UNIVERSITY of CALIFORNIA PRESS | Collabra

Nuijten, M. B. et al. (2017). Journal Data Sharing Policies and Statistical Reporting Inconsistencies in Psychology. *Collabra: Psychology,* 3(1): 31, pp. 1–22. DOI: https://doi.org/10.1525/collabra.102

ORIGINAL RESEARCH REPORT

Journal Data Sharing Policies and Statistical Reporting Inconsistencies in Psychology

Michèle B. Nuijten[*], Jeroen Borghuis[†], Coosje L. S. Veldkamp[*], Linda Dominguez-Alvarez[‡], Marcel A. L. M. van Assen[*,§] and Jelte M. Wicherts[*]

In this paper, we present three retrospective observational studies that investigate the relation between data sharing and statistical reporting inconsistencies. Previous research found that reluctance to share data was related to a higher prevalence of statistical errors, often in the direction of statistical significance (Wicherts, Bakker, & Molenaar, 2011). We therefore hypothesized that journal policies about data sharing and data sharing itself would reduce these inconsistencies. In Study 1, we compared the prevalence of reporting inconsistencies in two similar journals on decision making with different data sharing policies. In Study 2, we compared reporting inconsistencies in psychology articles published in PLOS journals (with a data sharing policy) and Frontiers in Psychology (without a stipulated data sharing policy). In Study 3, we looked at papers published in the journal Psychological Science to check whether papers with or without an Open Practice Badge differed in the prevalence of reporting errors. Overall, we found no relationship between data sharing and reporting inconsistencies. We did find that journal policies on data sharing seem extremely effective in promoting data sharing. We argue that open data is essential in improving the quality of psychological science, and we discuss ways to detect and reduce reporting inconsistencies in the literature.

Keywords: Statistical errors; data sharing; journal policy; meta-research

Most psychological researchers use Null Hypothesis Significance Testing (NHST) to evaluate their hypotheses (Cumming et al., 2007; Hubbard & Ryan, 2000; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). The results of NHST underlie substantive conclusions and serve as the input in meta-analyses, which makes it important that they are reported correctly. However, NHST results are often misreported. Several large-scale studies estimated that roughly half of psychology articles using NHST contain at least one *p*-value that is inconsistent with the reported test statistic and degrees of freedom, while around one in eight such articles contain a gross inconsistency, in which the reported *p*-value was significant and the computed *p*-value was not, or vice versa (Bakker & Wicherts, 2011; Caperos & Pardo, 2013; Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2016; Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014). In the medical sciences roughly one in three articles contains an inconsistent *p*-value (Garcia-Berthou & Alcaraz, 2004), and in psychiatry about one in ten articles (Berle & Starcevic, 2007).

There is evidence that inconsistent *p*-values are associated with reluctance to share data, especially when the inconsistencies concern statistical significance (Wicherts et al., 2011). Wicherts et al., speculated that it is possible that authors are reluctant to share data because they fear that other research teams will arrive at different conclusions, or that errors in their work will be exposed (see also Ceci, 1988; Hedrick, 1985; Sterling & Weinkam, 1990). Along these lines, one may expect that if authors intend to make their data available from the start, they will double-check their results before writing them up, which would result in fewer inconsistencies in the final paper. Wicherts et al., also offered the alternative explanation that the relation between data sharing and misreporting is caused by differences in the rigor with which data are managed; researchers who work more diligently in their handling and archiving of data are probably less likely to commit a reporting error.

In psychology, the availability of research data in general is already strikingly low (Vanpaemel, Vermorgen, Deriemaecker, & Storms, 2015; Wicherts, Borsboom, Kats, & Molenaar, 2006), although this problem is not limited to psychology (see e.g., Alsheikh-Ali, Qureshi,

[*] Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, NL

[†] Department of Developmental Psychology, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, NL

[‡] Ecorys, Rotterdam, NL

[§] Department of Sociology, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, NL

Corresponding author: Michèle B. Nuijten (m.b.nuijten@uvt.nl)

Al-Mallah, & Ioannidis, 2011). This is a worrying finding in itself, since the availability of original research data is essential to reproduce or verify analyses. However, this problem becomes worse if data are even less likely to be shared if the research article contained statistical inconsistencies, because in these cases verification of the analyses is even more important. Over the past few years there has been increasing awareness that the availability of research data is essential for scientific progress (Anagnostou et al., 2015; Nosek et al., 2015; Wicherts, 2011; Wilkinson et al., 2016), and several journals have started to request authors to share their data when they submit an article (e.g., in PLOS and Psychological Science; see Bloom, Ganley, & Winker, 2014; Lindsay, 2017; respectively). We theorized that such journal policies on data sharing could help decrease the prevalence of statistical reporting inconsistencies, and that articles with open data (regardless of journal policy) contained fewer inconsistencies.

In this paper, we present three retrospective observational studies that investigate the relation between data sharing and reporting inconsistencies. Our two main hypotheses were that 1) journals that encourage data sharing will show a (larger) decrease in inconsistencies and gross inconsistencies compared to similar journals that do not encourage data sharing (an open policy effect), and 2) articles that are accompanied with open data have fewer inconsistencies and fewer gross inconsistencies than articles without open data (an open data effect). We compared inconsistency rates between two similar journals on decision making with different data sharing policies (Study 1), between psychology articles from journals from the open access publisher PLOS that requires open data and Frontiers that has less strict data sharing policies (Study 2), and between papers in the journal Psychological Science with and without Open Practice Badges (Study 3). Studies 2 and 3 are pre-registered and the relevant registrations can be found at https://osf.io/538bc/. Exploratory findings across the three studies are reported in a final results section.

Study 1

In Study 1 we documented the prevalence of reporting inconsistencies in two similar journals on decision making that have different data sharing policies: the Journal of Behavioral Decision Making (JBDM; no data sharing policy) and Judgment and Decision Making (JDM; recommended data sharing). Furthermore, we compared the number of reporting inconsistencies in articles that actually did or did not include shared data, regardless of the journal they were published in. We hypothesized that JDM would show a (larger) decrease in inconsistencies and gross inconsistencies compared to JBDM after the introduction of the data sharing policy in JDM (open policy effect), and that articles that are accompanied with open data contain fewer inconsistencies and gross inconsistencies than articles that are not accompanied with open data (open data effect).

Method

Sample. We examined the relation between open data journal policy on statistical reporting inconsistencies in two similar psychological journals: JBDM (ISI impact factor in 2015: 2.768) and JDM (ISI impact factor in 2015: 1.856). Both journals focus on human decision processes and accept empirical research as well as theoretical papers. Furthermore, there is considerable overlap between their editorial boards: in 2015, seventeen researchers sat in the editorial boards of both JDM (51 members in total) and JBDM (125 members in total). A difference between the journals is that JDM is completely open access, whereas in JBDM the authors can pay a fee to make their article open access. The main difference of concern here, however, is that since 2011 JDM editors have started encouraging authors to submit their raw data at the time of review (Baron, 2011).[1] When the articles are accepted, these data are subsequently published on the web site along with the articles. Before 2011, there was no explicit data policy in JDM. JBDM did not adopt a similar data sharing policy in the relevant years.[2]

We downloaded the articles of JDM in the periods before and after their policy change, and we included articles from JBDM in the corresponding time periods. The first issue in JDM was published in 2006, and from April 2011 (Issue 3, 2011; corresponding to Issue 2 2011 of JBDM) onwards JDM started to implement the new data policy. We collected data in 2015, so we included papers up until the end of 2014 to include the most recent full year. Our final sample contained papers published in the years 2006 to February 2011 (T1), and in April 2011 to 2014 (T2). See **Table 1** for the number of articles collected per journal and time period. We included all research articles and special issue papers from these periods in both journals, but no book reviews and editorials. All articles of JDM were HTML files, whereas all articles of JBDM were PDF files because no HTML files were available in T1.

Procedure. For each article, we coded in which journal and time period it was published and whether the (raw) data were published alongside the articles. Published data files in matrix format with subjects in the rows (so no correlation matrices) as well as simulation codes and model codes were considered open data. The data had to be published either in the paper, an appendix, the journal's website, or a website with a reference to that website in the paper. Remarks such as "data are available upon request" were not considered open data (as such promises are often hollow; Krawczyk & Reuben, 2012).

Table 1: Number of articles (N) downloaded per journal and time period: 2006 to February 2011 (T1; published before open data policy of JDM), and from April 2011 to 2014 (T2; published after open data policy of JDM).

	N in T1	N in T2	Total N
JBDM	157	149	306
JDM	236	222	458
Total	393	371	764

Note that we did not assess whether any published data were also relevant, usable, and/or complete, which is by no means guaranteed (Kidwell et al., 2016).

We assessed the consistency of the reported statistical results through an automated procedure: an adapted version[3] of the R package "statcheck" (version 1.0.0; Epskamp & Nuijten, 2014). statcheck extracts NHST results and recomputes *p*-values in the following steps. First, statcheck converts PDF and HTML files into plain text files and extracts statistical results based on *t*-tests, *F*-tests, correlations, *z*-tests, and χ^2-tests that are reported completely (i.e., test statistic, degrees of freedom, and *p*-value) and according to the guidelines in the APA Publication Manual (American Psychological Association, 2010). Next, the extracted *p*-values are recomputed based on the reported test statistic and degrees of freedom. Finally, statcheck compares the reported and recomputed *p*-value, and indicates whether they are congruent. Incongruent *p*-values are marked as an inconsistency, and incongruent *p*-values that possibly change the statistical conclusion from significant to non-significant (and vice versa) are marked as a gross inconsistency.

The program statcheck contains an automated one-tailed test detection: if the words "one-tailed", "one-sided", or "directional" are mentioned somewhere in the article and a *p*-value would have been consistent if it was one-sided, it is considered consistent. Furthermore, statcheck takes rounding of the reported test statistic into account. Take for instance the result $t(48) = 1.43$, $p = .158$. Recalculation would give a *p*-value of .159, which seems incongruent with the reported *p*-value. However, the true *t*-value could lie in interval (1.425, 1.435), with *p*-values ranging from .158 to .161, statcheck will count any *p*-value within this range as consistent. We assumed that all studies retained an overall alpha of .05. We also counted results reported as $p = .05$ as significant, since previous research showed that over 90% of the instances in which $p = .05$ was reported, the authors interpreted the result as significant (Nuijten et al., 2016). Finally, note that when erroneously only one of the three components of an NHST result (test statistic, degrees of freedom, or *p*-value) is adjusted to correct for multiple testing, post-hoc testing, or violations of assumptions, the result becomes internally inconsistent and statcheck will flag it as such. However, in an extended validity study of statcheck, we found that such statistical corrections do not seem to cause the high estimates of the general prevalence of inconsistencies (for details, see Nuijten, Van Assen, Hartgerink, Epskamp, & Wicherts, 2017). For a more detailed explanation of statcheck, see Nuijten et al. (2016), or the statcheck manual at http://rpubs.com/michelenuijten/statcheckmanual.

