**The reporting of statistical results in sociology:**

**a systematic review**

**Research master thesis**

**Elise Schramkowski - 4067878**

**Utrecht University**

**Supervisor: Marcel van Assen**

**Research master programme:**

**Methodology and Statistics for the Behavioural, Biomedical and Social Sciences**

**Tab**

**11 August, 2021**

**Word count: approximately 5200 (excluding figures, tables and references)**

**Abstract**

High quality of reported results in scientific articles is vital for providing scientific and nonscientific communities with correct information on studied phenomena. In this article, the quality of different aspects of reported results in sociology is studied. Firstly, the adherence to statistical reporting guidelines by 143 sociology journals was studied. Furthermore, the presence of statistical reporting errors, publication bias, a ‘bump’ in just significant *p*-values and marginal significance among results of papers were studied. For this purpose, data were collected from the 2014-2016 volumes of five sociology journals, from which information was retrieved automatically using the package statcheck (Epskamp and Nuijten 2015). Furthermore, information on these topics was retrieved manually for the 2014-2016 volumes of three of these journals, which were previously used in a study by Gerber & Malhotra (2008) on publication bias in sociology. It was found that only 13 of these journals (9.1%) adhered to statistical reporting guidelines. No convincing evidence of the presence of publication bias and a ‘bump’ in *p*-values was found. Marginal significance was rather prevalent.

*More information will be added later*

**Keywords**

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

Statistical results in social science papers provide scientific and nonscientific communities with essential information about the social world. Therefore, examining statistical reporting standards and the quality of reported statistical results is highly relevant. Statistical results should comply with the following standards. Firstly, they should provide sufficient information for reproduction; this will make it easier for readers to assess the reported results of a study critically (Simera et al. 2010). Secondly, statistical results should not contain errors, because errors reflect inaccuracies in research and can lead to incorrect statistical conclusions, placing readers at risk of being misinformed about the nature of social phenomena. Finally, the reporting of statistical results in papers should be standardized at least within disciplines to enable authors to communicate them clearly and to enable critical evaluation of the quality of reported results. In this systematic review, we examined several aspects of reported results quality in sociology, namelythe adherence to statistical reporting guidelines, the prevalence of statistical reporting errors, evidence of publication bias and *p*-hacking, a ‘bump’ in just significant *p*-values, and *p*-values reported as marginally significant.

It has been suggested that the presence of statistical reporting guidelines in a discipline might lead to less reporting errors (Lang and Altman 2013). Furthermore, absence of clear statistical reporting guidelines leads to authors providing insufficient information when reporting statistics, which makes critical assessment of their results difficult(Simera et al. 2010). This suggests authors use statistical reporting guidelines when this is required by the journals they submit their papers to. In contrast with other disciplines (e.g., psychology), no general statistical reporting guidelines have been developed within sociology. Different sociology journals require authors to adhere to different style guidelines, such as the American Psychological Association (APA), American Sociological Association (ASA), Chicago, Harvard, and Oxford style guidelines. Of these style guidelines, only the APA guidelines contain statistical reporting guidelines. Thus, we examined which journals request authors to adopt these guidelines to evaluate if insufficient and incorrect reporting of results in sociology could be explained by sociology journals not requesting the use of clear statistical reporting guidelines.

Statistical reporting errors, also called *inconsistencies*, occur when there is an inconsistency between the following parameters belonging to a reported result: the test statistic, (if used) the degrees of freedom (*df*), and the *p*-value. Inconsistencies are undesirable for two reasons. Firstly, they reflect inaccuracies in reported results. Secondly, they can lead to changes in 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. If an inconsistency leads to changes in statistical conclusions, it is called a *gross inconsistency*. (implying not rejecting H­0), as *t*(50) = 2.05 implies *p* = .046 (actually rejecting H0). An example of an inconsistent APA-reported result 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’, suggesting a statistically significant result, but *t*(50) = 1.99 implies *p* = .052, which implies that the null-hypothesis should not be rejected. Another example of a gross inconsistency is ‘*t*(50) = 2.05, *p* = .055’. To our knowledge, no research on the prevalence of statistical reporting errors in sociology has been conducted at present. However, research in psychology suggests that 4% to 10% of reported results are inconsistently reported (Wicherts, Bakker, and Molenaar 2011; Nuijten et al. 2016). Gross inconsistencies have been found in about 0.8%-2.5% of reported results (e.g., Veldkamp et al. 2014; Nuijten et al. 2016; Hartgerink et al. 2016). Nuijten et al. (2016) found that gross inconsistencies corresponding to reported statistically significant results occur relatively often; due to gross inconsistencies, the percentage of significant *p*-valuesthey found among recalculated *p*-valueswas 2.2 percentage points lower than that found among reported *p*-values (it went from 76.6% to 74.4%). Similarly, Hartgerink et al. (2016) found that of all *p*-values reported as *p* = .05, 67.45% was actually larger than .05.This could point to authors using *p*-hacking and incorrect *p*-value rounding to obtain (false) significance (John et al. 2012; Hartgerink et al. 2016; Nuijten et al. 2016). Therefore, we studied the prevalence of statistical reporting errors manually as well as using the R package statcheck (Epskamp and Nuijten 2015), which checks the consistency of fully APA-reported results. The APA guidelines on reporting statistical results should be used by at least those sociology journals that follow the general APA guidelines.

Publication bias is the phenomenon that statistically significant results are published relatively more often than non-significant ones. It is one of the suboptimal research or publishing practices that can lead to a relatively high prevalence of just significant *p*-values. In various fields of research of science, and especially in the social sciences and the biomedical sciences, publication bias has been found to some extent (e.g., Dickersin, 1990; Easterbrook et al.1991; Fanelli 2011; Masicampo and Lalande 2012; Franco, Malhotra, and Simonovits 2014; De Winter and Dodou 2015; Franco, Malhotra, and Simonovits 2016). Potential causes of a high prevalence of just significant *p*-values are questionable research practices (QRPs) known as *p*-hacking (Hartgerink et al. 2016; John, Loewenstein, and Prelec 2012; Lakens 2015a; Masicampo and Lalande 2012) and the use of researcher degrees of freedom (Simmons et al. 2011), in which one collects or selects data and/or analyzes results until statistical significance is obtained. Just significant *p*-values can be defined as *p*-values in the range just below the most frequently used threshold for determining significance, *p* = .05. The presence of publication bias and *p*-hacking in sociology was studied by Gerber and Malhotra (2008), who looked at statistical results corresponding to hypotheses in all three volumes of the sociology journals *American Sociological Review* (*ASR*), *American Journal of Sociology* (*AJS*) and *Sociological Quarterly* (*SQ*) from 2003-2006. They compared, among others, the difference in the number of *z*-values in an interval that closely approximates the *p*-value interval (.04-.05], and an interval that closely approximates the *p*-value interval (.05-.06]. They found that the frequency of results corresponding to the *p*-value interval (.04-.05] was 3.25 to 4 times higher than that corresponding to the *p*-value interval (.05-.06], presenting strong evidence of publication bias. A similar method was used by Masicampo and Lalande (2012) in psychology and De Winter and Dodou (2015) across disciplines:, who studied just significant *p*-values in the range (.04-.05] and just non-significant *p-*values in the range (.05 - .06]. We studied these *p*-valueranges, and, like Gerber & Malhotra (2008), the *p*-value ranges (.03 - .05] and (.05 - .07], since larger intervals provide higher power for tests on publication bias.

