegen (als dat nog niet is gedaan), denk ik

Elise

Mentioned this in the discussion section: they will not promote reproducibility and checking because none of the other guidelines say anything specifically about statistical reporting guidelines.

Table 5: Numbers and percentages of sociology journals in 'SRG' requesting/allowing adherence to different types of statistical reporting guidelines.

	Number of journals (% of total)
Required	
APA	
Full manual	10 (7%)
Only references	10 (7%)
ASA	
Full manual	12 (8.4%)
Only references	3 (2.1%)
Chicago	
Full manual	7 (4.9%)
Only references	6 (4.2%)
Harvard	
Full manual	2 (1.4%)
Only references	9 (6.3%)
Oxford	1 (0.7%)
Style Manual for Authors, Editors and Printers	1 (0.7%)
Wiley	1 (0.7%)
Other	34 (23.8%)

Table 5: Numbers and percentages of sociology journals in 'SRG' requesting/allowing adherence to different types of statistical reporting guidelines. *(continued)*

	Number of journals (% of total)
Own	4 (2.8%)
Multiple options (one must be chosen)	1 (0.7%)
Multiple required	37 (25.9%)
Multiple required (one is full APA manual)	3 (2.1%)
Other, namely...	
Multiple allowed	1 (0.7%)
None mentioned	1 (0.7%)
Unknown*	1 (0.7%)
Total	143 (100%)

* We were unable to find which guidelines authors publishing in the journal *Society* are required or allowed to use. The link on the journal's website that should have provided access to this information gave a 'page not found' error.

Statistical reporting errors

Of the 505 'APA' results, 69 (13.7%) were inconsistent and 8 (1.6%) grossly inconsistent (see Table 6). All grossly inconsistent results had a statistically significant reported p-value and a non-significant recalculated p-value. Out of 168 results related to explicitly stated hypotheses, 22 (13.1%) were inconsistent and 4 (2.4%) grossly inconsistent. Out of 337 results not related to explicitly stated hypotheses, 47 (13.9%) were inconsistent and 4 (1.2%) grossly inconsistent. Of the recalculated p-values from 'APA', 416 (82.4%) were retrieved from the two APA journals. These journals, *JMF* and *CHQ*, had comparable percentages of inconsistencies (14.6% and 14.7%, respectively) and gross inconsistencies (1.6% and 1.7%, respectively). We found lower prevalences of inconsistencies among automatically retrieved results for non-APA journals *ASR* and *AJS* (2.3% and 4.9%, respectively). Our hypotheses that less (gross) inconsistencies would be observed for results on explicitly stated hypotheses

Elise
Moved this here so that I wouldn't have to mention these statistics in the discussion.

were not confirmed. As for H1, the odds of a result of an explicitly stated hypothesis being inconsistent were 1.076 times smaller than the odds that a result not related to an explicitly stated hypothesis was inconsistent, b = -.073, p = .793, OR = .930, 95% CI [.531, 1.585]. Regarding H2, the odds of a result of an explicitly stated hypothesis being grossly inconsistent were 2.030 two times larger than the odds that a result not related to an explicitly stated hypothesis was grossly inconsistent, but this difference was not statistically significant, b = .708, p = .321, OR = 2.030, 95% CI [.475, 8.685]. Full results of the logistic regressions used to test H1 and H2 can be found in Table S1. Among recalculated p-values from 'Hyp', 14 out of 404 were inconsistent (3.5%), and 2 (0.5%) were grossly inconsistent. Again, the grossly inconsistent results had a statistically significant reported p-value and a nonsignificant recalculated p-value. For a comprehensive overview of statistical reporting errors at the article and results level, see Table S4.

Table 6: Descriptive statistics on (gross) inconsistencies for 'APA' and 'Hyp'.

	Articles	Results	Inconsistencies	Gross inconsistencies
'APA'				
Relation to hypothesis				
Yes	29	168	22 (13.1%)	4 (2.4%)
No	68	337	47 (13.9%)	4 (1.2%)
Journal				
ASR	7	43	1 (2.3%)	1 (2.3%)
AJS	3	41	2 (4.9%)	0 (0%)
SQ	2	5	5 (100%)	0 (0%)
JMF	36	185	27 (14.6%)	3 (1.6%)
CHQ	28	231	34 (14.7%)	4 (1.7%)
Year				
2014	20	172	22 (12.8%)	1 (0.6%)
2015	21	136	15 (11.0%)	3 (2.2%)
2016	35	197	32 (16.2%)	4 (2.0%)
Total	76	505	69 (13.7%)	8 (1.6%)
'Hyp'				
Journal				
ASR	10	331	11 (3.3%)	1 (0.3%)
AJS	7	68	1 (1.5%)	1 (1.5%)
SQ	2	5	2 (40.0%)	0 (0%)
Year				
2014	11	312	11 (3.5%)	1 (0.3%)
2015	3	33	0 (0%)	0 (0%)
2016	5	59	3 (5.1%)	1 (1.7%)
Total	19	404	14 (3.5%)	2 (0.5%)

Note. The numbers of articles for results (not) related to explicitly stated hypotheses include articles with ≥ 1 result that is (not) related to an explicitly stated hypothesis.

