The reporting of statistical results in sociology: a systematic review

Elise Schramkowski Supervisor: Marcel van Assen

14-02-2022

Abstract

High quality statistical reporting in scientific articles is vital for providing scientific and nonscientific communities with correct information on studied phenomena. In this article, the quality of statistical reporting 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 automatically retrieved from the 2014-2016 volumes of five sociology journals using the R package statcheck (Epskamp & Nuijten, 2016). Furthermore, information on these topics was retrieved manually for the 2014-2016 volumes of three sociology journals previously studied in the context of publication bias in sociology by Gerber & Malhotra (2008). It was found that only 13 of these journals (9.1%) adhered to statistical reporting guidelines (i.e., the APA guidelines). No convincing evidence of the presence of publication bias and a 'bump' in p-values was found. Marginal significance was rather prevalent, especially when manually studying all results of articles related to explicitly stated hypotheses: then, marginal significance was assigned to 81.5% of results with p-values in the range (.05-.10]. Across our data, more than 40% of articles contained at least one p-values in the (.05-.10] range to which marginal significance was assigned.

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

Statistical results in scientific papers provide scientific and nonscientific communities with essential information about studied phenomena. Statistical results should comply with the following 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., 2010). Secondly, statistical results should not contain errors, because such inaccuracies in research 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 papers should be standardized at least within disciplines to enable authors to clearly communicate statistical results and to enable readers to critically evaluate them. In this systematic review, we examined several aspects of statistical results quality in sociology, namely adherence to statistical reporting guidelines, prevalence of statistical reporting errors, evidence of publication bias and p-hacking, 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). On the other hand, absence of clear statistical reporting guidelines leads to authors providing insufficient information when reporting statistics, making critical assessment of their results difficult (Simera et al., 2010). Thus, having statistical reporting guidelines may well lead to better statistical reporting quality in a discipline. Contrary to psychology, no 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. We examined which journals request authors to adopt the APA 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 inconsis-

tency 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 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. 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 '$t(50) = 1.99$, $p = .049$'. This suggests a statistically significant result, but $t(50) = 1.99$ implies $p = .052$, which implies that H_0 should not be rejected. To our knowledge, no research on the prevalence of statistical reporting errors in sociology has been conducted at present. However, research in psychology has found that 4%-10% of results are inconsistently reported (Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2016; Wicherts, Bakker, & Molenaar, 2011). Gross inconsistencies have been found in 0.8%-2.5% of reported results (e.g., Veldkamp, Nuijten, Dominguez-Alvarez, Van Assen, & Wicherts, 2014; Hartgerink, Van Aert, Nuijten, Wicherts, & Assen, 2016; Nuijten et al., 2016). Nuijten et al. (2016) found that gross inconsistencies occur relatively often in statistically significant reported results; due to gross inconsistencies, the percentage of significant *p*-values among recalculated *p*-values was 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 (Hartgerink et al., 2016; John, Loewenstein, & Prelec, 2012; Nuijten et al., 2016). We studied the prevalence of statistical reporting errors in a selection of APA and non-APA journals in two ways: manually, and with the R package statcheck (Epskamp & Nuijten, 2016), which automatically checks the consistency of fully APA-reported results.

Publication bias occurs when statistically significant results are published relatively

more often than non-significant ones. It is one of the suboptimal research/publishing practices that can lead to a relatively high prevalence of just significant p-values. In various fields of scientific research, especially in the social and biomedical sciences, publication bias has been found to some extent (e.g., Dickersin, 1990; De Winter & Dodou, 2015; Easterbrook, Gopalan, Berlin, & Matthews, 1991; Fanelli, 2011; Franco, Malhotra, & Simonovits, 2014, 2016; Masicampo & Lalande, 2012). 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 et al., 2012; Lakens, 2015a; Masicampo & Lalande, 2012) and the use of researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 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 & Malhotra (2008), who studied 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 number of z-values in an interval that closely approximates the p-value interval (.04-.05] to the number of z-values in an interval that closely approximates the p-value interval (.05-.06]. They found that the number 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. Similarly, the p-value intervals (.04-.05] and (.05-.06] were compared Masicampo & Lalande (2012) in psychology and by De Winter & Dodou (2015) across disciplines. We studied these p-value ranges too, and, following Gerber & Malhotra (2008), the p-value ranges (.03-.05] and (.05-.07], since larger intervals provide higher power.

A non-monotonic increase or a 'bump' in p-values occurs when there are more p-values in a just significant p-value interval than in the adjacent lower p-value interval. 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' (Leggett, Thomas, Loetscher, & Nicholls, 2013; Masicampo & Lalande, 2012). 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 implies 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 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 slightly larger than .01 will be much 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, arguing non-significant results represent true effects 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 p-value range (.05-.10]; Pritschet et al. (2016) found that of p-values reported as marginally significant in psychology, 92.6% were in this interval. Ohlsson Collentine et al. (2019) found that almost 40% of p-values in the (.05-.10] range 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 range (.05-.10] 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.