A non-monotonic increase or a ‘bump’ in *p*-values, which can be said to occur when *p*-values in the interval (.04-.05] occur more often than *p*-values in the interval (.03-.04], is also characterised by a predominance of just significant *p*-values. It is evidence of *p*-hacking, as publication bias cannot result in a ‘bump’ in *p*-values. Most discipline-specific research on a ‘bump’ in *p*-values has been conducted in psychology, where some studies focusing on *p*-values in the interval (.04-.05] claimed to have found evidence of a ‘bump’ (Masicampo and Lalande 2012; Leggett et al. 2013). However, according to Lakens (2015b), these studies had not modeled their *p*-value distributions correctly, as they did not take possible publication bias into account. Relatedly, Hartgerink et al. (2016) showed that *p*-hacking does not result in a ‘bump’ if true effect sizes are medium (Cohen’s *d* = 0.5) or larger. Although this shows that the absence of a ‘bump’ is no evidence of absence of *p*-hacking, the presence of a ‘bump’ can only be explained by *p*-hacking. Following Hartgerink et al. (2016), we studied the presence of a ‘bump’ using the intervals (.04 - .05] versus (.03 - .04] and (.03 - .05] versus (.01 - .03]. Larger intervals were again used because they may provide higher power of tests on the ‘bump’, although, power may also decrease because *p*-values of .01 are much more likely than *p*-values of .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., 2018; Ohlsson Collentine et al. 2019; Pritschet et al., 2016). Thus, arguing non-significant results represent true effects may result in (unwarranted) false positives. Since this can lead to audiences deciding 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 among *p*-values in the range (.05-.10]; Pritschet et al. found that of *p*-values that are reported as marginally significant in psychology, 92.6% were in interval (.05-.10]. Ohlsson Collentine et al. (2019) found that almost 40% of all *p*-values in the interval (.05-.10] 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 reporting *p*-values contained at least one *p*-value in interval (.05-.10] that was reported as marginally significant. Leahey (2005) found that in 10% of sociology articles from two unnamed top sociology journals from 1995-2000, a significance level of low evidential value of *p* < .10 was used.

Finally**,** we examinedthe prevalence of statistical reporting errors, publication bias and *p*-hacking, a ‘bump’ in *p*-values,and *p*-values reported as marginally significant among results of explicitly stated hypotheses (i.e., referred to in the paper’s text as hypotheses to be tested), and compared it to the prevalence of these phenomena among statistical results not related to explicitly stated hypotheses. One would hope that at least reported results of explicitly stated hypotheses would be without inaccuracies through careful checking by authors before submission and by reviewers and editors before accepting a paper. On the other hand, as publication bias/*p*-hacking is assumed to primarily operate on results corresponding to main hypotheses in papers (Gerber and Malhotra 2006), one would expect the prevalence of (gross) inconsistencies, a ‘bump’ in *p*-values, and marginal significance to be stronger among results corresponding to explicitly stated hypotheses. More specifically, we hypothesized the following:

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

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

*H3: The discrepancy between the amount of p-values in the interval (.04-.05] and the amount of p-values in the interval (.05-.06] among results of explicitly stated hypotheses is larger than among results not related to explicitly stated hypotheses.*

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

We did not construct a hypothesis regarding 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 relatively higher in sociology (Sedlmeier and Gigerenzer 1989; Cohen 1992). Assuming the same distribution of true effects examined, higher statistical power implies on average lower *p*-values in sociology than in psychology. Therefore, and because evidence of a ‘bump’ in psychology is weak at best, we expected neither a ‘bump’ in *p*-values in sociology in general, nor a difference in the presence or size of such a ‘bump’ between results related to hypotheses and results not related to hypotheses.

**METHOD**

*Data sources*

For our research on statistical reporting guidelines, we used the Web of Science’s Journal Citation Reports (2016) to create a list of sociology journals. For each journal on this list, we verified whether it adhered to the APA statistical reporting guidelines or not.

For our research on statistical reporting errors, publication bias/*p*-hacking, the ‘bump’ in *p*-values, and marginal significance, we collected data from several journals. Since statcheck only retrieves and recalculates fully APA-reported results, we collected articles from two journals from Web of Science’s Journal Citation Reports (2016) that require APA statistical reporting, namely *Cornell Hospitality Quarterly* (*CHQ*) and *Journal of Marriage and Family* (*JMF*). We selected *CHQ* and JMF because they had the highest impact factors (1st rank for *CHQ* with 2.657, and 3rd rank for JMF with 2.238) among sociology journals requiring APA statistical reporting1. We used all 310 articles from the 2014-2016 volumes of *CHQ* and *JMF* (100 and 210 articles, respectively). To compare differences in statistical reporting errors between APA journals and non-APA journals, we also retrieved results of articles (325 in total) from the 2014-2016[[1]](#footnote-1) volumes of three non­-APA journals from Web of Science’s Journal Citation Reports (2016): *ASR*, *AJS*, and *SQ*. These were used in Gerber & Malhotra’s (2008) study on publication bias, which we wanted to replicate. Fully APA-reported results retrieved from all five journals using statcheck were put into a data set called ‘*APA*’, and *p*-values retrieved from all five journals by statcheck were put into a data set called ‘*AllP*’. Finally, we created a data set called ‘*Hyp*’, which contained reported *p*-values and statistical results related to explicitly stated hypotheses which were manually retrieved from *ASR*, *AJS*, and *SQ*. This means results with *p*-values reported according to the APA guidelines are also included in ‘*Hyp*’, as this data set contains *all* reproducible results related to explicitly stated hypotheses.

*Data collection*

For each sociology journal in Web of Science’s Journal Citation Reports (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. 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 information using statcheck, we converted all relevant articles to HTML format, because statcheck 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 of articles (Epskamp and Nuijten 2015) to automatically retrieve APA-reported results, reported *p*-values, and recalculated *p*-values. The first data set, ‘*APA*’, contains information retrieved by statcheck on all aspects of fully APA-reported results of all five journals - test statistics (*t, z, F, χ*2, or *r*), *df*,andreported *p*-values - and recalculated *p*-values. Results with exactly reported *p*-values and results with *p*-values reported as ‘smaller than’, ‘greater than’, or ‘non-significant’ can be retrieved by statcheck. If *p*-values are reported as non-significant, statcheck assigns them the label ‘NA’. ‘*APA*’ also contains information retrieved by statcheck on whether reported *p*-values are (grossly) inconsistent with their recalculated counterparts. If a reported result seems inconsistent (and this cannot be due to correct rounding), statcheck applies a one-sided test to the reported result. If this leads to a consistent reported result, it keeps the one-sided test. Otherwise, it keeps 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, ‘H1, or ‘the first hypothesis’. This information was used to test our hypotheses on statistical reporting errors (H1 and H2). Since statcheck also retrieved other (incomprehensible) information besides fully APA-reported results, some rows of ‘*APA*’ were excluded. In total, 505 results from 78 articles could be used for descriptive purposes and hypothesis testing (see Table 1 and Table 2).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Table 1*. Overview of information provided by each separate data set. | | | | |
|  | | DATA SET | | |
|  | | AllP | APA | Hyp |
| Information source | |  |  |  |
|  | Journals | All | All | ASR/AJS/SQ |
|  | Part of article | Text | Text | Text/table/figure |
|  | Results related to explicitly stated hypotheses? | Partly | Partly | Yes |
|  | |  |  |  |
|  | # of articles (total/with information)1 | 471/314 | 80/78 | 91/91 |
|  | # of results (total/with information)1 | 7,280/2,959 | 524/505 | 4,849/4,849 |
| 1 ‘Total information’ refers to the numbers of articles from which results were retrieved and the total numbers of results retrieved. ‘With information’ refers to how many of these articles/results contained information relevant to a specific dataset. | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Table 2*. Overview of the numbers of results retrieved for all relevant parts of our study on sociology articles, for each data set separately. | | | | | |
|  | | | DATA SET | | |
| PART OF STUDY | | | AllP | APA | Hyp |
| Statistical reporting errors | | |  |  |  |
|  | Descriptive information | | - | 505 (76) | 353 (20) |
|  | Testing hypotheses (gross) inconsistencies (H1 & H2) | | - | 505 (76) | - |
| Bump in *p*-values | | |  |  |  |
|  | Descriptive information | |  |  |  |
|  |  | (.03 - .04] - (.04 - .05] | 64 | 38 | 14 |
|  |  | (.01 - .03] - (.03 - .05] | 184 | 88 | 37 |
| Publication bias | | |  |  |  |
|  | Descriptive information | |  |  |  |
|  |  | (.04 - .05] - (.05 - .06] | 73 | 28 | 14 |
|  |  | (.03 - .05] - (.05 - .07] | 127 | 56 | 26 |
|  | Testing hypothesis publication bias (H3) | |  |  |  |
|  |  | (.04 - .05] - (.05 - .06] | 73 | - | - |
|  |  | (.03 - .05] - (.05 - .07] | 127 | - | - |
| Marginal significance | | |  |  |  |
|  | Descriptive information | | 199 (107) | - | 130 (30) |
|  | Testing hypothesis marginal significance (H4) | | 199 (107) | - | - |
| *Note*. The numbers of articles from which results were used for different parts of the study are shown between parentheses. | | | | | |