Publication bias

In 'AllP', there was no evidence of publication bias (see Figure 2.1 and Table 7). Overall, for binwidth .01, 32 out of 73 results were just significant (43.8%), and for binwidth .02, 64 out of 127 results were just significant (50.4%) (Figures 2.1A and 2.1B). Splitting data among results not related to hypotheses (Figures 2.2A and 2.2B) and results that were related to hypotheses (Figures 2.3A and 2.3B), percentages were also around 50. Hence, we could not reject H_0 for our hypotheses on publication bias (H3a and H3b). For H3a, there were $\frac{1}{.877} \approx 1.140$ times less just significant p-values among results of explicitly stated hypotheses than among results not related to explicitly stated hypotheses for binwidth .01, b = -.132, p = .794, OR = .877, 95% CI [.321, 2.345]. In case of H3b for binwidth .02, H_0 could not be rejected either, since b = -.251, p = .521, OR = .778, 95% CI [.358, 1.674] (full results of the logistic regressions used to test H3a and H3b can be found in Table S2). Similarly, for 'Hyp', no evidence of publication bias was found among results related to explicitly stated hypotheses; we found (slightly) more just significant p-values than just nonsignificant ones, namely 9 out of 14 results (64.3%) for binwidth .01, and 14 out of 26 (53.8%) for binwidth .02 (see Figure 3 and Table 7).

Table 7: Descriptive statistics on publication bias for 'AllP' and 'Hyp'.

	Binwidth .01			Binwidth .02		
	(.04-.05]	(.05-.06]	Total	(.03-.05]	(.05-.07]	Total
'AllP'						
Relation to hypothesis						
Yes	10 (41.7%)	14	24	17 (45.9%)	20	37
No	22 (44.9%)	27	49	47 (52.2%)	43	90
Journal						

Table 7: Descriptive statistics on publication bias for 'AllP' and 'Hyp'. *(continued)*

	(.04-.05]	(.05-.06]	Total	(.03-.05]	(.05-.07]	Total
ASR	8 (40%)	12	20	11 (37.9%)	18	29
AJS	3 (42.9%)	4	7	8 (57.1%)	6	14
SQ	1 (20%)	4	5	3 (42.9%)	4	7
JMF	14 (43.8%)	18	32	28 (49.1%)	29	57
CHQ	6 (66.7%)	3	9	14 (70%)	6	20
Year						
2014	13 (44.8%)	16	29	27 (55.1%)	22	49
2015	10 (52.6%)	9	19	20 (52.6%)	18	38
2016	9 (36%)	16	25	17 (42.5%)	23	40
Total	32 (43.8%)	41	73	64 (50.4%)	63	127
'Hyp'						
Journal						
ASR	2 (40%)	3	5	5 (45.5%)	6	11
AJS	3 (100%)	0	3	4 (80%)	1	5
SQ	4 (66.7%)	2	6	5 (50%)	5	10
Year						
2014	4 (66.7%)	2	6	8 (61.5%)	5	13
2015	1 (50%)	1	2	1 (50%)	1	2
2016	4 (66.7%)	2	6	5 (45.5%)	6	11
Total	9 (64.3%)	5	14	14 (53.8%)	12	26