In Nuijten et al. (2016) we investigated the validity of statcheck and found that the interrater reliability between manual coding and statcheck was .76 for inconsistencies and .89 for gross inconsistencies. In an additional validity study, we found that statcheck's sensitivity (true positive rate) and specificity (true negative rate) were high: between 85.3% and 100%, and between 96.0% and 100%, respectively, depending on the assumptions and settings. The overall accuracy of statcheck ranged from 96.2% to 99.9%. For details, see Appendix A in Nuijten et al. (2016) and the additional validity study, see (Nuijten et al., 2017).

Using statcheck, we extracted 6,482 statistical results from 498 of the 764 articles (65.2%) that contained APA reported NHST results. Note that the conversion of articles to plain text files can be different for PDF and HTML files, which can cause statcheck to recognize or miss different statistical results. Since all articles for JBDM were PDF files, and all articles in JDM HTML files, we could not reliably compare overall inconsistency rates between the journals. However, since over time the file types for each journal stayed the same, we could compare change in inconsistencies over time between the journals. All tests in this study are two-tailed unless otherwise specified and we maintained an alpha level of .05.

Results

General Descriptives. In total, we extracted 6,482 NHST results, which is on average 13.0 NHST results per article. On average, the articles in JBDM contained more NHST results than JDM articles (15.4 and 10.9 results, respectively). We found that on average 9.3% of the reported NHST results within an article was inconsistent and 1.1% grossly inconsistent. These inconsistency rates are similar to what we found in previous research (9.7% and 1.4%, respectively; Nuijten et al., 2016).

Note that the general prevalence of inconsistencies can be estimated in several ways. A first way is to look at the complete set of NHST results, and calculate which percentage of these are inconsistent or grossly inconsistent. The downside of this method is that it does not take into account that results within one article may be statistically dependent. A second method is to calculate for each article which proportion of reported NHST results are inconsistent, and average this over all articles. The downside of this method is that articles with fewer results get as much weight in the calculations as articles with more results, whereas they contain less (precise) information. The third method is to use multilevel logistic models that estimate the probability that a single NHST result is inconsistent while including a random effect at the article level. The downsides of this method are that the assumption of normally distributed random effects may be violated and that the conversion of logits to probabilities in the tails of the distribution leads to inaccurate probability estimates. Taking into account the pros and cons of all these methods, we decided to focus on the second method: the average of the average percentage of inconsistencies within an article, which we call the "inconsistency rate". We retained this method throughout the paper to estimate the general prevalence of inconsistencies. To test relations between inconsistencies and open data or open data policies, we used multilevel models.

Confirmatory analyses. Our first hypothesis was that JDM would show a larger decrease in (gross) inconsistencies than JBDM after the introduction of the data sharing policy in JDM. However, the mean prevalence of (gross) inconsistencies actually shows a pattern opposite to what we expected: the inconsistency rate increased in JDM

after its open data policy from 9.7% to 11.0%, and the inconsistency rate decreased in JBDM from 9.1% to 7.0% (see **Table 2**). For illustration purposes, we also plotted the inconsistency rates in both journals over time in **Figure 1**. The Figure shows a drop in the inconsistency rate in JDM in 2013 onwards (two years after introduction of the data policy). However, there are only few inconsistencies in absolute sense in 2013 and 2014, which makes it hard to interpret this drop substantively; this decrease is in line with only random fluctuations from year to year. More details about the general trends in (gross) inconsistencies over time can be found in the Supplemental Materials at https://osf.io/5j6tc/.

We tested the interaction between journal and the period in which a paper was published with a multilevel logistic regression analysis in which we predicted the probability that a *p*-value was (grossly) inconsistent with Time (0 = before data sharing policy, 1 = after data sharing policy), Journal (0 = JBDM, 1 = JDM), and the interaction Time * Journal:

$$Logit\left[(gross)\ \ inconsistency\right] = b_0 + b_1 Time_i + b_2 Journal_i + b_3 Time_i * Journal_i + \ \theta_i, \quad (1)$$

Where subscript *i* indicates article, Time is the period in which an article is published (0 = published before JDM's data sharing policy, 1 = published after JDM's data sharing policy), Journal is the journal in which the article is published (0 = JBDM and JDM = 1), and θ_i is a random effect on the intercept b_0. We included a random intercept because the statistical results are nested within article, which means there can be dependency in the inconsistencies within the same article.

The interaction effect was not significant (b = 0.37, 95% CI = [–0.292; 1.033], $Z = 1.10$, $p = .273$), which means that there is no evidence that changes in inconsistencies over

Table 2: Number of (gross) inconsistencies per journal (JDM = Judgment and Decision Making and JBDM = Journal of Behavioral Decision Making) and time period (T1 = published in 2006–Feb 2011 and T2 = published in April 2011–2014). In April 2011 JDM started encouraging open data.

		# articles	# articles with APA reported NHST results	# articles with APA reported NHST results and open data	# APA reported NHST results	average # APA reported NHST results per article	average % inconsistencies per article	average % gross inconsistencies per article
JBDM	**T1**	157	117 (74.5%)	0	1,543	13.2	9.1%	1.4%
	T2	149	118 (79.2%)	2	2,074	17.6	7.0%	0.6%
JDM	**T1**	236	128 (54.2%)	11	1,313	10.3	9.7%	1.1%
	T2	222	135 (60.8%)	118	1,552	11.5	11.0%	1.1%
Total		764	498 (65.2%)	131	6,482	13.0	9.3%	1.1%

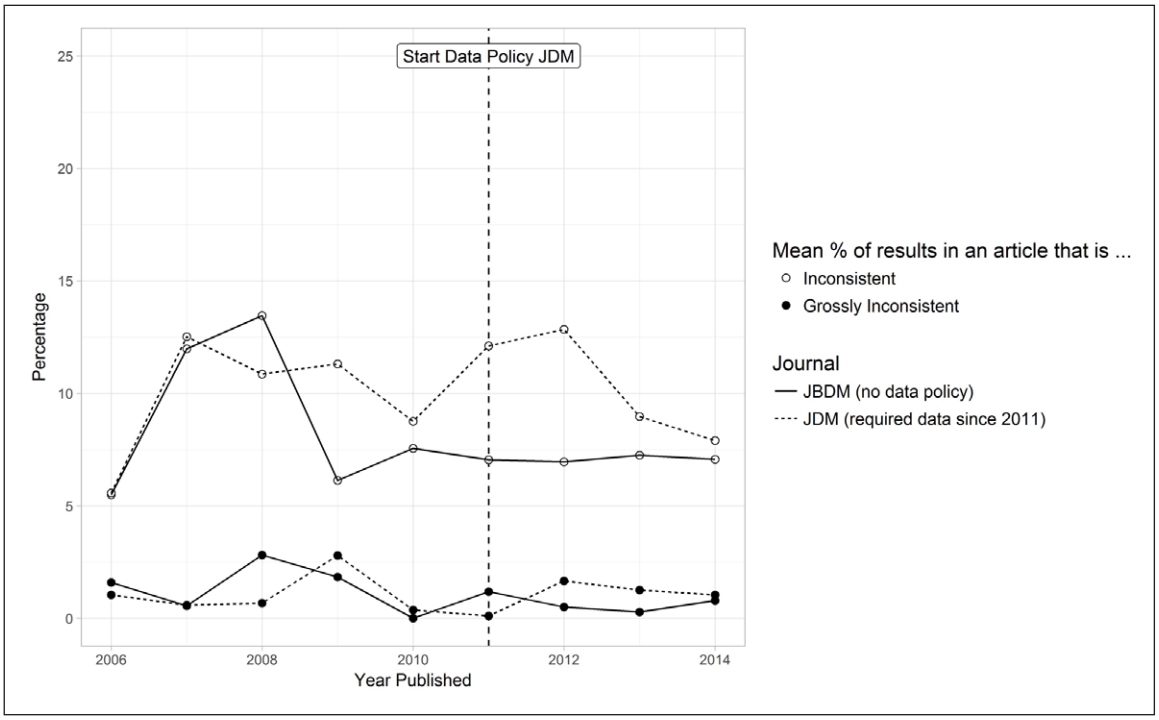

Figure 1: Per publication year and journal the average percentage of results within an article that was inconsistent or grossly inconsistent (the "inconsistency rate").

time differed for the journals. Second, we looked at the change in the prevalence of gross inconsistencies, but these showed patterns opposite to those expected as well. The gross inconsistency rate stayed at 1.1% in JDM after its open data policy, whereas the gross inconsistency rate in JBDM decreased from 1.4% to 0.6%. To test this finding, we performed the same multilevel logistic regression with Time, Journal, and Time * Journal as predictors, but this time we predicted the probability that a *p*-value was a *gross* inconsistency. Again, we included a random effect for article. In this analysis, too, we found that the interaction effect was not significant (b = 0.58, 95% CI = [−1.412; 2.580], $Z = 0.57$, $p = .566$), meaning that there is no evidence that any change in gross inconsistencies over time depends on journal.

Our second hypothesis was that articles that are accompanied with open data contain fewer (gross) inconsistencies than articles that are not accompanied with open data. Again, we observed the opposite pattern in the prevalence of inconsistencies: on average, in articles without open data 8.8% of the results was inconsistent as opposed to 10.7% in articles with open data (see **Table 3**). To test this pattern, we again fitted a multilevel logistic regression model in which we predicted the probability that a *p*-value was an inconsistency with Open Data (0 = the *p*-value is from an article without open data, 1 = the *p*-value is from an article with open data), and a random effect for article. Open Data did not significantly predict whether a *p*-value was inconsistent (b = 0.30, 95% CI = [–0.069; 0.672], $Z = 1.59$, $p = .111$). Next, we looked at the relation between gross inconsistencies and open data. We found a pattern in the predicted direction: articles with open data had on average a lower rate of gross inconsistencies than articles without open data (1.0% of the results versus 1.1%, respectively). To test this relation, we fitted a multilevel logistic regression model to see if Open Data predicts the probability that a *p*-value is a gross inconsistency, including a random effect for article. Again, Open Data was not a significant predictor (b = 0.001, 95% CI = [−1.150; 1.153], $Z = 0.002$, $p = .998$). A problem with this analysis is that the large majority of papers with open data were published in JDM, which makes this analysis a comparison of the inconsistency rates in both journals. Since we only have HTML files from JDM and only PDF files from JBDM, this comparison could therefore reflect differences in the performance of statcheck instead of an actual difference in inconsistency prevalence.

Based on our analyses we found no evidence for our two hypotheses: JDM did not show a larger decrease in inconsistencies and gross inconsistencies than JBDM after the introduction of the data sharing policy in JDM. We also did not find that articles that are accompanied by open data contained fewer inconsistencies or gross inconsistencies than articles without open data, but this analysis is possibly confounded.