The second dataset, ‘*AllP*’, consists of all *p*-values reported in all five journals retrieved by statcheck. We manually added information on whether results were related to an explicitly stated hypothesis or not in the same way as for ‘*APA*’. Of 7,280 results retrieved by statcheck, we removed 4,354 (59.8%) because they did not refer to reported *p*-values. Ultimately, 2,926 results from 308 articles were reported *p*-values. From these reported *p*-values, descriptive information on the ‘bump’ in *p*-values, publication bias, and marginal significance as assigned by authors to *p*-values in the range (.05-10] was obtained. Furthermore, these reported *p*-values were used to test H3 and H4 (see Table 1 and Table 2 for an overview). To determine if marginal significance was assigned to a reported *p*-value, we looked up *p*-values in the (.05-10] range in the text of articles[[2]](#footnote-2). Then, following Ohlsson Collentine et al. (2019), we decided that a *p*-value was assigned marginal significance by the authors if at least the expressions ‘margin\*’ or ‘approach\*’ were mentioned in relation to its significance. The text used to conclude that a *p*-value was (not) considered marginally significant was stored manually in a separate column of ‘*AllP*’. Finally, we examined the percentage of articles with at least one *p*-value in the range (.05 - .10] to which marginally significance was assigned.

A third data in set, *‘Hyp*’*,* was created to replicate Gerber and Malhotra’s (2008) research on publication bias by manually retrieving results from articles. 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 certain inclusion criteria. Firstly, like Gerber and Malhotra (2008), we only studied articles that explicitly stated one or more hypotheses before its results were presented. Of the 322 articles in *ASR*, *AJS* and *SQ*, 100 (31.1%) contained explicitly stated hypotheses. Secondly, articles had to contain at least one result corresponding to one or more explicitly stated hypotheses. Based on this additional criterion, 91 articles (91.0%) were included (see Figure 1 for an overview of the selection process). Following Gerber & Malhotra (2008), ‘*Hyp*’contains all relevant results from all models used to test explicitly stated hypotheses. Information from appendices was also included, but information from supplements was not, since only appendices are part of articles as published. There are also some differences between our study and that of Gerber & Malhotra (2008). Gerber & Malhotra (2008) used caliper tests for *z*-distributions consisting of *z*-values and *t*-values (converted to *z*-values)within 5%, 10%, 15% or 20% of *z* = 1.64 (one-sided testing) or *z* = 1.96 (two-sided testing). If *z*-values or *t*-values were unavailable, regression coefficients and standard errors were used to calculate accompanying *z*-values. We used exactly reported *p*-values in the ranges (.04 - .06] and (.03 - .07] instead, since it was often unknown what kind of distribution an analysis was based on, and this allowed us to include *p*-values based on *F*-values and χ2-svalues. Finally, Gerber & Malhotra (2008) excluded articles with more than 38 relevant coefficients because they could lead to certain articles having a disproportionate effect on analyses. We did not do so since we wanted to include *all p*-valuesrelevant for studying publication bias. If one or more articles would influence the results disproportionately, we would do extra analyses without these articles. We organized all aspects of a result of an explicitly stated hypothesis (*p*-values, regression coefficients, odds ratios, *z*-values, *t*-values, *F*-values,  *χ2*-values,standard errors, the phrasing of the hypothesis results belonged to as retrieved from the article, and, if applicable, text from the article in which particular results are mentioned) in the same way as in ‘*APA*’. In total, 4,849 results were included in ‘*Hyp*’. Where possible, we checked whether statistical results were (grossly) inconsistent by recalculating their *p*-values. For information on how this was done, see Table 3 and Table 4. We also manually added information on marginal significance of *p*-values in the range (.05 - .1] mentioned in text to ‘*Hyp*’as we did for ‘*AllP*’. For *p*-values in tables, we considered significance levels of *p* < .10 in captions of tables (indicated by, e.g., by an asterisk) to be assignment of marginal significance. Finally, we studied the percentage of articles containing marginally significant results in the range (.05 - .10] in ‘*Hyp*’. For an overview of information in ‘*Hyp*’, see Table 2. 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 datasets.

Text

Description automatically generated

*Figure 1*. Flowchart describing the process of selecting articles from which results were retrieved manually for ‘*Hyp*’. Percentages are conditional on the previous step.

|  |  |  |
| --- | --- | --- |
| *Table 3*. Conditions under which different types of reported *p*-values from ‘*Hyp*’ are considered (grossly) inconsistent. Results which are gross inconsistent are also automatically an inconsistency. | | |
| Type of reported *p*-value | Inconsistent if… | Grossly inconsistent if… |
| Insignificant (*ns*) | Cal*P* ≤ .05 | Cal*P* ≤ .05 |
| Smaller than a certain *p*-value (e.g., *p* < .03) | Cal*P* ≥ Rep*P* | Cal*P* is *p* > .05and Rep*P* is *p* ≤ .05 |
| Equal to or larger than a certain *p*-value (e.g., *p* ≥ .03) | Cal*P* < Rep*P* | Cal*P* is *p* < .05and Rep*P* is *p* ≥ .05 |
| Exactly reported (e.g., *p* = .03) | Cal*P* differs from reported *p*-value in ways that cannot be due to rounding.1 | Cal*P* differs from Rep*P* in ways that cannot be due to rounding, and CalP is significant while the Rep*P* is insignificant or vice versa.1 |
| *Note*. Cal*P* = recalculated *p*-value, Rep*P* = reported *p*-value. 1See 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 reported and recalculated *p*-values could be due to correct rounding or are indicative of an inconsistency among exactly reported results from ‘*Hyp*’. | |
| *b* & *SE* | Only used for recalculation if it was clear that a result was based on the *z*-distribution or the *t-*distribution.  *Example*   * a *b*-value and *SE* of a result are reported as *b* = 3.11, *SE* = 2.11. * the interval of correct rounding 🡪 statistics *b* = 3.11 and *SE* = 2.11 imply 3.105 ≤ *b* < 3.115 and 2.105 ≤ *SE* < 2.115. * Dividing the upper bound of *b* by the lower bound of *SE* and dividing the lower bound of *b* by the upper bound of *SE* 🡪 one finds the largest and smallest *t/z*-values, respectively, which are consistent with the reported *b* and *SE*. * Using lower and upper bound *t/z*-values and reported *df* to see if reported *p*-value is consistent . * Rounding recalculated *p*-value to the same number of decimals as its reported counterpart using the round() function of the R base package. * If a reported *p*-value is equal to the lowest or highest *p*-value that is possible under correct rounding, or is in between these two *p*-values, it is considered correct. Otherwise, it is considered incorrect. |
|  |  |
| Test statistics (*t, z, F, χ*2, *r*) | Functions used for the recalculation of *p*-values, the following functions of the ‘stats’ package in R were used: pt() for *t* and *r*, pnorm() for *z*, pf() for *F*, and chisq() for *χ*2. Note that all functions, except for pnorm(), require information on *df*.  *Example*   * a *t*-value reported as *t* = 3.11 is consistent with 3.105 ≤ *t* < 3.115. * Calculating the lowest and highest *p*-values that are consistent with this reported *t* using the pt() function. * Rounding the *p*-values to the same number of decimals as the p-value reported in the article using the round() function of the ‘base’ package in R. * If the reported *p-*value corresponding to the *t*-value is equal to the lowest or highest *p*-values possible under correct rounding, or is in between these *p*-values, it is considered correct. Otherwise, it is considered incorrect. |