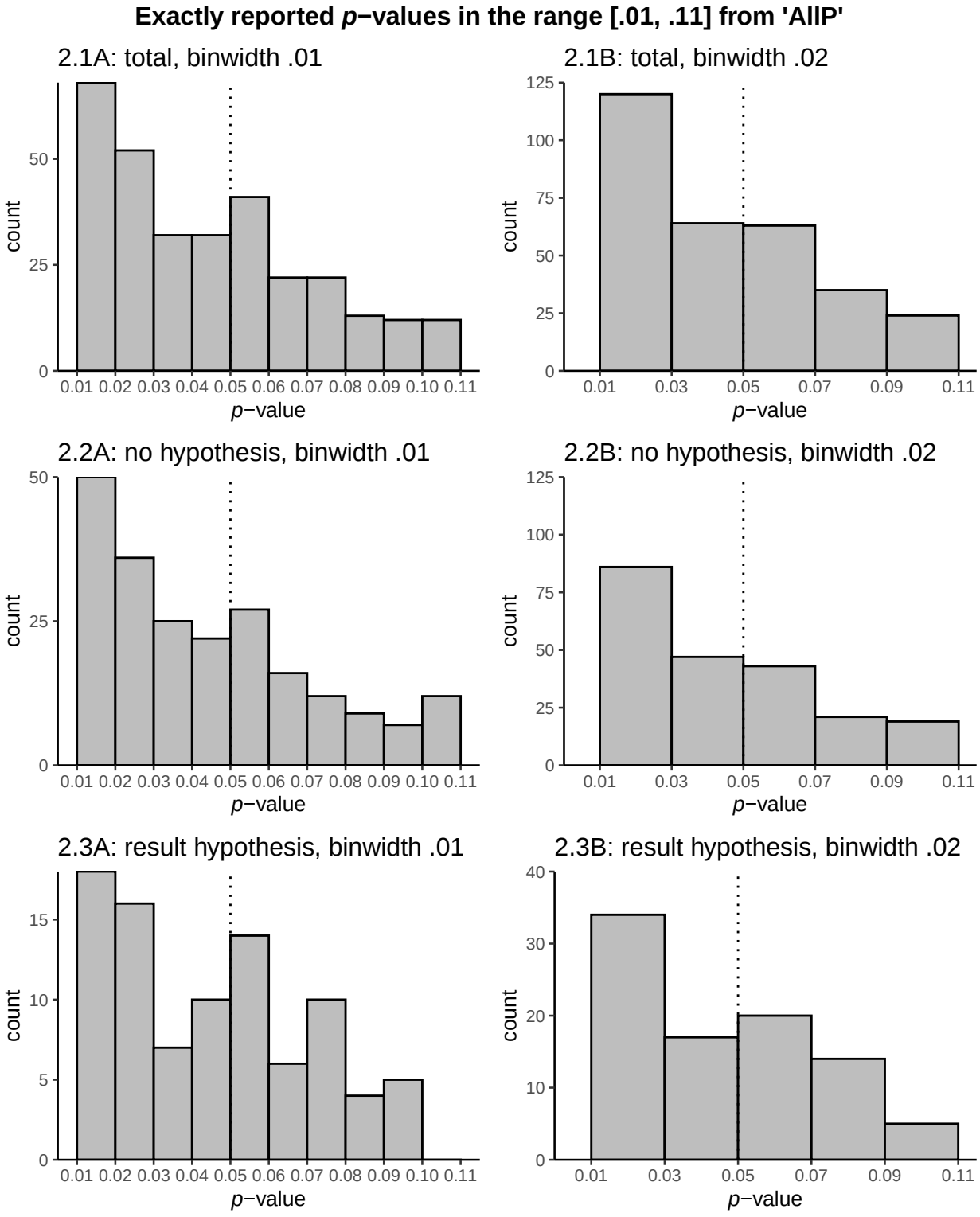

Figure 2: Histograms with binwidths .01 and .02 of exactly reported p-values in the range [.01-.11] from 'AllP'. Specifically, information is provided for the totals of exactly reported p-values (2.1A & 2.1B, for which $n = 331$), as well as for exactly reported p-values not related to explicitly stated hypotheses (2.2A & 2.2B, for which $n = 234$) and exactly reported p-values related to explicitly stated hypotheses (2.3A & 2.3B, for which $n = 97$).

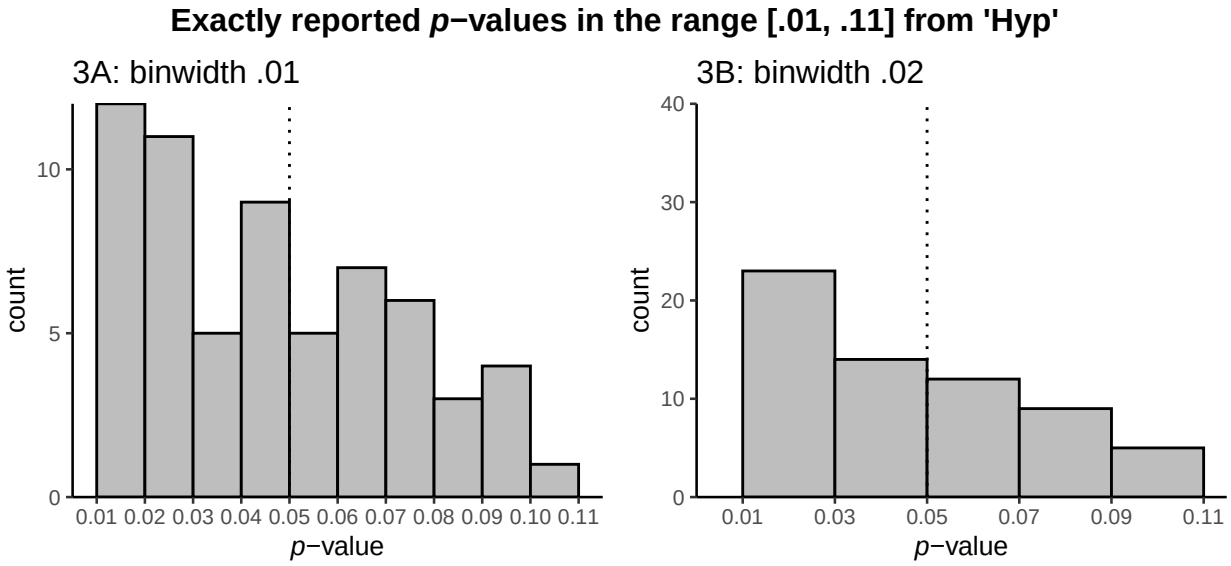

Figure 3: Histograms with binwidths .01 (3A) and .02 (3B) of exactly reported p-values related to explicitly stated hypotheses in the range [.01-.11] from 'Hyp' ($n = 67$).

Bump in p-values

Marcel: Ook deze kan efficienter en ik denk ook iets overzichtelijker.

Elise: Tried to improve!

Overall, Table 8, Figure 2, and Figure 3 show no 'bump' in p-values in 'AllP' or 'Hyp'. Using binwidth .01, lower p-value intervals contained 32 out of 64 p-values (50%) for 'AllP' overall (Figure 2.1A), 22 out of 47 (46.8%) (Figure 2.2A) for results not related to hypotheses from 'AllP', 10 out of 17 for results related to hypotheses from 'AllP' (Figure 2.3A), and 9 out of 14 (64.3%) for results related to hypotheses from 'Hyp' (Figure 3A). Correspondingly, for binwidth .02, the lower p-value intervals contained contained 64 out of 184 (34.8%) (Figure 2.1B), 47 out of 133 (35.3%)(Figure 2.2B), 17 out of 51 (33.3%) (Figure 2.3B), and 14 out of 37 (37.8%) results (Figure 3B), respectively.

Table 8: Descriptive statistics on the bump in just significant results for 'AllP' and 'Hyp'.

	Binwidth .01			Binwidth .02		
	(.03-.04]	(.04-.05]	Total	(.01-.03]	(.03-.05]	Total
'AllP'						
Relation to hypothesis						
Yes	7	10 (56.2%)	17	34	17 (33.3%)	51
No	25	22 (46.8%)	47	86	47 (35.3%)	133
Journal						
ASR	3	8 (72.7%)	11	25	11 (30.6%)	36
AJS	5	3 (37.5%)	8	23	8 (25.8%)	31
SQ	2	1 (33.3%)	3	1	3 (75%)	4
JMF	14	14 (50%)	28	44	28 (38.9%)	72
CHQ	8	6 (42.9%)	14	27	14 (34.1%)	41
Year						
2014	14	13 (48.1%)	27	51	27 (34.6%)	78
2015	10	10 (50%)	20	41	20 (32.8%)	61
2016	8	9 (52.9%)	17	28	17 (37.8%)	45
Total	32	32 (50%)	64	120	64 (34.8%)	184
'Hyp'						
Journal						
ASR	3	2 (40%)	5	9	5 (35.7%)	14
AJS	1	3 (75%)	4	11	4 (26.7%)	15
SQ	1	4 (80%)	5	3	5 (62.5%)	8
Year						
2014	4	4 (50%)	8	14	8 (36.4%)	22
2015	0	1 (100%)	1	7	1 (12.5%)	8
2016	1	4 (80%)	5	2	5 (71.4%)	7
Total	5	9 (64.3%)	14	23	14 (37.8%)	37