Table 3: Number of (gross) inconsistencies in articles with and without open data.

Open data?	# articles with APA reported NHST results	average % inconsistencies per article	average % gross inconsistencies per article
No	367	8.8%	1.1%
Yes	131	10.7%	1.0%

Conclusion

In this study, we investigated whether there is a relationship between recommended data sharing and statistical reporting inconsistencies, by comparing the number of inconsistencies over time in the journal JDM, which introduced a data sharing policy, and JBDM, that has no such policy. We hypothesized that JDM would show a stronger decrease in (gross) inconsistencies than JBDM (open policy effect), and that *p*-values from articles accompanied by open data were less likely to be inconsistent (open data effect). We found no evidence of an open policy effect or an open data effect.

It is worth noting that even though we found no relation between data sharing policy and reporting inconsistencies, the data sharing policy of JDM did result in the retraction of an article after anomalies in the (open) data were discovered.[4] This emphasizes the potential importance of open data (Simonsohn, 2013; Wicherts, 2011).

The main limitation of this study is its lack of power. Even though we downloaded a considerable number of articles for each cell in the design, statcheck did not retrieve statistics from every paper, and of the retrieved statistics only a small percentage was inconsistent, resulting in potentially underpowered regression analyses. Based on these data alone we cannot draw firm conclusions about the relation between data sharing and reporting inconsistencies. We therefore designed Studies 2 and 3 to obtain more power and thus more reliable results.

Study 2

In Study 2 we compared the prevalence of inconsistencies and gross inconsistencies in psychological articles the open access journal Frontiers in Psychology (FP) and in journals from the major open access publisher PLOS. From March 1st 2014 onwards PLOS required submissions to be accompanied with open data. Their online policy on data availability states that "The data underlying the findings of research published in PLOS journals must be made publicly available. Rare exceptions may apply and must be agreed to with the Editor." (https://www.plos.org/editorial-publishing-policies; retrieved October 2017). Furthermore, all submissions had to have an official Data Availability Statement explaining how the data were shared or why the data could not be shared. (Bloom et al., 2014). Not sharing data could affect the publication decision. The author guidelines of FP also state that data must be made available, but the guidelines are not as explicit as those of PLOS: FP does not require a standardized data availability statement, and it is not clear if not sharing data could affect the publication decision.[5] We again hypothesized that the inconsistencies and gross inconsistencies in articles from PLOS would show a stronger decrease (or less strong increase) over time than in FP. Furthermore, we again hypothesized that data sharing (regardless of whether it was required) is associated with fewer inconsistencies and gross inconsistencies in an article.

Method

Preregistration. The hypotheses as well as the design and analysis plan were preregistered and can be found at https://osf.io/a973d/. The hypotheses, procedure, and power analysis were registered in detail, whereas the analysis plan was registered more generally, and consisted of the regression equations we intended to test. We followed our preregistered plan, except for one detail: we did not preregister any data exclusion rules, but we did exclude one article from the analysis because it was unclear when it was received.

Sample. We downloaded all articles available in HTML from FP and all HTML articles with the topic "Psychology" from PLOS in two time periods to capture change in inconsistencies before and after the introduction of PLOS' requirement to submit raw data along with an article.

Articles from FP. We already had access to all FP articles published from 2010 to 2013 that were downloaded for the research in Nuijten et al. (2016). On top of that, in the period of 9 to 15 June 2015 we manually downloaded all FP articles published from January 1st 2014 up until April 30th 2015. In total we had 4,210 articles published from March 8th 2010 to April 30th 2015.

For our sample, we selected only the research articles, clinical (case) studies, and methods articles (excluding editorials, retractions, opinions, etc.). We used systematic text searches in R to automatically select these articles, which resulted in 2,693 articles. Next, we also used systematic text searches in R to extract whether the articles were received before or after PLOS' data sharing policy[6] that came into effect March 1st 2014.[7] 1,819 articles in the sample were received before the policy and 873 after the policy. One article was excluded because it was unclear when it was received. **Table 5** shows the number of downloaded articles per period and journal.

Articles from PLOS. PLOS has the option of selecting articles based on the date they were received, which made it straightforward to download articles and categorize them in received before or after PLOS' data sharing policy. Using the R package rplos (Chamberlain, Boettiger, & Ram, 2014) we first automatically downloaded all PLOS articles with the subject "Psychology" that were received before March 1st 2014, which rendered 7,719 articles. Next, we downloaded all "Psychology" articles received after March 1st 2014, rendering 1,883 articles. We restricted this sample to articles that were published in the same time span that the FP articles were published, which means that we excluded all PLOS articles published before March 8th 2010 (4 articles excluded) or after April 30th 2015 (376 articles excluded). Next, using systematic text searches in R we only selected the research articles from this sample,[8] rendering 7,700 articles from before the data sharing policy, and 1,515 articles from after the policy. The final sample size is described in **Table 4**.

Power analysis. Based on the number of downloaded articles and the previous results of Nuijten et al. (2016), we conducted a power analysis. We retained a baseline probability that a result in FP or PLOS was inconsistent of 6.4%.[9] We concluded that we had a power of .80 if the decrease in inconsistencies in PLOS over time is 2 to 3 percentage points steeper than in FP. The full details of the power analysis including all R code have been included in the preregistration and can be found at https://osf.io/ay6sh/.

Procedure. We used statcheck version 1.0.2 (Epskamp & Nuijten, 2015) to extract all APA reported NHST results from the PLOS and FP articles. Due to feasibility constraints, we decided not to check all the downloaded articles for open data, but only the ones that statcheck extracted results from (1,108 articles from FP and 2,909 articles from PLOS).[10]

For each downloaded article with detectable NHST results, we coded whether the (raw) data were available. Published data files in matrix format with subjects in the rows (so no correlation matrices) were considered open data. The data had to be published either in the paper, an appendix, or a website (with a reference to that website in the paper). Remarks such as "data are available upon requests" were not considered open data. Again, we did not assess whether any available data were relevant, usable, and/or complete.

Due to the large number of articles that needed to be coded with respect to data availability, we had seven coders: the six authors and a student assistant. We tested the coding protocol by assessing interrater reliability by coding 120 articles that were randomly selected from the full sample and calculating the intraclass correlation (ICC). In this set-up, every article was coded by two randomly selected coders. Per article three variables were coded. We coded whether the authors stated that the data was available (ICC(2,2) = .948), whether there actually was a data file available (ICC(2,2) = .861), and finally whether there was a URL linking to the data available (ICC(2,2) = .282). The last ICC was quite low. After further inspection of the coding it turned out that there was some confusion among coders whether a link to a data file that was also embedded in the article should be counted as a URL. Since this was not crucial for testing our hypotheses, we adapted the protocol to only code two variables: whether the authors state that the data were available, and whether

Table 4: Number of research articles downloaded from PLOS and FP before and after PLOS introduced obligatory data sharing. All articles were published between March 8th 2010 and April 30th 2015.

	Before PLOS' data sharing policy: Received before March 1st 2014	After PLOS' data sharing policy: Received after March 1st 2014	Total
FP	1,819 articles	873 articles	2,692 articles
PLOS	7,700 articles	1,515 articles	9,215 articles
Total	9,519 articles	2,388 articles	11,907 articles

the data actually were available. The final protocol is available on https://osf.io/yq4mt/.

The total sample was coded for open data with the help of an extra student assistant, resulting in eight coders in total. As a final reliability check, 399 articles (approximately 10% of all articles with APA reported NHST results) were coded twice by randomly assigned coders. The interrater reliability was high: for the data availability statement the ICC(2,2) was .900[11], and for whether the data was actually available the ICC(2,2) was .913.[12] Furthermore, the first author blindly recoded all cases in which a coder had added a remark, and solved any discrepancies by discussion. The first author also solved any discrepancies between coders when an article was coded twice. All coders were blind for the statcheck results, but not for the journal and time period in which the article was published.

Results

General Descriptives. Table 5 shows the descriptive results per journal and time. It turned out that statcheck extracted NHST results from more articles than expected based on the data of Nuijten et al. (2016). On average, 41.2% of the articles in FP and 31.6% of the articles in PLOS contained APA reported NHST results that statcheck could detect. This means that we obtained more power than expected based on our power analysis. Across journal and time, on average 13.0% of NHST results in an article was inconsistent, and 1.6% was grossly inconsistent. The average percentage of inconsistencies within an article in FP increased over time from 13.1% to 16.2%, whereas the inconsistency rate in PLOS increased from 12.5% to 13.5%. The percentage of gross inconsistencies in FP increased slightly from 1.7% to 2.0%, and the gross inconsistencies in PLOS increased from 1.4% to 1.7%. The steeper increase in inconsistencies in FP as compared to PLOS seems to be in line with our hypothesis that an open data policy influences the inconsistency rates, but we will test this in the next section. For the sake of completeness, we also added a plot that shows the trends over time in the inconsistency rates per journal (see **Figure 2**). Note that this plot shows the average inconsistency rates in the year the articles were published, not the years in which the articles were received. That means that even though some articles were published after PLOS introduced the data policy, they may have been submitted before the policy was implemented. Even so, the figure gives a good indication of the prevalence of (gross) inconsistencies in PLOS and FP over time. More details about the general trends in (gross) inconsistencies over time can be found in the Supplemental Materials at https://osf.io/5j6tc/.

Confirmatory analyses. For our first set of preregistered hypotheses we hypothesized that the probability that a result is inconsistent decreases more strongly in PLOS after they introduced a data sharing policy than in FP, where there was no data sharing policy (open policy effect). More specifically, we expected that there is a negative interaction effect of Time (0 = received before PLOS' data sharing policy, 1 = received after PLOS' data sharing policy) times Journal[13] (0 = FP, 1 = PLOS) on the probability that a result is inconsistent or grossly inconsistent. The raw probabilities of an inconsistency and gross inconsistency split up per time and journal can be found in **Table 4**. We tested our hypotheses by estimating the following multilevel logistic models:

$$Logit\left[(gross)\ \ inconsistency\right] = b_0 + b_1 Time_i + b_2 Journal_i + b_3 Time_i * Journal_i + \ \theta_i, \quad (2)$$

Where subscript *i* indicates article, Time is the period in which an article is published (0 = received before PLOS' data sharing policy, 1 = received after PLOS' data sharing policy), Journal is the outlet in which the article is published (0 = FP and PLOS = 1), and θ_i is a random effect on the intercept b_0. We included a random intercept because the statistical results are nested within article, which means there can be dependency in the inconsistencies within the same article. We hypothesized that in both models the coefficient b_3 is negative. We maintained an α of .05. We did not preregister that we would use one-tailed tests, so we tested our hypotheses two-tailed.