Statistical Analyses

In our descriptive analyses (which consist of frequencies and percentages), we reported how many journals require authors to adhere to the APA statistical reporting guidelines. For (gross) inconsistencies, descriptive results were based on ‘*APA*’ and ‘*Hyp*’. As for gross inconsistencies, we followed Nuijten et al. (2016) by studying the direction of gross inconsistencies: do errors make non-significant results significant, or vice versa? For publication bias, the ‘bump’ in *p*-values, and marginal significance, descriptive results were based on ‘*AllP*’ and ‘*Hyp*’. For marginal significance, these data sets also provided descriptive statistics at the article level. For all research topics but statistical reporting guidelines, descriptive results were split by explicitly stated hypothesis (yes/no), journal (*ASR*, *AJS*, *SQ*, and, for parts of the study using statcheck, *CHQ*, and *JMF* ), and year (2014, 2015, 2016). See Table 2 for an overview of which data sets were used for which parts of the research.

Nuijten et al. (forthcoming) 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. Thirdly, 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 in theory, the third method is most appropriate, 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. forthcoming). Therefore, following Wicherts et al. (2011) and Nuijten et al. (2016), we tested H1 and H2 using the second method, that is, using logistic regressions on individual *p*-values.

We conducted logistic regressions to test our hypothesis on publication bias (H3) with exactly reported *p*-values from ‘*AllP*’. Since statcheck interprets results with *p* = .05 as being statistically significant (Epskamp and Nuijten 2015), *p* = .05 was included in the interval of just significant *p*-values for the logistic regressions. H3 was tested using *p*-values from smaller intervals - (.04 - .05] versus (.05 - .06] – to obtain more precise results, and larger intervals - (.03 - .05] versus (.05 - .07] – for a potentially more powerful test.

To test our hypothesis on *p*-values reported asmarginally significant (H4), we used exactly reported *p*-values in the range (.05 - .10] from ‘*AllP*’. Logistic regressions were conducted to test this hypothesis.

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 we based our analyses on too few data for including multiple predictors.

**RESULTS**

In this section, we start by presenting our results regarding statistical reporting guidelines. Next, results on statistical reporting errors, publication bias/*p*-hacking, the ‘bump’ in *p*-values, and marginal significance are discussed. For each results-level topic, we first present the relevant descriptive results and (if applicable) results of hypothesis testing as obtained using automatic retrieval. Then, we discuss descriptive statistics of results related to explicitly stated hypotheses as obtained using manual retrieval. Results on specific years and journals that were of little theoretical interest or were based on too little data were not discussed in text but can be found in the corresponding tables. Full tables of the results of logistic regressions can be found in the supplement.

*Statistical Reporting Guidelines*

Of the 143 sociology journals mentioned by Web of Science in 2016, one journal (Society) did not seem to have any explicit guidelines authors are required or allowed to follow when preparing their manuscripts. Four journals (approximately 2.8%) explicitly required authors to follow guidelines established by the journal itself, and 102 (approximately 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. It is important to note that the APA guidelines are the only guidelines concerning statistical reporting used among the journals we studied. See Table 5 for an overview of the numbers of sociology journals requesting adherence to different categories of guidelines.

|  |  |  |  |
| --- | --- | --- | --- |
| *Table 5*. Numbers and percentages of journals requesting and allowing adherence to certain types of statistical reporting guidelines. | | | |
| Types of guidelinesauthors are requested/allowed to use by journals | | | #of journals (% of total) |
| Required | | |  |
|  | APA | Whole manual | 10 (7.0%) |
|  | Only references | 10 (7.0%) |
|  | ASA | Whole manual | 12 (8.4%) |
|  | Only references | 3 (2.1%) |
|  | Chicago | Whole manual | 7 (4.9%) |
|  | Only references | 6 (4.2%) |
|  | Harvard | Whole 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%) |
|  | Own | | 4 (2.8%) |
|  | Multiple options (of which one must be chosen) | | 1 (0.7%) |
|  | Multiple required | | 37 (25.9%) |
|  | Multiple required (one of which is full APA manual) | | 3 (2.1%) |
| Multiple allowed | | | 1 (0.7%) |
| None mentioned | | | 1 (0.7%) |
| Unknown1 | | | 1 (0.7%) |
| *Note*. Number of sociology journals from the 2016 Web of Science list is *N* = 143. 1We were unable to find out which guidelines authors publishing in the journal Society are required or allowed to use because the link provided on the website that should give access to this information led us to a ‘page not found’ error. | | | |

*Statistical reporting errors*

Of the 505 results in ‘*APA*’, 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 an insignificant recalculated *p*-value, making the percentage of significant *p*-values found among recalculated *p*-values 1.6 percentage points lower (30.3% rather than 31.9%) than that found among their reported counterparts. Out of 168 results corresponding to explicitly stated hypotheses, 22 (13.1%) were inconsistent and 4 (2.4%) grossly inconsistent. Out of 337 results corresponding to explicitly stated hypotheses, 47 (13.9%) were inconsistent and 4 (1.2%) grossly inconsistent. Our hypotheses that less (gross) inconsistencies will be observed for results on explicitly stated hypotheses are 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 significant, *b* = .708, *p* = .321, OR = 2.030, 95% CI [.475, 8.685]. Most recalculated *p*-values were retrieved from the two APA journals. These journals, *JMF* and *CHQ*, had very similar percentages of inconsistencies (14.6% and 14.7%, respectively) and gross inconsistencies (1.6% and 1.7%, respectively). As for ‘*Hyp*’, 15 out of 351 recalculated *p*-values were inconsistent (4.6%), and three (0.9%) were grossly inconsistent. Most recalculated *p*-values from ‘*Hyp*’ belonged to *ASR* articles (278 from 10 articles), and were published in 2014 (307 from 12 articles). For a more detailed overview of results, see Table 6.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Table 6*. Descriptive statistics on the number of articles, number of results and (gross) inconsistencies for the ‘*APA*’ and ‘*Hyp*’ data sets. | | | | | | |
|  | | | **‘APA’** | | | |
|  | | | #Articles | #Results | Inconsistencies | Gross inconsistencies |
| Relation to explicitly stated hypothesis1 | | |  |  |  |  |
|  | 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’** | | | |
|  | | | #Articles | #Results | Inconsistencies | Gross inconsistencies |
| Journal | | |  |  |  |  |
|  | | ASR | 10 | 278 | 12 (4.3%) | 2 (0.7%) |
|  | | AJS | 7 | 68 | 2 (2.9%) | 1 (1.5%) |
|  | | SQ | 2 | 5 | 2 (40.0%) | 0 |
| Year | | |  |  |  |  |
|  | 2014 | | 12 | 307 | 12 (3.9%) | 2 (0.7%) |
|  | 2015 | | 3 | 33 | 1 (3.0%) | 0 |
|  | 2016 | | 4 | 11 | 3 (27.3%) | 1 (9.1%) |
| Total | | | 19 | 351 | 17 (4.8%) | 3 (0.9%) |
| 1 The numbers of articles mentioned 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. | | | | | | |

*Publication bias*

Looking at ‘*AllP*’, there was no evidence for publication bias/*p*-hacking (see Figure 2A and Table 7). Overall, when using binwidth .01, 31 out of 73 results were just significant (43.8%). When using binwidth .02, 64 out of 127 results were just significant (50.4%). There was no evidence of publication bias among results related to explicitly stated hypotheses. A non-substantial indication of publication bias was found for results not related to explicitly stated hypotheses when using binwidth .02: out of 90 results, 47 (52.2%) were just significant (see Figure 2B). Next, we tested our hypothesis on publication bias (H3). There were 1/.877 ≈ 1.140 times less just significant *p*-values among reported results of explicitly stated hypotheses than among reported results not related to explicitly stated hypotheses for binwidth .01 (*b* = -.132, *p* = .794, OR = .877, 95% CI [.321, 2.345]) and binwidth .02 (*b* = -.251, *p* = .521, OR = .778, 95% CI [.358, 1.674]). For results from ‘*Hyp*’, we found slightly more just significant *p*-values than just insignificant ones - 9 out of 14 results (64.3%) for binwidth .01, and 14 out of 27 (53.8%) for binwidth .02, see Figure 3 and Table 7.