Marginal significance

In 46 out of 107 articles from 'AllP' (43%) with reported p-values in the interval (.05-.10], at least one p-value in this interval was reported as marginally significant (see Table 9). Out of 206 'AllP' results with reported p-values in the interval (.05-.10], 72 (35%) were reported as marginally significant. For results not related to hypotheses and results related to hypotheses, this was the case for 52 out of 136 (38.2%) and 20 out of 70 (28.6%) p-values in the interval (.05-.10], respectively. Among journals, the prevalence of p-values to which marginal significance was assigned was highest in *CHQ* (11 out of 21 results, or 52.4%) and lowest in *AJS* (4 out of 31 results, or 12.9%). The prevalence of assignment of marginal significance to p-values was comparable between years (33.8%-35.7%). Our hypothesis that assignment of marginal significance is less prevalent among results related to explicitly stated hypotheses in 'AllP' (H4) is not confirmed, b = -.437, p = .170, OR = .646, 95% CI [.342, 1.194]. Full results of the logistic regression used to test H4 can be found in Table S3.

Table 9 also shows the prevalence of marginal significance in 'Hyp'. In 19 of the 30 articles with p-values in the interval (.05-.10] from 'Hyp' (63.3%), at least one p-value in this interval was reported as marginally significant. At the results level, Table 9 shows that overall, 106 of 130 (81.5%) p-values in the interval (.05-.10] were reported as marginally significant. Reporting results as marginally significant was most prevalent in *AJS* (72 out of 76 results, or 94.7%) and least prevalent in *SQ* (5 out of 12 results, or 41.7%). Among the different years, reporting results as marginally significant was most prevalent in 2015 (71 out of 74 results, or 95.9%) and least prevalent in 2014 (11 out of 24 results, or 45.8%).

Interestingly, the percentage of p-values in interval (.05-.10] reported as marginally significant is much higher for manually retrieved statistical results related to explicitly stated hypotheses in 'Hyp' (81.5%) than for reported p-values in 'AllP' (35%) (z = 8.33, $p < .001$, z-test for two independent proportions).

Marcel

Ook iets over zeggen in discussie

Elise

Done

Table 9: Descriptive statistics on marginal significance for 'AllP' and 'Hyp'.

	Article level			Results level		
	Yes	No	Total	Yes	No	Total
'AllP'						
Relation to hypothesis						
Yes				20 (28.6%)	50	70
No				52 (38.2%)	84	136
Journal						
ASR	11 (44.0%)	14	25	19 (32.8%)	39	58
AJS	3 (20.0%)	12	15	4 (12.9%)	27	31
SQ	3 (37.5%)	5	8	4 (36.4%)	7	11
JMF	23 (47.9%)	25	48	34 (40%)	51	85
CHQ	6 (54.5%)	5	11	11 (52.4%)	10	21
Year						
2014	11 (35.5%)	20	31	19 (35.2%)	35	54
2015	17 (44.7%)	21	38	30 (35.7%)	54	84
2016	18 (47.4%)	20	38	23 (33.8%)	45	68
Total	46 (43.0%)	61	107	72 (35.0%)	134	206
'Hyp'						
Journal						
ASR	8 (72.7%)	3	11	29 (69.0%)	13	42
AJS	9 (60.0%)	6	15	72 (94.7%)	4	76
SQ	2 (50.0%)	2	4	5 (41.7%)	7	12
Year						
2014	6 (46.2%)	7	13	11 (45.8%)	13	24

Table 9: Descriptive statistics on marginal significance for 'AllP' and 'Hyp'. *(continued)*

	Yes	No	Total	Yes	No	Total
2015	6 (85.7%)	1	7	71 (95.9%)	3	74
2016	7 (70.0%)	3	10	24 (75.0%)	8	32
Total	19 (63.3%)	11	30	106 (81.5%)	24	130

Discussion

In this article, we studied different aspects of statistical reporting in sociology. Statistical reporting is of high quality if results are reproducible and do not contain errors, and if clear communication and critical evaluation of results within a discipline is possible due to standardized reporting. For this study, we created a data set on reporting guidelines in sociology with which we studied statistical reporting guidelines of 143 sociology journals. Also, we created three data sets to study statistical results reported in sociology articles. These data sets contained either results retrieved automatically by statcheck from the 2014-2016 volumes of *ASR*, *AJS*, *SQ*, *JMF*, and *CHQ*, or manually retrieved results related to explicitly stated hypotheses from the 2014-2016 volumes of *ASR*, *AJS*, and *SQ*. More specifically, we studied statistical reporting errors among 505 automatically retrieved APA-reported results and 404 manually retrieved results related to explicitly stated hypotheses. Furthermore, we studied publication bias, the 'bump' in p-values, and marginal significance among 2,960 automatically retrieved p-values and 4,929 manually retrieved p-values related to explicitly stated hypotheses.