When predicting the inconsistencies, we found a significant interaction effect of Time * Journal in the predicted direction, $b_3 = -0.43$, 95% CI = [–0.77; –0.085], $Z = -2.45$, $p = .014$. This indicates that the prevalence of inconsistencies decreased more steeply (or more accurately: increased *less* steeply) in PLOS than in FP. This finding is in line with the notion that requiring open data as a journal could decrease the prevalence of reporting errors.

When predicting gross inconsistencies, we did not find a significant interaction effect of Time * Journal; $b_3 = -0.12$, 95% CI = [–1.04; 0.80], $Z = -0.25$, $p = .804$. This means that there is no evidence that any change in

Table 5: Number of (gross) inconsistencies per journal (FP and PLOS) and time period (T1 = received before March 1st 2014 and T2 = received after March 1st 2014). PLOS required articles submitted after March 1st 2014 PLOS to be accompanied by open data.

		# articles	# articles with APA reported NHST results	# articles with APA reported NHST results and open data	# APA reported NHST results	average # APA reported NHST results per article	average % inconsistencies per article	average % gross inconsistencies per article
FP	T1	1,819	804 (44.2%)	11	11,079	13.8	13.1%	1.7%
	T2	873	304 (34.8%)	4	2,432	8.0	16.2%	2.0%
PLOS	T1	7,700	2,462 (32.0%)	110	33,064	13.4	12.5%	1.4%
	T2	1,515	447 (29.5%)	247	5,801	13.0	13.5%	1.7%
Total		11,907	4,017 (33.7%)	372	52,376	13.0	13.0%	1.6%

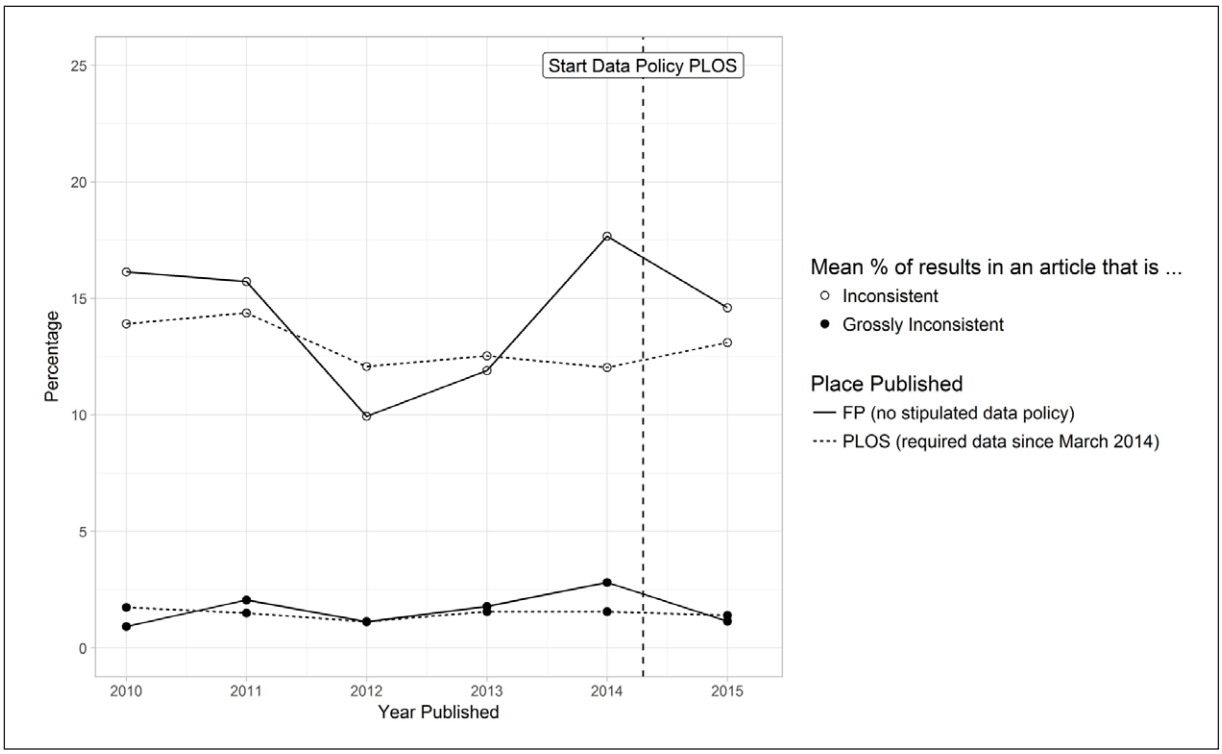

Figure 2: Per publication year and place published the average percentage of results within an article that was inconsistent or grossly inconsistent (the "inconsistency rate").

gross inconsistencies over time depended on the journal in which the result was published. This finding is not in line with our hypothesis. Since we found no significant interaction effect, we (exploratively) tested the model again without the interaction effect to see if there is a main effect for Time and/or Journal. We found no evidence for a main effect of Time (b_1 = 0.169, 95% CI = [–0.266; 0.605], Z = 0.762, p = .446) or a main effect of Journal (b_2 = –0.012, 95% CI = [–0.413; 0.438], Z = –0.063, p = .950). Note that our power analysis was based on the prevalence of inconsistencies, and not gross inconsistencies. The power of our analysis to find an effect of data sharing on the prevalence of gross inconsistencies is much lower since gross inconsistencies are much less prevalent.

For our second set of hypotheses we tested whether results in articles that are accompanied by open data have a lower probability of being inconsistent and grossly inconsistent than results in articles that are not accompanied by open data, regardless of the journal in which they were published (open data effect). We found that the average percentage of inconsistencies in an article was 13.7% when an article had open data, and 12.9% when an article did not have open data. The average percentage of gross inconsistencies in an article was 2.1% and 1.5% for articles with and without open data, respectively. These patterns are the opposite of what we expected. We tested whether there is a relationship between open data and the probability of a (gross) inconsistency by estimating the following two multilevel logistic models:

$$Logit\left[(gross)\ \ inconsistency\right] = b_0 + b_1 Open\ \ Data_i + \theta_i, \quad (3)$$

Where subscript i indicates article, Open Data indicates whether the data is published along with the article (0 = no open data, 1 = open data), and θ_i is a reffect on the intercept b_0. We hypothesized that in both models the coefficient b_1 is negative.

We found no effect of Open Data on the prevalence of inconsistencies (b_1 = 0.06, 95% CI = [–0.16; 0.27], Z = 0.50, p = .617) or the prevalence of gross inconsistencies (b_1 = 0.23, 95% CI = [–0.33; 0.79], Z = 0.79, p = .429). This finding is not in line with our hypothesis that articles accompanied by open data should have lower inconsistency rates.

Conclusion

In this study, we investigated the relation between required data sharing and statistical reporting inconsistencies using a larger dataset than in Study 1, by comparing the number of statistical reporting inconsistencies over time in open access articles. We compared psychology articles from journals in PLOS, which since March 2014 requires articles to be accompanied by open data, with articles in FP, which does encourage data sharing, but does not require it in the same strong terms as PLOS does. We hypothesized that PLOS would show a stronger decrease in (gross) inconsistencies than FP, and that *p*-values from articles accompanied by open data were less likely to be inconsistent. We found that the prevalence of inconsistencies over time increased less steeply in PLOS than in FP, which is in line with our hypotheses. However, we did not find evidence for our other hypotheses: there was no evidence that any change in gross inconsistency prevalence was different for PLOS and FP, and we also found no relationship between open data and *p*-value inconsistency.

Study 3

In Study 3, we examined the prevalence of reporting inconsistencies in the journal Psychological Science (PS). Before 2014, the policy of PS concerning data sharing

was simply the general policy of the APA, which roughly states that data should be available upon request. From 2014 onwards, however, PS has started to award so-called "Open Practice Badges" in recognition of open scientific practices (Eich, 2014). "Open Practice Badges" is a collective term for three types of badges: Authors can earn an Open Data Badge, an Open Materials Badge, and a Preregistration Badge. This simple intervention has proven to be very effective: the frequency of reported data sharing in PS increased almost ten-fold after introduction of the badges, compared to reported data sharing in PS before the badges, and data sharing in four comparable journals (Kidwell et al., 2016). Furthermore, articles in PS with an open data badge had a much higher probability of actually providing the data (93.8%) than articles without a badge that promised data (40.5%; Kidwell et al., 2016).

We again theorized that open practices in general and data sharing in particular would decrease inconsistencies and gross inconsistencies. To test this, we focused on articles published in PS from 2014 onwards, because in this time frame the Open Practice Badges enable a straightforward check for the availability of data and/or engagement in other open practices (sharing materials and preregistration). One of the main advantages of this study as compared to Study 1 and Study 2 in this paper, is that authors have to meet certain criteria before they are awarded any of the Open Practice Badges. For instance, for an Open Data Badge authors need to publish their data in an open-access repository that is time-stamped, immutable, and permanent.[14] Therefore, in this study we were better able to assess whether an article actually has (high quality) open data than in Study 1 or Study 2.[15] It is possible that articles published before 2014 also engaged in data sharing and other open practices, but due to feasibility constraints we did not attempt to code this. Furthermore, from July 2016 onwards, PS started using statcheck to screen articles for inconsistencies.[16] In our study, we only included PS articles published up until May 2016, because any drop in the prevalence of statistical reporting inconsistencies after May 2016 could have been caused by the use of statcheck in the review process instead of the introduction of the Open Practice Badges.

To investigate the relation between open practices in general and reporting inconsistencies, we tested the following two hypotheses (open practice effects), as stated in the preregistration at https://osf.io/6nujg/:

> "Statistical results in articles published in PS from 2014 onwards with one or more Open Practice Badges have a lower probability to be inconsistent (Hypothesis 1) and grossly inconsistent (Hypothesis 2) than statistical results in PS articles published from 2014 onwards without an Open Practice Badge."

These hypotheses concern an effect of open practices in general (including sharing materials and preregistration), but we were also interested in the effect of open data in particular on reporting inconsistencies. To that end we also focused on the Open Data Badges in specific, by testing the following two hypotheses (open data effects), as stated in the preregistration at https://osf.io/6nujg/:

> "Statistical results in articles published in PS from 2014 onwards with an Open Data Badge have a lower probability to be inconsistent (Hypothesis 3) and grossly inconsistent (Hypothesis 4) than statistical results in articles published from 2014 onwards without an Open Data Badge."

Finally, we theorized that PS' policy to award open practice with badges has caused the journal to become known as a journal focused on open, solid science. Because of this, we speculated that after the installation of the badge policy in 2014, the articles submitted to PS were of higher quality, regardless of whether they actually received a badge or not. Therefore, we also hypothesized that (open policy effects), as stated in the preregistration at https://osf.io/6nujg/:

> "Statistical results in articles published in PS before 2014 have a higher probability to be inconsistent (Hypothesis 5) and grossly inconsistent (Hypothesis 6) than statistical results in articles published in PS from 2014 onwards."