We examined the prevalence of statistical reporting errors, publication bias/p-hacking, a 'bump' in p-values, and p-values reported as marginally significant among results of explicitly stated hypotheses (hypotheses referred to in the paper's text as hypotheses to be tested) and results not related to explicitly stated hypotheses. For statistical reporting errors, publication bias/p-hacking, and assignment of marginally significancz, we formally tested whether there were differences in the prevalence of these phenomena between results of explicitly stated hypotheses and other results. 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 related to hypotheses (Gerber & Malhotra, 2006), one would expect the prevalence of (gross) inconsistencies, publication bias/p-hacking, and marginal significance to be higher among results corresponding to explicitly stated hypotheses. More specifically, we hypothesized the following:

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 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 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, 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 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 adhered to the APA statistical reporting guidelines or not.

For our study on statistical reporting errors, publication bias/p-hacking, the 'bump' in p-values, and marginal significance, we collected data from articles of several journals. Since statcheck only retrieves and recalculates fully 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 journals 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[1]. We studied 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 the 322 articles from the 2014-2016 volumes of three non-APA journals from Clarivate Analytics' Web of Science (2016): *ASR*, *AJS*, and *SQ*. Gerber & Malhotra (2008) used these in their study on publication bias, which we wanted to replicate. Fully APA-reported results retrieved using statcheck were put into a data set called 'APA', and p-values retrieved 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 some fully APA-reported results are also included in 'Hyp', as this data set contains all statistical results related to explicitly stated hypotheses.

[1]Initially, we had collected articles from *CHQ* and Work and Occupations (*WOX*), which had the second highest impact factor (2.355). However, extracting results from *WOX* articles with statcheck was not possible due to compatibility issues. For an unknown reason, no results could be extracted from neither the HTML nor PDF versions of *WOX* articles.

Data collection

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 several different style guides. This information was put into a data set called 'SRG'. 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 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 (Epskamp & Nuijten, 2016) to automatically retrieve APA-reported results, reported p-values, and recalculated p-values. '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, and r), df, and reported p-values. Results with exactly reported p-values and results with p-values reported as '<', '>', or 'non-significant' were retrieved by statcheck. If p-values were reported as non-significant, statcheck assigned them the label 'NA'. 'APA' also contains recalculated p-values as retrieved by statcheck, as well as information on whether reported p-values are (grossly) inconsistent with their recalculated counterparts. If a reported result seemed inconsistent (and this cannot be due to correct rounding), statcheck applied a one-sided test to it. If this lead 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'. 'APA' 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 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
Results related to explicitly stated hypotheses?	partly	partly	yes
Total number of articles	471	80	91
Number of articles used	314	76	91
Total number of results	7280	524	4849
Number of results used	2959	505	4849

Table 2: Overview of the numbers of results used in the analyses of article-focused topics for 'AllP', 'APA', and 'Hyp'.

	'AllP'	'APA'	'Hyp'
Statistical reporting errors			
Descriptive information	-	505 (76)	399 (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 hypotheses publication bias (H3a & H3b)			
(.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. Numbers of articles from which results were used in analyses are shown between parentheses.

The third dataset, 'AllP', consists of all reported *p*-values of all five journals retrieved by statcheck. 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 statcheck, we removed 4,354 (59.8%) because they did not refer to reported *p*-values. Thus, 'APA' contained 2,926 reported *p*-values from 308 articles. From these data, descriptive information on publication bias, the 'bump' in *p*-values, and marginal significance was obtained. Furthermore, these data were used to test H3a, H3b, 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. Then, following Ohlsson Collentine et al. (2019), we decided that a p-value 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 articles with at least one p-value in the range (.05-.10] to which marginal significance was assigned.

A fourth data set, 'Hyp', was created to replicate the study of Gerber & Malhotra (2008) 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 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 in *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). 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 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.

Also, this allowed us to include p-values based on F-values, r-values, and χ^2-values. 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. We organized all 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, 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. We used 'Hyp' to study all our results-level phenomena (see Table 2). 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 assignment of marginal significance to in-text p-values in the range (.05-.10] to 'Hyp' as we did for 'AllP'. For p-values in tables, we considered significance levels of $p < .10$ in captions (indicated by, e.g., an asterisk) to be assignment of marginal significance. Finally, we studied the percentage of articles in 'Hyp' containing marginally significant results in the range (.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.