One important result of our study is that for all our hypotheses, no differences between

Marcel

Mag ook bij benadering in 100 of 1000 tallen

Elise

Tried to improve!

results related to explicitly stated hypotheses and other results were found. Potentially, sociology authors are not persuaded by publication bias and publication pressure to ensure that especially results related to their articles' hypotheses are significant in ways that lead to less just nonsignificant results (and, thereby, a spike in just significant results), a bump in just significant results (created through QRPs), or more statistical reporting errors among results of hypotheses that make nonsignificant results significant. Furthermore, they seemingly do not see the need to 'boost' the evidential value of p-values related to hypotheses in the interval (.05-.10] in particular to create an (unwarranted) impression of a true effect. It should be noted that, in hindsight, we do not see why inconsistencies that do not change a reported result's statistical significance would occur more often among results related to hypotheses than among other results. Instead, a more plausible hypothesis would be that authors pay more attention to the accuracy of their most important results. Alternatively, authors could be more precise in writing down statistically significant results of hypotheses than they are in writing down non-significant results of hypotheses, since the latter might be less interesting to them due to publication bias/pressure.

Marcel: Dat mag niet, die moet je juist WEL bediscussieren! En misschien wel als eerste ….
Wat betekent het en waarom is dat zo denk je?

Elise: tried to improve

Marcel: Maak hier maar een alinea van die dit belangrijkste resultaat bediscussieerd.

Marcel: Wat betreft marginal significance moet je dan wel uitleggen waarom het verschil zo groot is tussen manueel en allp (die z=8.33)

Elise: This has been incorporated later on in the discussion

We found a lack of requested adherence to statistical reporting guidelines within sociology journals: only 9.1% required adherence to the APA statistical reporting guidelines.

Marcel: Hier mist nog een belangrijk statement dat bewijst dat de andere guidelines niet ok zijn, dus niet reproduceerbaar.?

Marcel
Idealiter past alles over dit onderwerp in EEN alinea. Ik heb het geprobeerd, jij mag het afmaken.

Elise
tried to improve

Elise: Tried to improve. Do I need to cite all other guidelines included in the study in here, as direct proof?

This is especially troublesome because none of the other guidelines authors were allowed or required to follow by sociology journals had statistical reporting guidelines that ensure reproducibility of results. For this reason, a lack of adherence to the APA statistical reporting guidelines is may hamper reproducibility of results, which in turn may negatively influence statistical reporting quality in sociology. Reproducible results can namely be easily assessed by third parties and may thereby motivate authors to thoroughly check if their results are reported correctly. Thereby, reproducibility of results could reduce the prevalence of statistical reporting errors. Concretely, for statistical reporting guidelines to ensure reproducible reporting, they should require that authors explicitly state for each analysis or result the test statistic, *df* (if applicable), *p*-value, and whether the test was one- or two-tailed (if applicable). All these requirements are covered by the APA statistical reporting guidelines (see American Psychological Association, 2022a).

We found that among automatically retrieved statistical APA-reported results from sociology articles, 13.7% were inconsistent and 1.6% grossly inconsistent. Previous research in psychology found a slightly lower but comparable prevalence of inconsistencies of 4.3%-12.8% (Bakker & Wicherts, 2011; Krawczyk, 2015; Nuijten et al., 2016; Veldkamp et al., 2014; Vermeulen et al., 2015; Wicherts et al., 2011). The prevalence of gross inconsistencies was comparable to the 0.8%-2.5% previously found in psychology (Bakker & Wicherts, 2011; Nuijten et al., 2016; Veldkamp et al., 2014; Vermeulen et al., 2015). Also, all gross inconsistencies from both 'APA' and 'Hyp' had a significant reported *p*-value and a nonsignificant recalculated *p*-value. The phenomenon of there being less significant recalculated *p*-values than reported ones is in line with research by Vermeulen et al. (2015) and Nuijten et al. (2016) in psychology. Although these results suggest that the prevalence and direction of statistical reporting errors is similar in both fields, we did not examine the prevalence of errors in the 11 other APA-journals of sociology. In any case, both fields would profit from

Elise: Added this back in because this is also a result we can compare to psychology

using statcheck in the review or editorial phase to prevent statistical reporting errors in the published literature (see Nuijten et al., 2016, 2017). Finally, and interestingly, we found lower prevalences of inconsistencies among automatically retrieved results for non-APA journals *ASR* and *AJS*, suggesting that the absence of APA statistical reporting guidelines does not seem to negatively impact the prevalence of statistical reporting errors in sociology. However, we should be careful interpreting this finding, since these two non-APA-journals are not representative but are considered top journals in the field .

Overall, we found no evidence of publication bias in sociology. This contrasts with Gerber & Malhotra (2008), who did find evidence of publication bias among results related to explicitly stated hypotheses in *ASR*, *AJS*, and *SQ*. Of course, it could be that publication bias was simply not present in the articles we studied, but difference in methodology may also be at least partially responsible for the different findings. Gerber & Malhotra (2008) retrieved substantially more results because they used z-values and t-values converted to z-values, and they calculated z-values using regression coefficients and standard errors (thereby mixing reported and recalculated results). However, it is not clear why this different methodology would affect results of analyses on publication bias. Because mixing recalculated and reported p-values is not recommended in analyses of reporting practices (see Hartgerink et al., 2016), we decided to study only reported p-values, resulting in relatively little data and hence low power to detect publication bias (if present). As both direct and indirect evidence of publication bias in the related field of psychology is strong (see Lakens, 2015b reevaluation of Masicampo & Lalande, 2012; Kühberger et al., 2014), and Gerber and Malhotra (2008) also found publication bias in sociology, we recommend more research in sociology, both on the process from conception to submission and acceptance of articles - as has previously been done in psychology by, e.g., Coursol & Wagner (1986) and Franco et al. (2014) - and using more data on reported statistical results.