Method

Preregistration. The hypotheses and analysis plan (including the full R code) of this study were preregistered. The preregistration can be found at https://osf.io/8j56r/. All elements of the preregistration were written up in a high level of detail. We followed our preregistered plan, except for one aspect of the analysis. We preregistered the R code for the intended analyses, but did not take into account convergence problems. Our solutions to deal with these problems were ad hoc.

Sample. To investigate the prevalence of inconsistencies and gross inconsistencies in PS, we looked at HTML articles published in PS from 2003 to 2016. We already downloaded the articles published from 2003 to 2013 in previous research, which resulted in a sample of 2,307 articles (Nuijten et al., 2016). In June 2016, a research assistant downloaded all HTML articles except editorials published from January 2014 up until May 2016, which resulted in 574 articles (see **Table 6** for details).

Power Analysis. As we did in Study 2, we conducted power analyses for all hypotheses based on the number of downloaded articles and the results of Nuijten et al. (2016). We concluded that for hypothesis 1 and 3 we have 80% power if the probability of an inconsistency drops with about 50% after introduction of the badges (from .049[17] to .024), and if the probability of an inconsistency drops with about 25% for hypothesis 5 (from .049 to .036; see the preregistration for details). Furthermore, we concluded that we probably do not have sufficient power to detect predictors of a reasonable size of gross inconsistencies (hypotheses 2, 4, and 6). Consequently, we do not trust the test results on gross inconsistencies.

However, we still reported the results of the multilevel logistic regression analyses of gross inconsistencies for the sake of completeness. The full details of this power analysis including all R code has been included in the preregistration and can be found at https://osf.io/xnw6u/.

Procedure. For the articles published from 2014 onwards a research assistant coded which (if any) badges accompanied the article. A detailed protocol (in Dutch) with instructions for the research assistant on which articles to download and how to code the open practice badges is available on OSF: https://osf.io/kktk5/. For full sample details, see **Table 6**.

We used statcheck version 1.2.2 (Epskamp & Nuijten, 2016) to extract all APA reported NHST results from the downloaded PS articles and check them on internal consistency.

Results

General Descriptives. Of the 2,879 downloaded articles, 2,106 (73.2%) contained APA reported NHST results. In total, we extracted 20,926 NHST results, which is on average 9.9 NHST results per article. Per article we found that on average 9.3% of the reported NHST results was inconsistent and 1.1% grossly inconsistent. These inconsistency rates are similar to what we found in Study 1 and 2, and in previous research (Nuijten et al., 2016).

Hypotheses 1 & 2: Open Practice Badges. Hypothesis 1 and 2 focused on whether the probability that a result is a (gross) inconsistency was lower if the article had one or more Open Practice Badges. We found that 574 articles were published in the period from 2014 onwards when PS started to award badges. In our sample, the probability that a result was inconsistent is slightly higher for articles with a badge (11.8%) than articles without a badge (9.7%), but the probability that a result was a gross inconsistency is equal in the two groups (see **Table 7** for details).

We tested hypothesis 1 and 2 with the following logistic multilevel models:

$$Logit\left[(gross)inconsistency\right] = b_0 + b_1 OpenPracticeBadge_i + \theta_i, \quad (4)$$

Where subscript i indicates article, OpenPracticeBadge indicates whether an article had one or more of the three available Open Practice Badges (1) or not (0), and is a random effect on the intercept b_0. We hypothesized that in both models the coefficient b_1 is negative. We tested these hypotheses maintaining an α of .05, and we tested one-sided ($b_1 < 0$).

Consistent with our preregistered analysis plan, we took into account the possibility that the year in which the paper was published could cause a spurious relation between having a badge and the prevalence of (gross) inconsistencies: it is imaginable that a gradual change in research culture caused both the prevalence of open practice badges to increase and the prevalence of (gross) inconsistencies to decrease (although **Figure 3** does not seem to show such a trend in inconsistencies, see the next sections for more details). We therefore first intended to test whether there was an interaction effect between OpenPracticeBadge and Year on the prevalence of (gross) inconsistencies. Due to convergence problems, we re-estimated this model by altering the number of nodes in the Gauss-Hermite quadrature formula to 0 and 0.9. The results of these analyses revealed no effect of the year in which an article was published. Therefore, we proceeded with fitting the originally hypothesized models. Based on our analyses, we found no evidence for an effect of OpenPracticeBadge on the probability that a result is inconsistent (b_1 = –0.349, 95% CI = [–0.867; 0.169], Z = –1.320, p = .093, one-tailed) or grossly inconsistent (b_1 = –0.894, 95% CI = [–3.499; 1.711], Z = –0.673, p = .250, one-tailed).

Hypotheses 3 & 4: Open Data Badges. In Hypotheses 3 and 4, we looked at the relation between whether articles had an Open Data Badge or not and the probability that a result in that article was inconsistent. Of the 574 articles published in PS from 2014 onwards, 97 had an Open Data Badge and 477 did not. The average percentage of both inconsistencies and gross inconsistencies per article in this sample was higher in articles with an Open Data Badge than in articles without one (see **Table 8** for details).

Table 6: Total number of downloaded research articles published before and after PS introduced the Open Practice Badges, and how many of these articles were accompanied by the different badges.

Year published	Total # articles downloaded	Open Data Badge	Open Material Badge	Preregistration Badge
2003–2013	2,305[18]	0	0	0
2014–2016	574[19]	97	69	4

Table 7: Number of (gross) inconsistencies for articles published in PS after 2014 with at least one Open Practice Badge and without any badges.

	# articles downloaded	# articles with APA reported NHST results (%)	# APA reported NHST results	average # APA reported NHST results per article	average % inconsistencies per article	average % gross inconsistencies per article
No Badges	469	351 (74.8%)	4240	9.7	9.7%	1.5%
Open Practice Badge(s)	105	75 (71.4%)	1039	10.3	11.8%	1.5%
Total	574	426 (74.2%)	5279	9.8	10.0%	1.5%

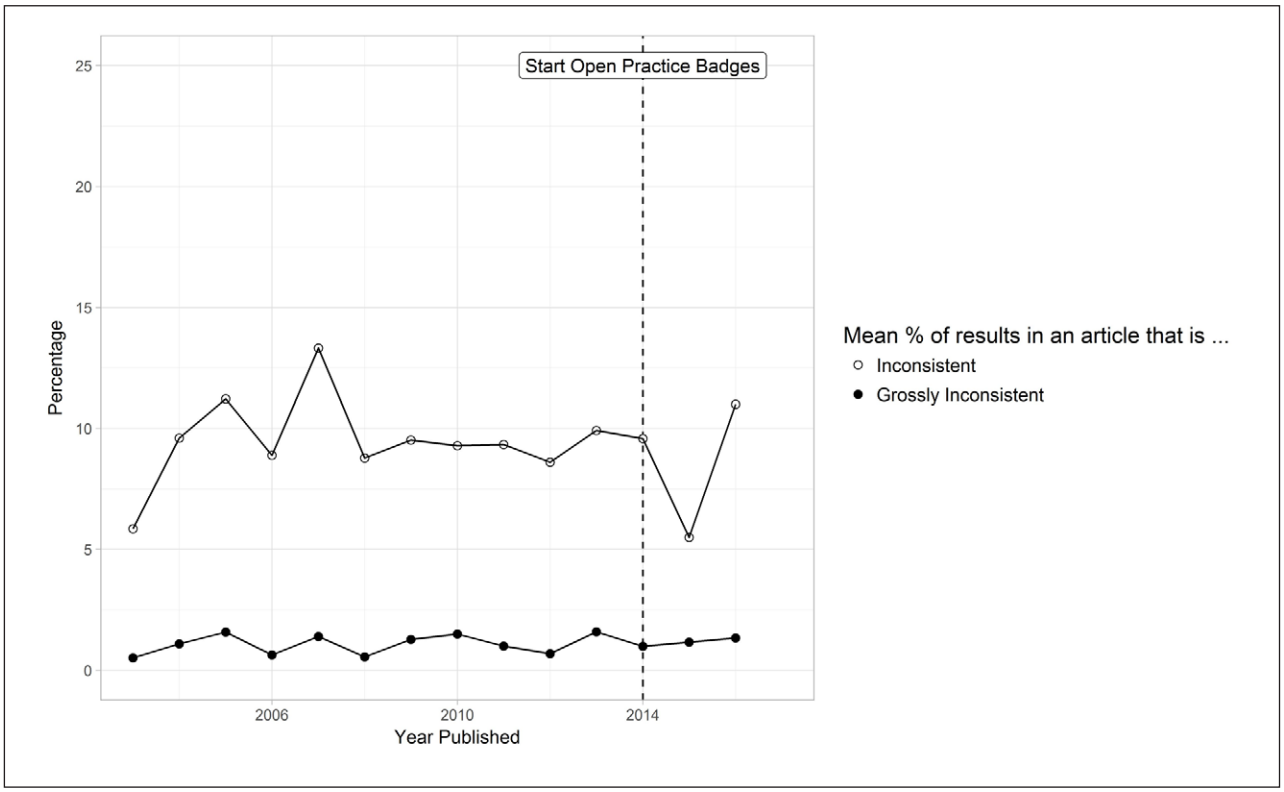

Figure 3: The average percentage of results within an article that was inconsistent or grossly inconsistent (the "inconsistency rate") per publication year.

Table 8: Number of (gross) inconsistencies for articles published in PS after 2014 with and without an Open Data Badge.

	# articles downloaded	# articles with APA reported NHST results (%)	# APA reported NHST results	average # APA reported NHST results per article	average % inconsistencies per article	average % gross inconsistencies per article
No Open Data Badges	477	354 (74.2%)	4,259	9.8	9.6%	1.5%
Open Data Badge	97	72 (74.2%)	1,020	9.8	12.0%	1.6%
Total	574	426 (74.2%)	5,279	9.8	10.0%	1.5%

We estimated the following logistic multilevel models to test Hypothesis 3 and 4:

$$Logit\left[(gross)\,inconsistency\right] = b_0 + b_1 OpenDataBadge_i + \theta_i, \quad (5)$$

Where OpenDataBadge indicates whether an article had an Open Data Badges (1) or not (0). We hypothesized that in both models the coefficient b_1 is negative. We tested these hypotheses maintaining an α of .05, and we tested one-sided ($b_1 < 0$).