Graphical user interface

Description automatically generated with medium confidence

Figure 2. Histograms of numbers of p-values in the range 0 - .10 based on data from ‘AllP’. The histograms provide information on exactly reported p-values in general (2A), reported p-values that are not related to explicitly stated hypotheses (2B), and reported p-values that are related to explicitly stated hypotheses (2C) (n between brackets).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Table 7*. Numbers of *p*-values in the ranges (.04 - .06] and (.03 - .07], also split by explicitly stated hypothesis, journal, and year, which are relevant for studying publication bias among automatically and manually retrieved reported results. | | | | | | | |
|  |  | **AllP** | | | | | |
|  |  | Total | (.04 - .05] | (.05 - .06] | Total | (.03 - .05] | (.05 - .07] |
| Result hypothesis | |  |  |  |  |  |  |
|  | Yes | 24 | 10 (41.7%) | 14 | 37 | 17 (45.9 %) | 20 |
|  | No | 49 | 22 (44.9%) | 27 | 90 | 47 (52.2%) | 43 |
| Journal | |  |  |  |  |  |  |
|  | ASR | 20 | 8 (40%) | 12 | 29 | 11 (37.9%) | 18 |
|  | AJS | 7 | 3 (42.9%) | 4 | 14 | 8 (57.1%) | 6 |
|  | SQ | 5 | 1 (20%) | 4 | 7 | 3 (42.9%) | 4 |
|  | JMF | 32 | 14 (43.8%) | 18 | 57 | 28 (49.1%) | 29 |
|  | CHQ | 9 | 6 (66.7%) | 3 | 20 | 14 (70%) | 6 |
| Year | |  |  |  |  |  |  |
|  | 2014 | 29 | 13 (44.8 %) | 16 | 49 | 27 (55.1%) | 22 |
|  | 2015 | 19 | 10 (52.6%) | 9 | 38 | 20 (52.6%) | 18 |
|  | 2016 | 25 | 9 (36%) | 16 | 40 | 17 (42.5%) | 23 |
| Total | | 73 | 32: (43.8%) | 41 | 127 | 64 (50.4 %) | 63 |
|  | |  |  |  |  |  |  |
|  | | **Hyp** | | | | | |
|  | | Total | (.04 - .05] | (.05 - .06] | Total | (.03 - .05] | (.05 - .07] |
| Journal | |  |  |  |  |  |  |
|  | ASR | 5 | 2 (40.0%) | 3 | 11 | 5 (45.5%) | 6 |
|  | AJS | 3 | 3 (100.0%) | 0 | 5 | 4 (80.0%) | 1 |
|  | SQ | 6 | 4 (66.7%) | 2 | 10 | 5 (50.0%) | 5 |
| Year | |  |  |  |  |  |  |
|  | 2014 | 6 | 4 (66.7%) | 2 | 13 | 8 (61.5%) | 5 |
|  | 2015 | 2 | 1 (50.0%) | 1 | 2 | 1 (50.0%) | 1 |
|  | 2016 | 6 | 4 (66.7%) | 2 | 11 | 5 (45.5%) | 6 |
| Total | | 14 | 9 (64.3%) | 5 | 26 | 14 (53.8%) | 12 |

Chart, histogram

Description automatically generated

Figure 3. Histogram of number of exactly p-values related to explicitly stated hypotheses in the range [0 - .10] from ‘Hyp’. The n is the total number of exactly reported p-values in the range [0, .10] of ‘Hyp’.

*Bump in p-values*

Table 8 shows there is no bump in *p*-values in ‘*AllP*’: the lower *p*-value intervals contained *p*-values 50% when using binwidth .01 and 34.8% when using binwidth .02. For results from ‘*Hyp*’, 9 out of 14 reported *p*-values (64.3%) were just significant for binwidth .01, see Figure 3 and Table 8. However, this provides no clear indication for the presence of a bump in *p*-values since little data were retrieved.For binwidth .02 in ‘*Hyp*’, 14 out of 37 reported *p*-values (37.8%) were just significant. For a more detailed overview, see Table 8.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Table 8*. Numbers and percentages of *p*-values in the range (.03, .05] and (.01 - .05], which are relevant for studying a bump in *p*-values. Data sets used are ‘*AllP*’ and ‘*Hyp*’. | | | | | | | |
|  |  | **AllP** | | | | | |
|  |  | Total | (.03 - .04] | (.04 - .05] | Total | .01 - .03 | .03 - .05 |
| Result hypothesis | |  |  |  |  |  |  |
|  | Yes | 17 | 7 | 10 (56.2%) | 51 | 34 | 17 (33.3%) |
|  | No | 47 | 25 | 22 (46.8%) | 133 | 86 | 47 (35.3%) |
| Journal | |  |  |  |  |  |  |
|  | ASR | 11 | 3 | 8 (72.7%) | 36 | 25 | 11 (30.6%) |
|  | AJS | 8 | 5 | 3 (37.5%) | 31 | 23 | 8 (25.8 %) |
|  | SQ | 3 | 2 | 1 (33.3%) | 4 | 1 | 3 (75%) |
|  | JMF | 28 | 14 | 14 (50%) | 72 | 44 | 28 (38.9%) |
|  | CHQ | 14 | 8 | 6 (42.9%) | 41 | 27 | 14 (34.1%) |
| Year | |  |  |  |  |  |  |
|  | 2014 | 27 | 14 | 13 (48.1%) | 78 | 51 | 27 (34.6%) |
|  | 2015 | 20 | 10 | 10 (50.0%) | 61 | 41 | 20 (32.8%) |
|  | 2016 | 17 | 8 | 9 (52.9%) | 45 | 28 | 17 (37.8%) |
| Total | | 64 | 32 | 32 (50.0%) | 184 | 120 | 64 (34.8%) |
|  | |  |  |  |  |  |  |
|  | | **Hyp** | | | | | |
|  | | Total | .03 - .04 | .04 - .05 | Total | .01 - .03 | .03 - .05 |
| Journal | |  |  |  |  |  |  |
|  | ASR | 5 | 3 | 2 (40.0%) | 14 | 9 | 5 (35.7%) |
|  | AJS | 4 | 1 | 3 (75.0%) | 15 | 11 | 4 (26.7%) |
|  | SQ | 5 | 1 | 4 (80.0%) | 8 | 3 | 5 (62.5%) |
| Year | |  |  |  |  |  |  |
|  | 2014 | 8 | 4 | 4 (50.0%) | 22 | 14 | 8 (36.4%) |
|  | 2015 | 1 | 0 | 1 (100.0%) | 8 | 7 | 1 (12.5%) |
|  | 2016 | 5 | 1 | 4 (80.0%) | 7 | 2 | 5 (71.4%) |
| Total | | 14 | 5 | 9 (64.3%) | 37 | 23 | 14 (37.8%) |



*Marginal significance*

As Table 9 shows for ‘*AllP*’, in 46 of the 107 articles (43.0%) with reported *p-*values in the range (.05 - .10], at least one such *p-*value was reported as marginally significant. For results split out by journal and year, see Table 9. Out of 206 results with reported *p-*values in the range (.05 - .10], 72 (35.0%) were reported as marginally significant. Among results not related to hypotheses, 52 out of 136 (38.2%) *p*-values in the range (.05 - .10] were reported as marginally significant. Thus, our automatically retrieved results suggest assigning marginal significance occurs regularly in sociology articles. Next, our hypothesis that assignment of marginal significance will be more prevalent among results related to explicitly stated hypotheses (H4) is not confirmed. The odds of a result with a *p*-value in the range (.05 - .10] not related to an explicitly stated hypothesis being assigned marginal significance were 1.548 times higher than the odds that a result related to an explicitly stated hypothesis with a *p*-value in this range was assigned marginal significance, *b* = -.437, *p* = .170, OR = .646, 95% CI [.342, 1.194].