We also did not find evidence of a 'bump' in p-value distributions of statistical results in sociology. This is in line with our own expectations and with Hartgerink et al. (2016), who

Elise
Rewrote the last phrase a little, since ASR and AJS are not the only non-APA journals we studied (we studied SQ as well, but retrieved very little results from it).

Marcel
Probeer kort te zijn (1 alinea), geen cijfers, wel redenen. Ik heb wat geschoven en geschrapt.

Elise
Bedankt! I have added the necessary references.

did not find consistent indications of a 'bump' in several psychology journals, and in accordance with the reanalysis of Masicampo & Lalande (2012) by Lakens (2015b). Following Hartgerink et al. (2016), we conclude that the absence of a 'bump' does not say anything, whereas a 'bump' suggests the (extensive) use of p-hacking in a field, probably accompanied by zero or small true effect sizes.

Assignment of marginal significance to p-values in the interval (.05-.10] in articles was rather common. Among automatically retrieved results, its prevalence of 35% at the results level in sociology was somewhat lower than the almost 40% found by Ohlsson Collentine et al. (2019) in psychology. This suggests there is no large systemic difference in the reporting of marginal significance in both fields. Noteworthy is the much higher prevalence of marginal significance among manually retrieved results related to hypotheses (81.5%). We mainly attribute this difference to authors being aware that p-values in the interval (.05-.10] are of low evidential value, which prevents them from assigning marginal significance in the articles' text (Ohlsson Collentine et al., 2019), whereas assigning marginal significance in tables and figures may seem as harmful to them. This explanation is corroborated by Leahey (2005), who found a mere 10% of articles using '$p < .10$' in two sociology journals, whereas the 'three-star system' for assigning significance (i.e., $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$), which is often used in table captions in sociology.

Marcel
Ook hier: NIET herhalen resultaten, graag discussie

Elise
check whether this reference actually belongs here

Elise
Rephrasing necessary, see word

To conclude,

Marcel: Deze alinea HELEMAAL AAN HET EIND. En combineren met de alinea hierna

Elise: have combined these two

we advise the sociological field to adopt the APA statistical reporting guidelines or establish its own set of statistical reporting guidelines. Besides improving reproducibility and the information value of results, this would enhance standardization, and thus, comparability of statistical results in sociology, and allow for preventing reporting inconsistent statistical results after applying statistical checking programs such as statcheck. Enforcing statistical

Marcel
Dit stukje kan evt ook worden gebruikt (maar korter) in de eerste alinea van de discussie die je nog moet schrijven. Edit: Maar kan ook hier! Hier beter denk ik

reporting guidelines would also be helpful for meta-research on statistical reporting, publication bias, and research practices. However, for statistical reporting guidelines to be most effective, it would be very helpful if journals and the American Sociological Association would actively enforce them. Otherwise, as Krawczyk (2015) has shown for inexactly reported *p*-values in experimental psychology APA-journals, suboptimal research practices may still seep through.

Limitations & suggestions for future research

Our study has some limitations related to amounts of retrieved data. Firstly, we retrieved relatively little data on publication bias and the bump in *p*-values, rendering us unable to provide firm conclusions on the presence of these phenomena in sociology. This is likely due to most *p*-values in sociology being reported inexactly and relatively many results not being reported in-text (also compared to psychology), but in tables and figures instead. In hindsight, we should have included more articles to obtain enough data to properly study publication bias and the 'bump' in *p*-values. By doing this, we might also have been able to include potentially relevant control variables in our logistic regression analyses. Finally, for statistical reporting errors, we had data in which the occurrence of events was quite rare, which means there was a risk of underestimating the occurrence of statistical errors inin logistic regression analyses (see, e.g., King & Zeng, 2001). This issue could be solved by applying Firth's correction to rare events logistic regression analyses (Puhr, Heinze, Nold, Lusa, & Geroldinger, 2017; Šinkovec, Geroldinger, & Heinze, 2019) .

Marcel
from 'Finally' onwards, 1st paragraph: Evt weg, weet ik nog niet goed

Methodologically, there were disadvantages to automatically retrieving data in our study. As mentioned before, most *p*-values in sociology articles are reported in tables and can thus only be retrieved manually. Furthermore, many results are reported in text, but are not

APA-reported, rendering statcheck unable to retrieve them. This was especially the case for the non-APA journals in our study. Given the tendency to report p-values and results in tables, it is quite likely that the in-text reported p-values and APA-reported results retrieved by statcheck are a selective set of all reported p-values and APA-reported/reproducible results: they are likely the most theoretically relevant results reported in a study. This has consequences for comparisons of our findings to previous studies in psychology, where it seems to be more common to also report more mundane p-values and APA-reported/reproducible results in text: it is likely wise not to make one-to-one translations from the parts of our study using statcheck to studies in psychology that use statcheck.

Considering the above, it seems that at present, automatic retrieval from the text of articles is methodologically not the best way of collecting data on statistical reporting quality in sociology. Instead, future research on this topic would ideally focus on retrieving statistical data from text, tables, and figures. At present, this can only be done manually, which is a very tedious process. It is therefore vital for reproductive and meta-analytical research purposes that software to extract p-values and reproducible results from tables and figures will be developed and/or improved. Only the ability to automatically extract statistical data from figures and tables will enable high-quality non-tedious assessment of statistical reporting quality in sociology. Steps towards automatically extracting data from figures for meta-analytical purposes have been undertaken in recent years. Pick, Nakagawa, & Noble (2018), e.g., have developed a promising R package called metaDigitise, which extracts descriptive statistics such as standard deviations, means, correlations, and CIs from several often-used types of plots from articles without major problems. While metaDigitise does not retrieve all types of individual statistics and fully reported results from all types of figures, its applicability can probably be broadened. Automatic data extraction from tables, on the other hand, seems to have its fair share of problems. In a review article of proposals for data extraction from the increasingly popular HTML-tables, Roldán, Jiménez, & Corchuelo (2020) found that extraction from tables is inhibited by, e.g., the inability to identify multi-

part cells, context-data cells, and split headers, and the inability to analyze the structure of a cell's contents. This leads to data extraction only being possible for certain relatively simple types of tables, which is suboptimal in a meta-analytical context. Thus, substantial work will have to be done to enable automatic extraction of statistical data from tables.