Similar to Hypotheses 1 and 2 and following the preregistration, we first tested the models including two extra control variables: in which year the article was published and whether the article had a badge other than an Open Data Badge. We included the latter control because we wanted to distinguish between effects of open practice in general and open data in particular. We first intended to test a three-way interaction between Open Data Badge, other badges, and year published, because if there would be a three-way interaction, any two-way interactions or main effects could not be interpreted. However, these models were too complex to fit and did not converge. We therefore continued to fit the models with three two-way interactions. Similar to hypotheses 1 and 2, we fit the models with the node-parameter set to 0 and 0.9. Based on these analyses, we continued to estimate the simple effects of the following model:

$$Logit\left[inconsistency\right] = b_0 + b_1 OpenDataBadge_i + b_2 OtherBadge_i + b_3 Year_i + b_4 OpenDataBadge_i * Year_i + \theta_i, \quad (6)$$

Where we looked at the coefficients of the model when Year was centered on 2014, 2015, and 2016. The results show that the negative relation between whether an article had an Open Data Badge and the probability that a result was inconsistent was stronger for articles published in 2014 than in 2015 or 2016 (see **Table 9** for details). This finding would be in line with a scenario in which open data (badges) led to a lower prevalence of reporting

Table 9: Results of the simple effects analysis to predict the probability that a result is inconsistent when Year is centered on 2014, 2015, and 2016. The Table shows the regression coefficients and their standard errors. The main predictor of interest, Open Data Badge, is printed in bold.

	b (SE)		
Year centered on	**2014**	**2015**	**2016**
Intercept	−2.96 (.15)***	−3.18 (.15)***	−3.40 (.27)***
Open Data Badge	**−1.60 (.78)***	**−0.65 (.48)**	**0.29 (.56)**
Other Badge	0.22 (.50)	0.22 (.50)	0.22 (.50)
Year	−0.22 (.16)	−0.22 (.16)	−.22 (.16)
Year * Open Data Badge	0.94 (.48)*	0.94 (.48)*	.94 (.48)*

* $p < .05$; ** $p < .01$; *** $p < .001$.

Table 10: Number of (gross) inconsistencies for articles published in PS before 2014 (Period 1) and from 2014 onwards (Period 2). From 2014 onwards PS started to award Open Practice Badges.

	# articles downloaded	# articles with APA reported NHST results (%)	# APA reported NHST results	average # APA reported NHST results per article	average % inconsistencies per article	average % gross inconsistencies per article
Period 1	2,305	1,680 (72.9%)	15,647	10.0	9.1%	1.1%
Period 2	574	426 (74.2%)	5,279	9.8	10.0%	1.5%
Total	2,879	2,106 (73.2%)	20,926	9.9	9.3%	1.1%

inconsistencies in 2014, but that this effect decreased over time.

Then, to predict the probability that a result was grossly inconsistent, we fitted a model including the two-way interactions to compare it with a model with only the main effects. However, the model with the two-way interactions was too complex to fit, and failed to converge. We therefore continued with the model with only main effects, which we again fitted with the node-parameter set to 0 and 0.9. We compared these models with a model with only Open Data Badge as a predictor and found that adding control variables did not significantly improve the model ($\chi^2(2) = .531$, $p = .767$). Based on the final model including only Open Data Badge as a predictor, we found that there was no significant relation between the probability that a result was grossly inconsistent and whether the article had an Open Data Badge or not (b = −.869, 95% CI = [−3.481; 1.743], $Z = -0.652$, $p = .257$, one-tailed).

Hypotheses 5 & 6: Time Period. For Hypothesis 5 and 6 we were interested if there was a change in the probability that a result was (grossly) inconsistent when PS started to award badges, so we looked at articles published in PS before and after 2014 when the badge system was introduced. In our sample, we had 2,305 downloaded articles from before 2014, and 574 articles from 2014 onwards. The prevalence of inconsistencies and gross inconsistencies was slightly higher in the second period (see **Table 10** for details). We were interested in the difference in inconsistency rates before and after the introduction of the badges, but to sketch a more complete picture we also plotted the inconsistency rates per year (see **Figure 3**). This figure shows that there is a steep drop in the inconsistency rate in articles that were published after the Open Practice Badges were introduced, but this is not a consistent trend. More details about the general trends in (gross) inconsistencies over time can be found in the Supplemental Materials at https://osf.io/5j6tc/.

We tested our hypotheses using the following multilevel logistic models:

$$Logit\left[(gross)\,inconsistency\right] = b_0 + b_1 Period_i + \theta_i, \quad (7)$$

Where Period indicates the time period in which the article was published (0 = T1, published before 2014 and the badge policy; 1 = T2, published from 2014 onwards when the badge policy was installed). Again we included a random intercept to account for dependencies of results within articles. We hypothesized that in both models the coefficient b_1 is negative. We tested this hypothesis maintaining an α of .05 using a one-sided ($b_1 < 0$) test.

Following the strategy from the previous hypotheses, we first intended to test the models controlling for possible effects of whether an article had any of the badges, and the specific year in which the article was published. Again, we first intended fit the models including a three-way interaction between Period, Badges, and Year, and in case there was no significant three-way interaction continue with a model with all two-way interactions, as we preregistered. However, we later realized that testing an interaction between Period and Badges does not make sense because badges were always awarded in T2. Similarly, any interaction between Year and Period also does not make sense, because all years up to 2014 were per definition T1 and from 2014 onwards T2.[20] We therefore ran models with a main effect for Period and only one two-way interaction between Badges and Year. Including this two-way interaction did not improve the models, so we continued to fit the models including all main effects and compared them to the models with only Period as predictor. The models that included all main effects did not significantly improve in fit as compared to the models with only Period

as predictor when predicting inconsistencies ($\chi^2(2)$ = 2.244, p = .326) or gross inconsistencies ($\chi^2(2)$ = 0.263, p = .877). We therefore proceeded with fitting the originally hypothesized models.

In line with our hypothesis, we found evidence that a result has a lower probability of being inconsistent when it was published from 2014 onwards (b_1 = –0.204, 95% CI = [–0.424; 0.015], Z = –1.823, p = .034, one-tailed). Note that this conclusion differs from the descriptives in **Table 10** that show that the average percentage of inconsistencies actually increased from Periods 1 to 2 (from 9.1% to 10.0%). These differences in results arise because these analyses reflect different ways to estimate the prevalence of inconsistencies, each with its own advantages and disadvantages (see the section General Descriptives in Study 1 for details). However, despite these seemingly discrepant results for both methods, the effect of open data policy was invariably very small at best. When we looked at gross inconsistencies, we found no evidence for an effect of Period on the probability that a result is grossly inconsistent (b_1 = –0.186, 95% CI = [–1.140; 0.768], Z = –0.382, p = .351, one-tailed).

The full details on the analyses of hypotheses 1 through 6 and the ad-hoc solutions to the convergence problems can be found in the Supplemental Information at https://osf.io/4gx53/ and in the R code at https://osf.io/8e3gr/.

Conclusion

In Study 3, we documented the prevalence of reporting inconsistencies in the journal *Psychological Science*. We hypothesized that articles with any of the Open Practice Badges had a lower prevalence of inconsistencies and gross inconsistencies than articles without any badges, but we found no evidence to support this. Furthermore, we hypothesized that articles with an Open Data Badge in particular had a lower prevalence of inconsistencies and gross inconsistencies than articles without an Open Data Badge. We found that for articles published in 2014 there was a lower probability that a result was inconsistent if an article had an open data badge, but this pattern did not hold for other years or for gross inconsistencies. Finally, we hypothesized that the prevalence of inconsistencies and gross inconsistencies was lower from 2014 onwards, when PS installed the badge policy. We found evidence that the prevalence of inconsistencies was indeed lower from 2014 onwards than before 2014, but this only held when we looked at the multilevel logistic models and were not in line with the descriptives in **Table 10**. Furthermore, we did not find a similar pattern for gross inconsistencies. Our results indicate that if there is any effect of the introduction of the policy on reporting inconsistencies, it is very small at best.

Exploratory Findings across Studies 1, 2, and 3

We distinguish between confirmatory and exploratory analyses. Confirmatory analyses are intended to test a priori formulated hypotheses, as opposed to exploratory analyses, which are more data-driven. Although confirmatory findings are more reliable than exploratory findings, exploratory findings can be important in formulating new hypotheses. As long as the distinction is made clear, both confirmatory and exploratory findings have their own merits (see also Wagenmakers, Wetzels, Borsboom, Maas, & Kievit, 2012).

The results of Studies 2 and 3 in the sections above can be considered purely confirmatory, since we preregistered the hypotheses, procedure, and analysis plans. This also means that the results of Study 1 cannot be considered purely confirmatory, because this study was not preregistered. Beside confirmatory analyses, we also performed several additional, more explorative analyses. We looked at cases in which data were promised but not delivered, the effectiveness of journal policies on data sharing, and whether articles with different types of gross inconsistencies also differ in how often they have open data. Finally, we also looked at the prevalence of inconsistencies over time, but since we did not find clear trends (similar to the findings of Nuijten et al., 2016), we only included these results in the Supplemental Information at https://osf.io/5j6tc/. We did not test any of the exploratory findings for statistical significance, because *p*-values are only interpretable in confirmatory tests (see Wagenmakers et al., 2012).

Data missing when promised

A large part of this study focuses on the availability of research data. Ideally, open data should follow the FAIR Guiding Principles (Wilkinson et al., 2016), which state that data should be Findable, Accessible, Interoperable, and Reusable. Here, we only focused on the first and least stringent of these principles: findability. However, in Study 2 (PLOS vs. FP) we noticed that in many cases articles stated that all data were available, whereas in fact this was not the case. We analyzed these cases in detail below.

We recorded 134 cases in articles from PLOS journals where data were promised but not available. This is as much as 29.0% of all PLOS articles that promised data. This is in line with the findings of Chambers (2017, p. 86), who found that 25% of a random sample of 50 PLOS papers employing brain-imaging methods stated their data was available, whereas in fact it was not. In FP, we found a similar percentage: of the twelve articles that promised data, three articles (25.0%) did not have available data. In **Table 11** we categorized all articles from Study 2 on whether data were promised and whether data were actually available, split up by journal.

We examined papers that promised but did not deliver data according to the type of "missing" data. In a minority of the cases (N = 11), the data were hard or impossible to find due to broken URLs, links to Chinese websites, or directions to general data websites (e.g., http://osf.io). The large majority of cases (N = 126, all in PLOS) were articles that only reported summary data, such as tables with means and standard deviations or bar plots, instead of actual raw data files. All but two of these cases were published after PLOS started requiring open data and every published article contained an explicit data availability statement. These data availability statements roughly fell in two categories: "Data Availability: The authors confirm that all data

underlying the findings are fully available without restriction. All data are included within the manuscript" (N = 9) and "Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All *relevant* data are within the paper" (italics added; N = 115).

Based on our findings, we speculate that there are two likely causes for the high rate of "missing" promised data in PLOS. Firstly, it is possible that the definition of "data" is unclear to the authors, PLOS editorial staff, or both. Perhaps summary data are considered enough information to comply with PLOS' open data regulations. Secondly, a lot of flexibility is introduced by allowing the data statement to promise all "relevant" data to be available. The word "relevant" is open to interpretation and might lead to underreporting of actual raw data files. We note that this high rate of missing promised open data is by no means unique for PLOS. A recent study found that as much as 40.5% of articles published in the journals *Clinical Psychological Science, Developmental Psychology, Journal of Experimental Psychology: Learning, Memory, and Cognition, and Journal of Personality and Social Psychology* that promised open data did not deliver (Kidwell et al., 2016). Whatever the cause may be, we are concerned about the high percentage of papers with missing open data.