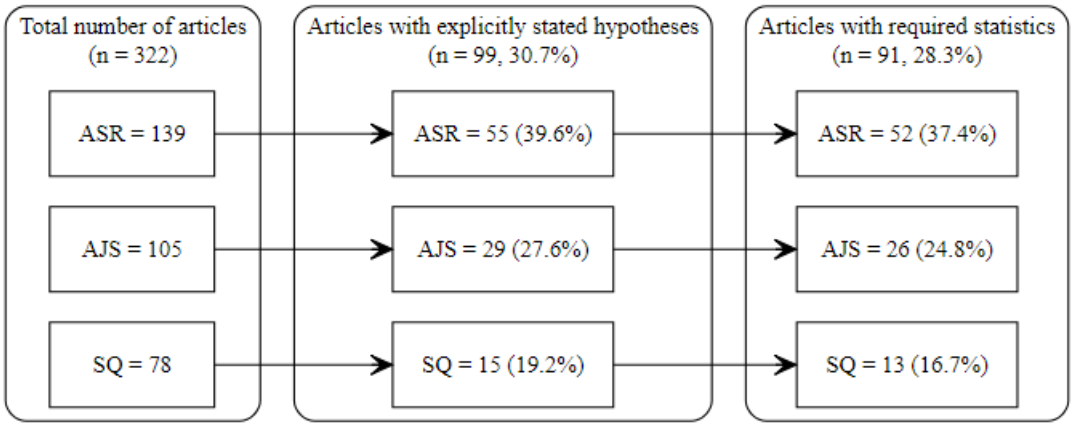

Figure 1: Flowchart describing the process of selecting articles from which results were retrieved for 'Hyp'.

Table 3: Conditions under which different types of reported p-values from 'Hyp' were considered (grossly) inconsistent.

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 are indicative of an inconsistency.

b & SE

Only used for recalculation if a result was explicitly based on the z-distribution or t-distribution.

We use $b = 3.11$, SE $= 2.11$, $p = 0.07$, from a z-distribution as an example:

- Correct Cal$P$ will 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 $z_{\text{lb}} = \frac{3.105}{2.115} = 1.468085$).

- Use t/z_{lb} and t/z_{ub} (and, in case of t, reported df) to calculate CalP_{lb} and CalP_{ub} (boundaries of correctly rounded Rep$P$). For this, the R stats package pt() function - in case of t - or the pnorm() function - in case of z - is used. Then, CalP_{lb} and CalP_{ub} are rounded 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 for recalculation of p-values: pt() for t and r, pnorm() for z, pf() for F, and chisq() for χ^2. All functions, except pnorm(), require information on df. We use the example of $t(61) = 3.11$, $p = 0.0001$:

- Correct Cal$P$ will 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, and 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$P$ is smaller than CalP_{lb} and 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_{lb} = lower bound SE, SE_{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 results were based on 'APA' and 'Hyp'. 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).

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. 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., 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 conducted logistic regressions to test our hypothesis 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. H3a was tested using the p-value intervals (.04-.05] and (.05-.06] to obtain more precise results, and H3b was tested using p-value intervals (.03-.05] and (.05-.07] for a potentially more powerful test.

To test our hypothesis on p-values reported as marginally significant (H4), we conducted logistic regressions with exactly reported p-values in the range (.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.

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 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 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 in 'SRG', one journal (Society) did not seem to have any explicit guidelines authors are 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 categories of 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%)

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)
Other	34 (23.8%)
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 provided on the website that should give 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, making the percentage of significant p-values in recalculated p-values 1.6 percentage points lower (30.3% rather than 31.9%) than that in reported p-values. 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. 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]. Of the recalculated p-values, 416 (82.4%) 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 reproducible results from 'Hyp', 17 out of 399 recalculated p-values were inconsistent (4.3%), and three (0.8%) were grossly inconsistent. For a comprehensive overview of results, see Table 6.

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	11	313	13 (4%)	2 (0.6%)
AJS	7	68	2 (2.9%)	1 (1.5%)
SQ	2	5	2 (40.0%)	0 (0%)
Year				
2014	12	307	12 (3.9%)	2 (0.7%)
2015	3	33	1 (3.0%)	0 (0%)
2016	5	59	3 (5.1%)	1 (1.7%)
Total	20	399	17 (4.3%)	3 (0.8%)

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.

Publication bias

In 'AllP', there was no evidence of 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%). 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). There was no evidence of publication bias among results related to explicitly stated hypotheses (see Figure 2C). Next, we tested our hypotheses on publication bias (H3a and H3b). As for H3a, there were $\frac{1}{.877} \approx 1.140$ 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]. H3b for binwidth .02 could not be rejected either, since 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 2D and Table 7.