Next, we discuss the prevalence of marginal significance in ‘*Hyp*’. Looking at Table 9, in 19 of the 30 articles with reported *p*-values in the range (.05 - .10] from ‘*Hyp*’ (63.3%), at least one *p*-value was reported as marginally significant. For results at the article level split out by journal and year, see Table 9. At the results level, Table 9 shows that overall, 106 (81.5%) of 130 reported *p*-values in the range (.05 - .10] were reported as marginally significant. Looking at the different journals, 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%). Thus, results from ‘*Hyp*’ suggest assigning marginal significance occurs regularly among results related to explicitly stated hypotheses in sociology articles.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Hyp* | | | | | | | | |
|  |  |  | | | | | | |
|  |  |  | | | | | | |
|  |  |  | | |  |  | | |
|  |  |  |  |  |  |  |  |  |
|  | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | |  |  |  |  |  |  |  |
|  | | | | | | | | |
|  | | **Hyp** | | | | | | |
|  | |  | | |  |  | | |
|  | |  |  |  |  |  |  |  |
|  | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | |  |  |  |  |  |  |  |
| (2019). | | | | | | | | |

*Discussion and conclusions*

In this article, we studied different aspects of the quality of statistical reporting in sociology. High-quality statistical reporting entails that results are reproducible (and thus, easy to check), that results do not contain errors, and that clear communication and critical evaluation of results within a discipline is possible due to standardized reporting. We used different data sets to study statistical reporting quality in sociology. Manual data consisted of a data set with information on which journals request authors to adhere to the APA statistical reporting guidelines and of ‘*Hyp*’, which contains *p*-values and fully reproducible results related to explicitly stated hypotheses from the 2014-2016 volumes of ‘*ASR*’, ‘*AJS*’, and ‘*SQ*’. Furthermore, we created two data sets with data retrieved by statcheck from the 2014-2016 volumes of ‘*ASR*’, ‘*AJS*’, ‘*SQ*’, ‘*JMF*’ and ‘*CHQ*’: ‘*APA*’, which contains all in-text APA-reported results, and ‘*AllP*’, which contains all in-text automatically retrievable *p*-values. In this section, we will first discuss our conclusions regarding the aspects of statistical reporting we studied (requested adherence by sociology journals to the APA statistical reporting guidelines, statistical reporting errors, publication bias/*p*-hacking, the bump in *p*-values, and marginal significance). If differences between years and journals are not mentioned for specific topics, this is because they were not substantial or were based on too few data. Next, we will discuss the limitations of our study and provide some recommendations for future research.

Firstly, there is a lack of requested adherence to statistical reporting guidelines within sociology journals: only 9.1% required adherence to the APA statistical reporting guidelines. This could lead to lower quality of statistical reporting than that generally found in psychology, where journals commonly require authors to follow the APA statistical reporting guidelines. Statistical reporting guidelines namely help authors in reporting results such that all information necessary for their recalculation is present. This reproducibility allows third parties to quickly determine whether a result is reported (in)correctly, which, through social control measures, likely motivates authors to thoroughly check whether their results are reported correctly. This, in turn, likely increases statistical reporting quality.

Among results from ‘*APA*’, 13.7% were inconsistent and 1.6% grossly inconsistent. This is roughly comparable to the finding of Veldkamp et al. (2014) that 10.6% of results in psychology papers were inconsistent and 0.8% grossly inconsistent. In our study, approximately 82.4% of APA-reported results belonged to the two APA-journals. We were likely unable to recalculate some reproducible in-text results from other journals because they were not APA-reported. Contrary to expectations, we could not confirm H1 and H2, since there was no difference in (gross) inconsistencies for results related to explicitly stated hypotheses and results not related to explicitly stated hypotheses. Our inability to confirm our hypothesis on gross inconsistencies (H2) was possibly hampered by only 8 gross inconsistencies being observed. Among results from non-APA journals in ‘*Hyp*’, 4.2% were inconsistent and 0.8% grossly inconsistent.

Next, we studied the presence of publication bias and/or *p*-hacking in sociology, which can be indicated by there being more just significant than just non-significant *p*-values. In ‘*AllP*’, no evidence of publication bias/*p*-hacking was found. This is in line with the findings of Lakens’ (2015b), who did not find evidence of publication bias in psychology when reevaluating Masicampo & Lalande (2012). Furthermore, when testing H3, no evidence of a difference in publication bias between results of explicitly stated hypotheses and results not related to explicitly stated hypotheses was found. Few *p*-values related to explicitly stated hypotheses in the binwidths relevant for studying publication bias were present in ‘*Hyp*’(14 for binwith .01 and 26 for binwidth .02). This made it impossible to replicate Gerber and Malhotra’s (2008) finding that there were 3.25 to 4 more *z*-valuescorresponding to the *p*-value interval (.04-.05] than *z*-values corresponding to the *p*-value interval (.05-.06] in ‘*ASR*’, ‘*AJS*’, and ‘*SQ*’. Although Gerber and Malhotra (2008)’s data set contained only 46 articles and ours 91 articles, they retrieved more results relevant for studying publication bias: 56 for the 5% caliper (which approximates binwidth .01), and 106 for the 10% caliper (which approximates binwidth .02). Perhaps Gerber and Malhotra (2008) had more relevant data because they used *z*-values and *t*-values, which tend to be exactly reported. Thus, they did not, like us, lose data due to most *p*-values in sociology being reported inexactly. Furthermore, Gerber and Malhotra (2008) might have had more relevant data because they included *z*-values calculated using regression coefficients and standard errors, while we only included reported *p*-values. Finally, Gerber and Malhotra (2008) did not specify whether they only calculated *z*-values if it was explicitly clear that a *z*-distribution or *t*-distribution was used, or if they also calculated *z*-values based on an ‘educated guess’ that one of these distributions was used in a particular analysis. The latter method, though somewhat risky, would have likely yielded relatively many *z*-values and *t*-values relevant for studying publication bias.

Subsequently, we studied the bump in *p*-values. Often, we had too few data to properly study the bump, and when enough data was available, clear indications of a ‘bump’ were not found. This lack of clear indications of a ‘bump’ is in line with Lakens’ (2015b) findings in psychology and with our expectation of finding neither a ‘bump’ in *p*-values in general, nor a difference in the presence/size of a ‘bump’ between results related to hypotheses and results not related to hypotheses. However, the absence of a ‘bump’ in *p*-values alone does not indicate an absence of *p*-hacking. While this study has not found convincing evidence of *p*-hacking, it would be interesting to conduct a study containing more sociology journals from several years to see if this holds more generally.