Effectiveness open data policy

We noted that journal policy on sharing data seems highly effective. **Figure 4** shows that the percentage of articles with open data increased dramatically right after JDM, PLOS, and PS introduced a data sharing policy (in 2011, 2014, and 2014, respectively), whereas JBDM and FP without a data policy did not show such an increase. Specifically, in Study 1 we saw that the percentage of articles in JDM with open data increased dramatically from 8.6% to 87.4% after the introduction of their data policy (see **Table 12**). Moreover, in 2013 and 2014, 100% of the articles in JDM contained open data (see **Figure 4**). In the similar journal JBDM that did not introduce a data policy, none of the articles had open data in period 1, and only 1.7% of the articles had open data in period 2 (see **Table 12**). We found a similar pattern in Study 2. There, the articles in PLOS that were accompanied by open data increased from 4.5% to 55.9% after PLOS introduced a data sharing policy. In the comparable open access journal FP without such a stringent policy, we see no such increase (1.4% to 1.3%; see **Table 12**). Note that these percentages reflect whether data are actually available or not, so despite the worrying finding that roughly a third of the articles in Study 2 that promised data did not deliver (see the previous section), we still see a steep increase in the prevalence of open data in PLOS. In Study 3, we found that after the introduction of Open Practice Badges in PS, 16.9% of the articles earned an Open Data Badge. Previous research investigating the effectiveness of the badges in more detail found that after the introduction of the badges, data was more often available, correct, usable, and complete (Kidwell et al., 2016). These results are in line with the finding that journal submission guidelines in general can inspire desirable change in authors' behavior (Giofrè, Cumming, Fresc, Boedker, & Tressoldi, 2017; but see also Morris & Fritz, 2017).

Table 11: Number of articles in which data were promised or not and data were actually available or not, split up per journal. The cases in which data were promised but not available are printed in bold.

Journal	Data Available	Data Promised		
PLOS		**Yes**	**No**	**Total**
	Yes	328	32	360
	No	**134**	2,415	2,549
	Total	462	2,447	2,909
FP		**Yes**	**No**	**Total**
	Yes	9	6	15
	No	**3**	1,090	1,093
	Total	12	1,096	1,108

Table 12: Percentage of articles that was accompanied by open data, split up per journal and period. The periods were decided per study based on the dates that one of the journals implemented their open data policy.

		% Articles with open data			
		Before implementation		**After implementation**	
Study 1		**Up to April 2011**		**From April 2011**	
	JBDM (no data policy)	0%	N = 0/117	1.7%	N = 2/118
	JDM (data policy)	8.6%	N = 11/128	87.4%	N = 118/135
Study 2		**Up to March 2014**		**From March 2014**	
	FP (no stipulated data policy)	1.4%	N = 11/804	1.3%	N = 4/304
	PLOS (data policy)	4.5%	N = 110/2462	55.9%	N = 250/447
Study 3		**Up to 2014**		**From 2014**	
	PS (data policy)	Not coded		16.9%	N = 72/426

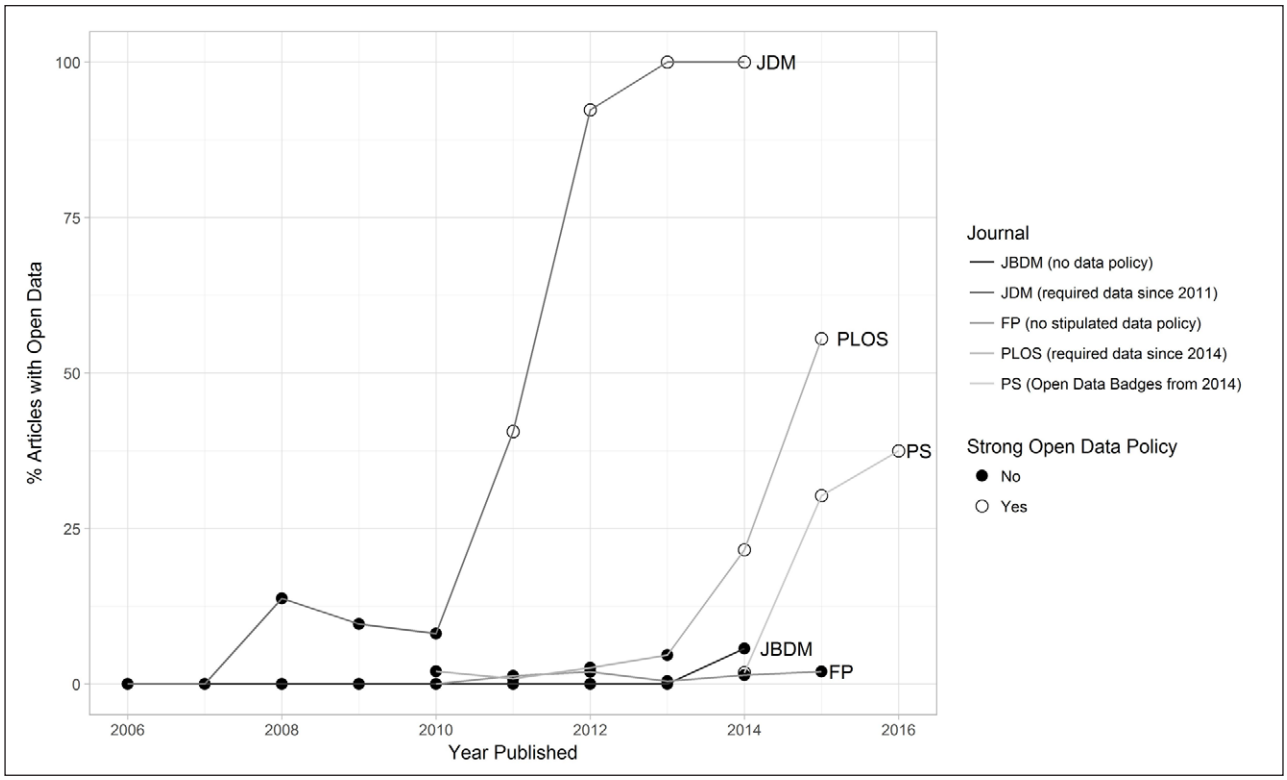

Figure 4: The percentage of articles per journal and year that had open data. A solid circle indicates that there was no (stipulated) open data policy at this point, and an open circle indicates that there was. The different line colors indicate the different journals. The journal abbreviations indicate the following: JBDM = Journal of Behavioral Decision Making, JDM = Judgment and Decision Making, FP = Frontiers in Psychology, PLOS = Public Library of Science, and PS = Psychological Science.

Note, however, that our design is observational, which does not allow us to draw a causal conclusion. It is imaginable that there is an alternative explanation for the increase in data availability after data policies were introduced. For instance, it is possible that the introduction of data policies changed the image of these journals, which inspired "open-science-minded" researchers who always share their data to submit to these journals instead of elsewhere. In that case, it would not be the policy per se that increased data availability, but the way these journals present themselves. We would need an experimental design to be able to investigate whether data policies actually lead to higher data availability. For instance, one way to investigate this would be to have one or multiple journals randomly assign submissions to a "required data sharing condition" and a control condition in which no explicit requests concerning data sharing are made. This way, any systematic difference in the prevalence of statistical reporting inconsistencies between conditions is likely to be due to the presence or absence of a data sharing request.

Open data and inconsistencies in significant vs. non-significant findings

In previous research we found that gross inconsistencies were more common in results reported as significant (1.56%) than as non-significant (0.97%), suggesting evidence for a systematic bias towards finding significance (Nuijten et al., 2016). This finding can have several causes, ranging from deliberately rounding down non-significant *p*-values (see also John, Loewenstein, & Prelec, 2012) to publication bias, which would primarily cause the *p*-values that are wrongly rounded *down* to be published. Because of this apparent emphasis on finding significant results, we looked in more detail at the difference between gross inconsistencies in results reported as significant and reported as non-significant.

We first tried to replicate our previous finding that there seems to be a systematic bias towards significant findings, using the aggregated data of Studies 1, 2, and 3. Interestingly, we found no clear evidence for such a bias in the current data. Of all 56,716 results reported as significant, 1.26% was flagged as a gross inconsistency, as opposed to 1.23% of the 22,344 results reported as non-significant.[21]

Furthermore, we looked at whether the probability of data sharing was related to the type of gross inconsistencies in a paper. Specifically, we looked at the proportion of articles sharing data that 1) did not contain a gross inconsistency, 2) contained at least one gross inconsistency in general, 3) contained a gross inconsistency in a result reported as non-significant, and 4) contained a gross inconsistency in a result reported as significant. We speculated that if gross inconsistencies in favor of finding significant results as opposed to non-significant results would be intentional, authors would be reluctant to share data. We therefore expected that articles with gross inconsistencies, especially those in the direction of statistical significance, would be accompanied by open data less often than articles without any gross inconsistencies.

Interestingly, in the aggregated data of Studies 1, 2, and 3 we found no such pattern (see **Table 13**). Articles without gross inconsistencies shared data in 8.6% of the cases,

Table 13: Categorization of all papers from Study 1, 2, and 3 with or without at least one (type of) gross inconsistency and whether they were accompanied by open data.

Articles that contain...			Data Available		% Articles with Data Available
			No	Yes	
No gross inconsistencies			5,473	516	8.6%
At least one gross inconsistency...			573	59	10.3%
	... in a result reported as n.s.		198	18	8.3%
	...in a result reported as sig.		402	44	9.9%
		...in a result where the recomputed *p*-value is .06	99	7	6.6%

whereas articles with gross inconsistencies shared data slightly more often: in 10.3% of the cases. We also found that articles with gross inconsistencies in the direction of finding a significant result shared data *more* often (9.9%) than articles with gross inconsistencies in the direction of non-significance (8.3%). This finding is not in line with the notion that authors are more reluctant to share data when their articles contain gross inconsistencies in favor of finding significant results.

We also looked at a special case of gross inconsistencies in favor of significance: *p*-values that were reported as significant, but upon recalculation turned out to be $p = .06$. This case most closely resembles the questionable research practice (QRP) of wrongly rounding down *p*-values as defined in (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; John et al., 2012). If such cases in our data were indeed the result of intentional QRPs, we would expect articles with such gross inconsistencies to be less likely to share data than articles without gross inconsistencies. Our findings seem to be in line with this notion (see **Table 13**). We found that articles that contained a *p*-value wrongly rounded down from $p = .06$ to $p < .05$ shared data in only 6.6% of the cases, as compared to articles without gross inconsistencies that shared data in 8.6% of the cases. Note that the sample sizes of these subgroup analyses are small, and these results should be interpreted with caution.