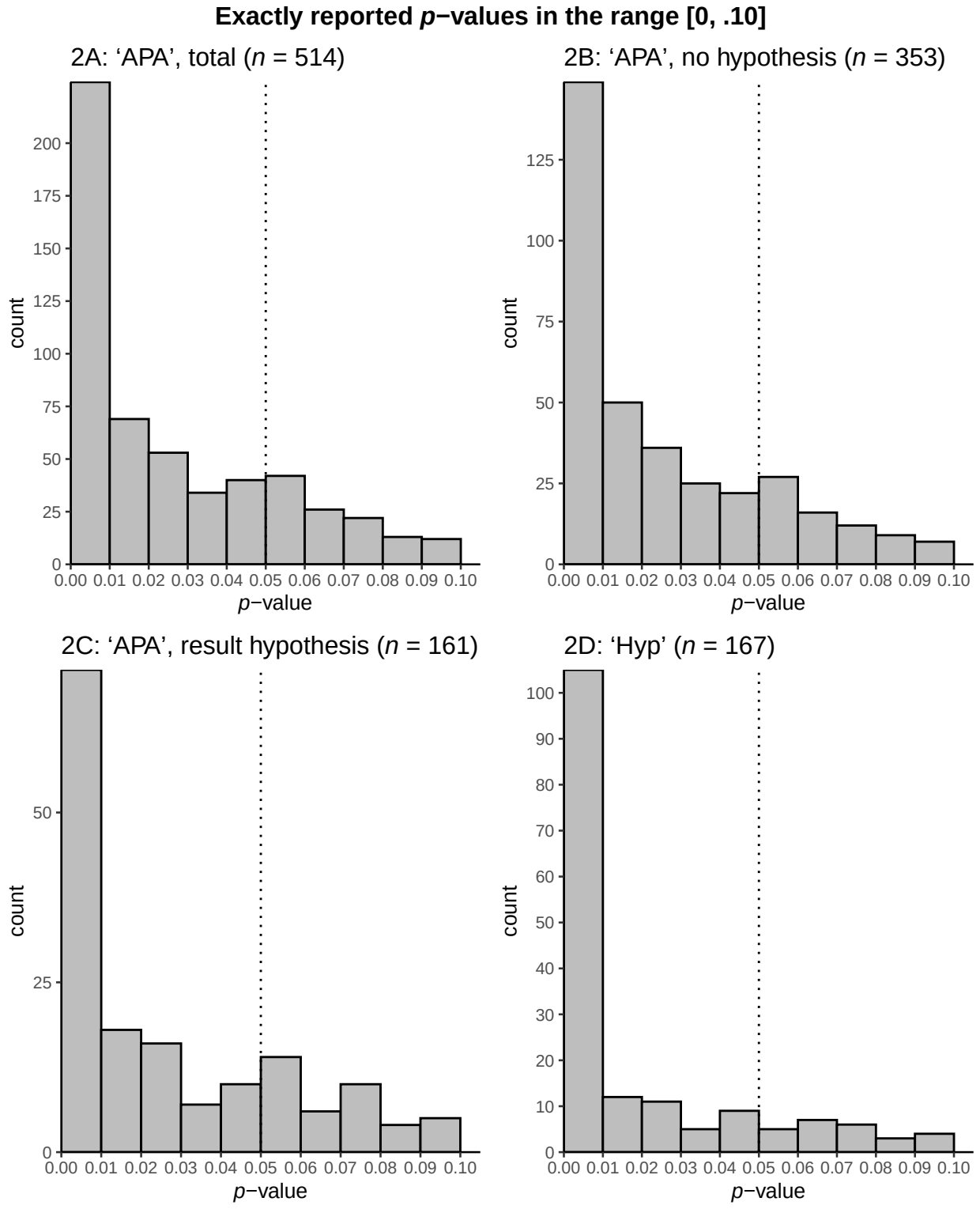

Figure 2: Histograms of exactly reported p-values in the range [0-.10] from 'AllP' and 'Hyp'. Specifically, information on 'AllP' is provided for all exactly reported p-values in the range [0-.10] (2A), as well as for exactly reported p-values in this range (not) related to explicitly stated hypotheses (2B and 2C). Finally, information on exactly reported p-values in the range [0-.10] from 'Hyp' (2D) is provided.

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%)	49	43	47 (52.2%)	43	90
Journal						
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%)	16	40
Total	32 (43.8%)	41	73	64 (50.4%)	41	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

References

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.

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.

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.

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867–872.

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

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904.

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

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.

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.

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

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.

Lakens, D. (2015a). On the challenges of drawing conclusions from p-values just below 0.05. *PeerJ*, (3), 1–14.

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.

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.

Leahey, E. (2005). Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces*, *84*(1), 1–24.

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.

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.

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.

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.

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. (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, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

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.

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]