Assignment of marginal significance to *p*-values in the range (.05 - .10] by authors was rather common. In 43.0% of ‘*AllP*’ and 63.3% of ‘*Hyp*’ articles with *p*-values in this range, marginal significance was assigned at least once. The 63.3% from ‘*Hyp*’ is largely based on results from tables and is much higher than the 10% of articles using ‘*p* < .10’ found by Leahey (2005) for articles from two sociology journals published in 1995-2000. Thus, it seems the ‘three-star system’ for assigning significance (i.e., \**p* < = .05, \*\**p* < = .01, \*\*\**p* < = .001), which is often used in captions of tables in sociology (see Leahey 2005), is frequently used with the likely unwarranted addition of ‘*p* < .10’. Of all results with *p*-values in the range (.05 - .10] in ‘*AllP*’, 35.0% were assigned marginal significance. This is only slightly lower than the almost 40% found by Ohlsson Collentine et al. (2019) in psychology. However, Ohlsson Collentine et al. (2019) studied many more journals and articles than we did, and it is unknown if our results would generalize to the entire sociological field. It should be noted, though, that there are no indications that in-text reporting practices regarding assignment of marginal significance differ much throughout sociology. In ‘*Hyp*’, 81.5% of results with *p*-values in the (.05 - .10] range were assigned marginal significance. This is much higher than the 35% of ‘*AllP*’. Possibly, this can be explained by potential awareness among authors that *p*-values in the range (.05 - .1] are of low evidential value, which prevents them from assigning marginal significance explicitly in articles’ text (Ohlsson Collentine et al., 2019). Assignment of marginal significance in tables and figures, however, is done more implicitly through symbols, which might make it seem less harmful to authors. Finally, we found no support for H4 based on ‘*AllP*’: authors did not seem more likely to assign marginal significance to results related to explicitly stated hypotheses than to results not related to explicitly stated hypotheses.

Concluding, we identified two statistical reporting issues in sociology: a lack of requested adherence to statistical reporting guidelines by journals and common usage of marginal significance by authors. A lack of statistical reporting guidelines leads to reproducibility issues, as became evident during our data collection process: if a single set of adequate statistical reporting guidelines (such as the APA-guidelines) would have been in place in sociology, we would have been able to retrieve substantially more reproducible results from articles. Having one set of statistical reporting guidelines in a scientific discipline namely provides a standardized way of presenting statistical information needed for the recalculation of results. Thereby, it allows for the development and use of statistical checking programs to check for errors and to retrieve *p*-values and fully reported results for reproductive and meta-analytical purposes. If the APA-guidelines were adopted, statcheck could immediately be used to automatically retrieve such information. To ensure reproducibility of results, any chosen set of statistical reporting guidelines should require that authors mention for each analysis what distribution and sidedness of testing is used, and mention for each statistical result what its test statistic is, what the degrees of freedom are, and what the *p*-value is. Presently, this information is often only partially available in sociology articles. Statistical reporting guidelines should also require reporting *p*-values exactly unless *p* < .001, because exactly reported *p*-values are more informative than inexactly reported *p*-values and *p*-values reported as ‘ns’, which are common in sociology. Furthermore, exactly reported *p*-values can be recalculated with more precision and can be relevant for meta-analytical research on publication bias and the bump in *p*-values (when studying these topics, inexactly reported *p*-values cannot be used). Finally, introducing one set of statistical reporting guidelines in sociology could provide a solution to the assignment of marginal significance by journals explicitly prohibiting it.

*Limitations & suggestions for future research*

One limitation of our study is that we retrieved relatively little data on publication bias/*p*-hacking and the bump in *p*-values, rendering us unable to provide firm conclusions on the presence of these phenomena in sociology. As mentioned previously, this is likely due to most *p*-values in sociology being reported inexactly. Also, we did not include potentially relevant control variables in our logistic regression analyses because we had too few data for including multiple predictors. These data issues could be mitigated in future research by the sociological discipline enforcing proper statistical reporting guidelines, which will make more reproducible results and exactly reported *p*-values available.

Another limitation concerns comparisons of our automatically retrieved findings to automatically retrieved findings in psychology. Within psychology, it is much more common to report *p*-values and APA-reported results in text than within sociology, where such data are more commonly found in tables. Thus, even though enough data were automatically retrieved to study statistical reporting errors and marginal significance, comparing these data to data retrieved in psychology may have led to comparisons of rather selective sets of in-text APA-reported results and *p*-values (likely those deemed most important by authors) from sociology to more complete sets from psychology. Thus, it is unlikely one-to-one translations from the parts of our study using automatic retrieval to studies with similar methods in psychology can be made.

Automatic retrieval of statistical information using statcheck in sociology was not necessarily efficient in our study. Since most reported *p*-values in sociology papers can be found intables and figures, we retrieved only a relatively small, and likely selective, set of in-text *p*-values using statcheck. Furthermore, APA-reported results retrieved by statcheck logically belonged mostly to APA-journals. Using manual retrieval, however, quite some reproducible results related to explicitly stated hypotheses from *ASR* and *AJS* were retrieved. This is due to two factors: some in-text reproducible results in these journals were not APA-reported, and most reproducible results in these journals were reported in tables and figures. Thus, future research on statistical reporting quality in sociology would ideally focus on retrieving statistical information from text, tables, and figures. For reproductive and meta-analytical purposes, it is vital that software for extracting *p*-values and reproducible results from tables and figures will be developed and/or improved. Only the ability to automatically (rather than manually) extract statistical data from figures and tables will enable high-quality non-tedious assessment of statistical reporting quality in sociology. Such assessment can help safeguard statistical reporting quality through social control mechanisms. Steps towards automatically extracting data from figures for meta-analytical purposes have been undertaken in recent years. Pick, Nakagawa and 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, however, 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 and 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.

It would also 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**

Clarivate Analytics (2017, October 10). Journal Citation Reports: Sociology, 2016.

Retrieved from -com.proxy.library.uu.nl/JCRJour

nalHomeAction.action?SID=A1-DQgg6Fhwox2FYqx2BFbS9RlcedDgx2FWFGn

HWI-18x2dv1xxye6CUOW8Va0lgrsEC0Qx3Dx3DqHp2KCJGPsLx2BdZTCxxa4

jqgx3Dx3D-YwBaX6hN5JZpnPCj2lZNMAx3Dx3D-jywguyb6iMRLFJm7wHskH

Qx3Dx3D&refineString=null&SrcApp=IC2LS&timeSpan=null&Init=Yes&wsid=N1CIgDR1OV5XpHOAk77.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2018). “Redefine statistical significance.” *Nature Human Behaviour*, *2*(1), 6.

Cohen, Jacob. 1992. “A power primer.” *Psychological bulletin*, 112(1): 155-159.

Epskamp, Sacha and Michèle. B. Nuijten. 2015. “statcheck: Extract statistics from

articles and recompute p values*.* R package version 1.0.1.” Retrieved October 10,

2017 (http://CRAN.R-project.org/package=statcheck).

Fanelli, Daniele. 2011. “Negative results are disappearing from most disciplines and

countries.” Scientometrics, 90(3): 891-904.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication bias in the social

sciences: Unlocking the file drawer." *Science*, 6203(345): 1502-1505.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2016. "Underreporting in psychology

experiments: Evidence from a study registry." *Social Psychological and Personality*

*Science*, 7(1): 8-12.

Gerber, Alan S. and Neil Malhotra. 2006. “Can political science literatures be

believed? A study of publication bias in the APSR and the AJPS.” Presented at

the Annual Meeting of the Midwest Political Science Association, Chicago, IL,

April.

Gerber, Alan. S. and Neil Malhotra. 2008. “Publication bias in empirical sociological

research: Do arbitrary significance levels distort published results?” *Sociological Methods & Research*, 37(1): 3-30.

Hartgerink, Chris. H. J. 2017. “2016 marginal, extraction functions: extract results.”

**Retrieved 13 December, 2017 (https://github.com/chartgerink/2016marinal/blob/m**

**aster/).**

Hartgerink, Chris. H. J., Jelte M. Wicherts, & Marcel A. L. M. Van Assen. 2017.