Finally, it would certainly be useful for future research to track the progress made in sociology in implementing statistical reporting guidelines. If substantial progress is made, it would be interesting to study whether progress seems to have a positive impact on statistical reporting quality in sociology. Then, future research can provide recommendations on how the quality of statistical reporting within sociology could be improved even more.

References

American Psychological Association. (2022a). APA style numbers and statistics guide. https://apastyle.apa.org/instructional-aids/numbers-statistics-guide.pdf.

American Psychological Association. (2022b). Browse journals by title. https://www.apa.org/pubs/journals/browse?query=Title:*&type=journal.

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Psychological Methods*, *132*(2), 235–244.

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

Bakker, M., & Wicherts, J. M. (2014). Outlier removal and the relation with reporting errors and quality of psychological research. *Psychological Methods*, *19*(3), 409–427. https://doi.org/10.1037/met0000014

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., … Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6. https://doi.org/10.1038/s41562-017-0189-z

Clarivate Analytics' Web of Science. (2016). Journal citation reports: Sociology, 2016. com.proxy.library.uu.nl/JCRJournalHomeAction.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/14805-018

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates. *Professional Psychology: Research and Practice*, *17*(2), 136–137. https://doi.org/10.1037/0735-7028.17.2.136

De Winter, J. C. F., & Dodou, D. (2015). A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*, (3), 1–44.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *Jama*, *263*(10), 1385–1389. https://doi.org/10.1001/jama.1990.03440100097014

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867–872. https://doi.org/10.1016/0140-6736(91)90201-Y

Epskamp, S., & Nuijten, M. B. (2016). *Statcheck: Extract statistics from articles and recompute p values*. https://cran.r-project.org/src/contrib/Archive/statcheck/.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS One*, *5*(4), 1–10. https://doi.org/10.1371/journal.pone.0010068

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *6203*(345), 1502–1505. https://doi.org/10.1126/science.1255484

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*, *7*(1), 8–12. https://doi.org/10.1177/1948550615598377

Gerber, A. S., & Malhotra, N. (2006). Can political science literatures be believed? A study of publication bias in the APSR and the AJPS. Paper presented at the annual meeting of the Midwest Political Science Association.

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods and*

Research, *37*(1), 3–30. https://doi.org/10.1177/0049124108318973

Ginsel, B., Aggarwal, A., Xuan, W., & Harris, I. (2015). The distribution of probability values in medical abstracts: An observational study. *BMC Research Notes*, *8*(1), 1–5. https://doi.org/10.1186/s13104-015-1691-x

Hartgerink, C. H. J. (2017). Reanalyzing Head et al. (2015): Investigating the robustness of widespread p-hacking. *PeerJ*, (5), 1–10. https://doi.org/10.7717/peerj.3068

Hartgerink, C. H. J., Van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & Van Assen, M. A. L. M. (2016). Distributions of p-values smaller than. 05 in psychology: What is going on? *PeerJ*, (4), 1–28. https://doi.org/10.7717/peerj.1935

Head, M. L., Holman, L., Lanfear, R., Khan, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), 1–15. https://doi.org/10.1371/journal.pbio.1002106

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Knight, J. (2003). Negative results: Null and void. *Nature*, *422*(6932), 554–555. https://doi.org/10.1038/422554a

Krawczyk, M. (2015). The search for significance: A few peculiarities in the distribution of p values in experimental psychology literature. *PloS One*, *10*(6), 1–19. https://doi.org/10.1371/journal.pone.0127872

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PloS One*, *9*(9), 1–8.

https://doi.org/10.1371/journal.pone.0105825

Lakens, D. (2015a). On the challenges of drawing conclusions from p-values just below 0.05. *PeerJ*, (3), 1–14. https://doi.org/10.7717/peerj.1142

Lakens, D. (2015b). What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *The Quarterly Journal of Experimental Psychology*, *68*(4), 829–832. https://doi.org/10.1080/17470218.2014.982664

Lang, T. A., & Altman, D. G. (2013). Basic statistical reporting for articles published in clinical medical journals: The SAMPL Guidelines. In P. Smart, H. Maisonneuve, & A. Polderman (Eds.), *Science editors' handbook* (pp. 24–30). Split, Croatia: European Association of Science Editors.

Lawrence, P. A. (2003). The politics of publication - authors, reviewers and editors must act to protect the quality of research. *Nature*, *422*(6929), 259–261. https://doi.org/10.1038/422259a

Leahey, E. (2005). Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces*, *84*(1), 1–24. https://doi.org/10.1353/sof.2005.0108

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. R. (2013). The life of p: "Just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology*, *66*(12), 2303–2309. https://doi.org/10.1080/17470218.2013.863371

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below. 05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. https://doi.org/10.1080/17470218.2012.711335

Maxwell, C. (1981). Clinical trials, reviews, and the journal of negative results. *British Journal of Clinical Pharmacology*, *11*(1), 15–18.

Nuijten, M. B., Borghuis, J., Veldkamp, C. L. S., Dominguez Alvarez, L., Van Assen, M. A. L. M., & Wicherts, J. M. (2017). Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra: Psychology*, *3*(1), 1–22.

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.

Ohlsson Collentine, A., Van Assen, M. A. L. M., & Hartgerink, C. H. J. (2019). The prevalence of marginally significant results in psychology over time. *Psychological Science*, *30*(4), 576–586.

Pick, J. L., Nakagawa, S., & Noble, D. W. A. (2018). Reproducible, flexible and high throughput data extraction from primary literature: The metaDigitise package. *Methods in Ecology and Evolution*, *10*(3), 426–431.

Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, *27*(7), 1036–1042.

Puhr, R., Heinze, G., Nold, M., Lusa, L., & Geroldinger, A. (2017). Firth's logistic regression with rare events: Accurate effect estimates and predictions? *Statistics in Medicine*, *36*(14), 2302–2317.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; http://www.R-project.org/.

Roldán, J. C., Jiménez, P., & Corchuelo, R. (2020). On extracting data from tables that are encoded using html. *Knowledge-Based Systems*, *190*, 105157.

Sedlmeier, P., & Gigerenzer, G. (1989). A power primer. *Psychological Bulletin*, *105*(2), 309–316.

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F., & Altman, D. G. (2011). Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR Network. *BMC Medicine*, *22*(11), 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547.

Song, F., Parekh, S., Hooper, L., Ryder, Y. K., Sutton, A. J., Hing, C., … Harvey, I. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment*, *14*(8), 1–236.

Šinkovec, H., Geroldinger, A., & Heinze, G. (2019). Bring more data!—a good advice? Removing separation in logistic regression by increasing sample size. *International Journal of Environmental Research and Public Health*, *16*(23), 1–12.

Ulrich, R., & Miller, J. (2015). P-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General*, *144*(6), 1137–1145.

Veldkamp, C. L., 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), 1–19.

Vermeulen, I., Beukeboom, C. J., Batenburg, A., Avramiea, A., Stoyanov, D., Van de Velde, B., & Oegema, D. (2015). Blinded by the light: How a focus on statistical "significance" may cause p-value misreporting and an excess of p-values just below .05 in communication science. *Communication Methods and Measures*, *9*(4), 253–279.

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), 1–7.

.

Supplement

Table S1: Logistic regressions for H1 and H2, which concern the difference in prevalence of (gross) inconsistencies in results related to hypotheses versus results not related to hypotheses.

	b	SE	p	OR [95% CI]
Inconsistencies (H1)				
Intercept	-1.820	.157	$< .001$	.162 [.118, .218]
Result hypothesis	-.073	.278	.793	.930 [.531, 1.585]
Gross inconsistencies (H2)				
Intercept	-4.422	.503	$< .001$	.012 [.004, .028]
Result hypothesis	.708	.714	.321	2.030 [.475, 8.685]

Table S2: Logistic regressions for H3a and H3b, which concerns the difference in prevalence of publication bias in results related to hypotheses versus results not related to hypotheses.

	b	SE	p	OR [95% CI]
Binwidth .01 (H3a)				
Intercept	-.205	.287	.476	.815 [.456, 1.428]
Result hypothesis	-.132	.504	.794	.877 [.321, 2.345]
Binwidth .02 (H3b)				
Intercept	.089	.211	.673	1.093 [.723, 1.658]
Result hypothesis	-.251	.392	.521	.778 [.358, 1.674]

Table S3: Logistic regression for H4, which concerns the difference in prevalence of marginal significance in results related to hypotheses versus results not related to hypotheses.

	b	SE	p	OR [95% CI]
Intercept	-.480	.177	.007	.619 [.436, .871]
Result hypothesis	-.437	.318	.170	.646 [.342, 1.194]

Table S4: Descriptive statistics on (gross) inconsistencies for 'APA' and 'Hyp'.

	Article level			Results level		
	Total	With inconsistencies	With gross inconsistencies	Total	Inconsistencies	Gross inconsistencies
'APA'						
Relation to hypothesis						
Yes	29			168	22 (13.1%)	4 (2.4%)
No	68			337	47 (13.9%)	4 (1.2%)
Journal						
ASR	7	1 (14.3%)	1 (14.3%)	43	1 (2.3%)	1 (2.3%)
AJS	3	1 (33.3%)	0 (0%)	41	2 (4.9%)	0 (0%)
SQ	2	2 (100%)	0 (0%)	5	5 (100%)	0 (0%)
JMF	36	12 (33.3%)	2 (5.6%)	185	27 (14.6%)	3 (1.6%)
CHQ	28	13 (46.4%)	2 (7.1%)	231	34 (14.7%)	4 (1.7%)
Year						
2014	20	7 (35%)	1 (5.0%)	172	22 (12.8%)	1 (0.6%)
2015	21	7 (33.3%)	2 (9.5%)	136	15 (11.0%)	3 (2.2%)
2016	35	15 (42.9%)	2 (5.7%)	197	32 (16.2%)	4 (2.0%)
Total	76	29 (38.2%)	5 (6.6%)	505	69 (13.7%)	8 (1.6%)
'Hyp'						
Journal						
ASR	10	4 (40.0%)	1 (10.0%)	331	11 (3.3%)	1 (0.3%)
AJS	7	1 (14.3%)	1 (14.3%)	68	1 (1.5%)	1 (1.5%)
SQ	2	1 (50.0%)	0 (0%)	5	2 (40.0%)	0 (0%)
Year						
2014	11	4 (36.4%)	1 (9.1%)	312	11 (3.5%)	1 (0.3%)
2015	3	0 (0%)	0 (0%)	33	0 (0%)	0 (0%)
2016	5	2 (40.0%)	1 (20.0%)	59	3 (5.1%)	1 (1.7%)
Total	19	6 (31.6%)	2 (10.5%)	404	14 (3.5%)	2 (0.5%)

Note. The numbers of articles for the results (not) related to explicitly stated hypotheses reflect the numbers of articles that contain at least one result that is (not) related to an explicitly stated hypothesis.