“Too good to be false: Non-significant results revisited.” *Collabra: Psychology*, 3(1): 1-18. https://doi.org/10.1525/collabra.71.

Hartgerink, Chris. H. J., Robbie C. M. Van Aert, Michèle. B. Nuijten, Jelte M.

Wicherts and Marcel A. L. M. Van Assen. 2016. “Distributions of p-values

smaller than. 05 in psychology: what is going on?” *PeerJ*, 4: 1–28. [doi:10.7717/peerj.1935](https://doi.org/10.7717/peerj.1935).

John, Leslie. K., George Loewenstein and Drazen Prelec. 2012. “Measuring the

prevalence of questionable research practices with incentives for truth telling.”

*Psychological science*, 23(5): 524-532.

Lakens, Daniël. 2015a. “On the challenges of drawing conclusions from p-values just

below 0.05.” *PeerJ*, *3*: 1-14. doi:10.7717/peerj.1142.

Lakens, Daniël. 2015b. “What p-hacking really looks like: A comment on Masicampo

and LaLande (2012).” *The Quarterly Journal of Experimental Psychology*, *68*(4):

829-832.

Lang, Thomas A. and Douglas G. Altman. 2013. “Basic statistical reporting for

articles published in clinical medical journals: the SAMPL Guidelines.” Pp. 175–182 in *Science Editors' Handbook*, edited by P. Smart, H. Maisonneuve and A. Polderman. Split, Croatia, European Association of Science Editors.

Leahey, Erin. 2005. “Alphas and asterisks: the development of statistical significance

testing standards in sociology.” *Social Forces*, 84(1): 1-24.

Leggett, Nathan C., Nicole A. Thomas, Tobias Loetscher, and Michael E. R. Nicholls.

2013. “The life of p: “just significant” results are on the rise.” *The Quarterly Journal of Experimental Psychology*,66(12): 2303-2309.

Masicampo, E. J. and Daniel R. Lalande. 2012. “A peculiar prevalence of p values

just below. 05.” *The Quarterly Journal of Experimental Psychology*, 65(11): 2271-2279.

Nuijten, Michèle. B., Chris H. J. Hartgerink, Marcel A. L. M. Van Assen, Sacha

Epskamp and Jelte M. Wicherts. 2016. “The prevalence of statistical reporting

errors in psychology (1985–2013).” *Behavior research methods*, 48(4): 1205-1226.

Nuijten, Michèle. B., Jeroen Borghuis, Coosje L. S. Veldkamp, Linda Dominguez

Alvarez, Marcel A. L. M. Van Assen and Jelte M. Wicherts. 2017.

“Journal data sharing policies and statistical reporting inconsistencies in

psychology”. Collabra: Psychology, 3(1): 1–22.

Ohlsson Collentine, Anton, Marcel A. L. M. Van Assen M. and Chris H. J.

Hartgerink. 2019. “The prevalence of marginally significant results in

psychology over time.” *Psychological science*, 30(4): 576-586.

Pick, Joel L., Shinichi Nakagawa and Daniel W. A. Noble. 2018. “Reproducible,

flexible and high throughput data extraction from primary literature: The

metaDigitise package.” *Methods in ecology and evolution*, 10(3): 426-431.

Pritschet, Laura, Derek Powell and Zachart Horne. 2016. “Marginally significant

effects as evidence for hypotheses: Changing attitudes over four decades.”

*Psychological science*, 27(7): 1036-1042.

R Core Team. 2017. “R: The R Stats Package”. Vienna, Austria: R Core Team.

Retrieved October 10, 2017 (<https://stat.ethz.ch/R-manual/R-devel/library/stats/>

html/TDist.html).

Roldán, Juan C., Patricia Jiménez, and Rafael Corchuelo. 2020. "On extracting data

from tables that are encoded using HTML." *Knowledge-Based Systems*, 190:

105157.

Sedlmeier, Peter & Gerd Gigerenzer. 1989. “Do studies of statistical power have an

effect on the power of studies?.” *Psychological bulletin*, 105(2): 309-316.

Simera, Iveta, David Moher, Allison Hirst, John Hoey, Kenneth F. Schulz, and

Douglas G. Altman. 2010. “Transparent and accurate reporting increases

reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network.” *BMC medicine*, 8(1): 24-30.

Simmons, Joseph P., Leif D. Nelson & Uri Simonsohn. 2011. “False-positive

psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant.” *Psychological science*, *22*(11): 1359-1366.

Veldkamp, Coosje. L., Michèle B. Nuijten, Linda Dominguez-Alvarez, Marcel A. L.

M. Van Assen and Jelte M. Wicherts. 2014). “Statistical reporting errors and

collaboration on statistical analyses in psychological science.” *PloS one*, 9(12): 1-

19. doi:10.1371/journal.pone.0114876.

Wicherts, Jelte M., Marjan Bakker and Dylan Molenaar. 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. doi:10.1371/journal.pone.0026828.

De Winter, Joost C. F. and Dimitra Dodou. 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. doi:0.7717/peerj.73.

**APPENDIX**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Table A1*. Logistic regressions used to test hypotheses related to (gross) inconsistencies using the ‘*APA*’ dataset (*n* = 505). | | | | | |
| **Inconsistencies (H1)** | | | | | |
|  | *B* | SE | *p*-value | OR | 95% CI OR |
| Intercept | -1.820 | .157 | <.001 | .162 | [.118, .218] |
| Hypothesis | -.073 | .278 | .793 | .930 | [.531, 1.585] |
| **Gross inconsistencies (H2)** | | | | | |
|  | *B* | SE | *p*-value | OR | 95% CI OR |
| Intercept | -4.422 | .503 | <.001 | .012 | [.004, .028] |
| Hypothesis | .708 | .714 | .321 | 2.030 | [.475, 8.685] |
| *Note*. OR = odds ratio, SE = standard error. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Table A2*. Logistic regressions used to test hypothesis related to publication bias (H3) using exactly reported *p*-values from ‘*AllP*’. | | | | | |
| **Binwidth .01, (.04 - .05] vs (.05 - .06]** | | | | | |
|  | *B* | SE | *p*-value | OR | 95% CI OR |
| Intercept | -.205 | .287 | .476 | .815 | [.460, 1.428] |
| Hypothesis | -.237 | .515 | .645 | .789 | [.281, 2.147] |
| *N* | 72 | | | | |
| **Binwidth .02, (.03 - .05] vs (.05 - .07]** | | | | | |
|  | *B* | SE | *p*-value | OR | 95% CI OR |
| Intercept | .089 | .211 | .673 | 1.093 | [.723, 1.658] |
| Hypothesis | -.312 | .396 | .431 | .732 | [.333, 1.588] |
| *N* | 126 | | | | |
| *Note*. OR = odds ratio, SE = standard error. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Table A3*. Logistic regression used to test the hypothesis related to marginal significance (H4) using reported *p*-values1 from ‘*AllP*’. | | | | | |
|  | *B* | SE | *p*-value | OR | 95% CI OR |
| Intercept | -.514 | .181 | .004 | .598 | [.417, .848] |
| Hypothesis | -.453 | .324 | .163 | .636 | [.332, 1.188] |
| *N* | 206 | | | | |
| *Note*. 1 Reported *p*-values used were those in smaller than or equal to a *p*-value in the (.05 - .10] range. OR = odds ratio, SE = standard error. | | | | | |

1. At first, we had collected articles from *CHQ* and *Work and Occupations* (*WOX*), which has an impact factor of 2.355 and thus ranks second. However, extracting results from *WOX* articles turned out to be unachievable due to compatibility issues with statcheck. For an unknown reason, no results could be extracted from neither the HTML nor PDF versions of the *WOX* articles. [↑](#footnote-ref-1)
2. Note that this includes both exactly reported *p*-values, such as, ‘*p* = .07’ and inexactly reported *p*-values, such as ‘*p* < .07’. [↑](#footnote-ref-2)