Discussion

We conducted three retrospective observational studies to test the hypotheses that data sharing and data sharing policy are negatively related to statistical reporting inconsistencies. Overall, we found that on average the prevalence of statistical inconsistencies was in line with the estimates of previous research (see Nuijten et al., 2016 for an overview). In Study 1, on average 9.3% of the *p*-values in an article were inconsistent and 1.1% grossly inconsistent, in Study 2 these numbers were 13.0% and 1.6%, respectively, and in Study 3, 9.3% and 1.1%, respectively. Contrary to what we hypothesized, we did not find consistent evidence that these inconsistencies were related to data sharing or data sharing policies. In Study 2, we did find that the probability of an inconsistency increased less steeply over time in PLOS after they installed a data policy, as compared to FP, that did not install such a policy. However, we did not find a similar pattern for gross inconsistencies, or for the other journals in Studies 1 and 3. Although we considered meta-analyzing the findings of our three studies, we decided not to, for two reasons. First, the results of three studies do not consistently point to a positive or negative effect. Second and most importantly, the three contexts are very different, which questions the use of combining them in one meta-analysis. Note that a random-effects meta-analysis with just three studies is generally also considered not to be very useful.

We ran several exploratory analyses and found some interesting results. First and foremost, we found that installing an open data policy seems to be highly effective: the proportion of articles with open data increased rapidly after the journals started requiring or recommending open data, as compared to the prevalence of open data in journals without an open data policy over time. This is in line with previous research that shows evidence that journal policy can encourage desirable change in research practices (Giofrè et al., 2017; Kidwell et al., 2016). Even though these results seem promising, they should be interpreted with care. These findings are not based on experimental data but on observational data, which only allow for correlational conclusions.

Even though data availability increased after open data policies were introduced, we did find that a surprisingly high number of cases in which an article stated the data were available, whereas in fact they were not. We found that roughly one third of the articles in PLOS and FP that promised open data did not deliver. This is comparable to the findings of Chambers (2017, p. 86), and Kidwell et al. (2016). Kidwell et al. (2016) showed that of the articles from journals without badges that promised open data, only 40.5% actually had data available. Kidwell et al., also found that articles in Psychological Science with an Open Data Badge had a much higher probability of the data being available, usable, and complete. These data suggest that even though installing an open data policy increases the availability of open data, there needs to be an extra check at the journal to verify if open data statements are justified.

Finally, contrary to previous findings (Nuijten et al., 2016), we found that gross inconsistencies in this sample

do not seem to be biased towards finding significant results. Furthermore, we found no evidence that articles with gross inconsistencies were less likely to have open data than articles without gross inconsistencies. Interestingly, we did find that articles were less likely to share data, when it contained a gross inconsistency in which a recalculated *p*-value of .06 was reported as < .05. This finding could indicate that some of the gross inconsistencies are intentionally wrongly rounded down *p*-values, which would lead to reluctance in sharing data. However, these findings are exploratory and based on a relatively small sample, so they should be interpreted with caution.

We recognize three main limitations in our studies. The first limitation is that our choice of retrospective observational designs limits the internal validity of the three studies, and prevents us from drawing causal conclusions. Because we did not randomly assign manuscripts to an "open data condition" and a control condition while keeping everything else constant, we were by definition not able to rule out alternative explanations for any relation between open data and reporting inconsistencies.

A second limitation is the lack of statistical power. Even though we downloaded a considerable number of articles for each study, the relatively low prevalence of inconsistencies dramatically decreases power to detect small effects. That said, we ran several power analyses that showed that if data sharing had a reasonable effect on the prevalence of inconsistencies, we should have had enough power to detect that. This means that even if data sharing or data sharing policy decreases inconsistencies, the effect is probably not strong enough to be of much practical value. However, the situation was more problematic for detecting any effects on the prevalence of gross inconsistencies. Our power analyses in Study 3 revealed serious shortcomings of multilevel analyses to analyze low incidence rates (as with gross inconsistencies) when based on a small number of observations per level-2 unit (article, in our case). More specifically, in our power analysis we used the baseline probability for gross inconsistencies as found in previous research (1.2% in PS; Nuijten et al., 2016), and found that in this case the type I error does not equal .05 but approaches zero instead, and the power to detect extremely large effects may not even exceed .05. This problem holds for Studies 1, 2, and 3, and consequently we do not put too much trust in the results of the multilevel logistic analyses concerning gross inconsistencies. We decided to still include them in the paper for the sake of completeness and because we preregistered these analyses. More generally, we recommend against using multilevel logistic regression analyses as a statistical method to analyze nested data characterized by a low incidence rate (e.g., less than 5%) in combination with level-2 units having few observations (e.g., eight observations per level-2 unit).

The third main limitation is that we used automated software to detect reporting inconsistencies. Even though statcheck was extensively validated (Nuijten et al., 2016; Nuijten et al., 2017), it will never be as accurate as a manual search. The main problem is that statcheck does not find all statistical results in a paper, due to variations in reporting style or problems in recognizing characters because of a journal's copy-editing process. It is possible that there is a systematic difference in the inconsistency rate between results that were or were not recognized by statcheck. For instance, maybe if researchers make an effort to report their results in APA style (which statcheck can detect), there is a lower probability of making a typo as compared to researchers who do not attempt to adhere to a strict reporting style. However, in statcheck's validity study there was no evidence for a systematic difference in reporting inconsistencies between results that were and were not picked up by statcheck, so we have no reason to assume that statcheck's estimates of the prevalence of inconsistencies is biased.

Taking these limitations into account, the results from these three studies are evidence against our hypotheses that data sharing and data sharing policies lead to fewer statistical reporting inconsistencies. We theorized that the precision needed to archive data in such a way that it is accessible and usable to others would also make typos and other errors in statistical reporting less likely. Additionally, we theorized that authors who are unsure about the quality of their analysis or know that there are errors in their work would be more reluctant to submit their work to a journal that requires data sharing. However, our data suggest that this is not the case; requiring data sharing in itself might not be enough to decrease the prevalence of statistical reporting inconsistencies in psychology.

Our findings are not directly in line with Wicherts et al. (2011), who found that reluctance to share data was related to, among other things, an increased rate of reporting inconsistencies. A meaningful difference between our studies is that we looked at whether data sets were published alongside the articles, whereas Wicherts et al., looked at (reluctance in) data sharing when explicitly requested. However, our findings are in line with those of Veldkamp et al. (2014) and Veldkamp, Hartgerink, Van Assen, and Wicherts (2017), who did not find support for their suggested "co-pilot" model in which they theorize that if multiple authors work on the analyses, the probability for reporting inconsistencies should decrease. Their rationale was that shared responsibility for the analysis and results section should (partly) eliminate human error and therefore increase accuracy of the reported results. However, they did not find a relation between co-piloting and the prevalence of statistical reporting inconsistencies. The combined evidence of our three studies and previous literature seems to point to the conclusion that strategies to increase more rigorous data management such as sharing data and collaborating on analyses is not enough to prevent statistical reporting inconsistencies. Even though this collection of findings is based on a limited set of journals, we see no immediate reason to expect differences in other journals. To find out which strategies could be effective in preventing statistical reporting inconsistencies, we need more research to investigate what causes them.

One way to help decreasing reporting inconsistencies is to use programs and apps such as statcheck (Epskamp & Nuijten, 2016; http://statcheck.io), or p-checker (Schönbrodt, 2015; http://shinyapps.org/apps/p-checker/) to quickly and easily check results for internal consistency. These programs can be used by authors themselves before submitting a paper in order to avoid mistakes in the published paper and having to file a correction. Similarly, journals themselves can also include these extra checks during peer review. The journal Psychological Science started using statcheck in their peer review process last year to prevent inconsistencies from ending up in the literature (http://www.psychologicalscience.org/publications/psychological_science/ps-submissions; retrieved on June 1, 2017), and the use of statcheck is recommended by the journals Stress & Health (Barber, 2017) and the new journal Advances in Methods and Practices in Psychological Science (http://www.psychologicalscience.org/publications/ampps/ampps-submission-guidelines; retrieved on June 1, 2017). Another solution to decrease the prevalence of reporting errors is to make use of Analytic Review (AR; Sakaluk, Williams, & Biernat, 2014), in which reviewers also check the analysis scripts and accompanying data files. The advantage of AR over automated programs is that a (human) reviewer can also check if the reported statistical analyses were the appropriate ones.

Even though we found no evidence that (recommended) data sharing is related to a decreased prevalence of statistical reporting inconsistencies, we still want to emphasize the importance of open data. Some of the greatest advantages of sharing data include, but are not limited to, the possibility to run secondary analyses to answer new questions, verify analyses of published work or examine the robustness of the original analyses, and compute specific effect sizes for meta-analyses (see Wicherts, 2013). Stating that "data are available upon request", as is APA policy, is often not enough to ensure availability (Vanpaemel et al., 2015; Wicherts et al., 2006). On top of that, sharing data upon request is not robust to time: how likely is it that the data are actually still available after ten years? Or fifty? Or even longer? Vines et al. (2014) found that the odds of data actually being available upon request dropped by 17% per year. To ensure availability over time it is necessary to publish data in online repositories. An example of a platform for doing so is the Open Science Framework (http://osf.io). Availability of raw data does not guarantee usability or completeness, so it is desirable to build in checks or review of data sets. For instance, it is possible to publish your data in the Journal of Open Psychology Data, in which your data is reviewed to see if it is archived well. There have been concerns about data sharing pointing at issues such as privacy (Finkel, Eastwick, & Reis, 2015), or the risk that "freeriders" will take advantage of your painstakingly collected data (but see Longo & Drazen, 2016). These are valid concerns, but in most cases, it is easy to come up with solutions tailored to the situation. For instance, the majority of experiments in psychology do not concern sensitive data and can easily be anonymized, and there are options to publish data online privately, and only make it public after a pre-specified period of time in order to first publish findings from these data yourself. Moreover, there is evidence that data sharing is associated with an increased citation rate (Piwowar, Day, & Fridsma, 2007).

In this paper, we used empirical methods to investigate one possible solution to the high prevalence of inconsistently reported statistical results. Reporting inconsistencies are only a small part of the problems related to the current "replication crisis" that psychology is facing (for an overview of these problems, see e.g., Shrout & Rodgers, 2017). Even so, we think that it is useful to treat problems in our scientific system (no matter how small) as empirical questions that we can solve by applying the scientific method. Research that aims to do so, such as this paper, adds to a growing body of literature on "meta-science" (Ioannidis, Fanelli, Dunne, & Goodman, 2015; Munafò et al., 2017). Improving the quality of our research is a complex endeavor and we will need much more research to understand where the biggest problems lie, what caused them, and how we can solve them. Even though we still have a long way to go, it is encouraging to see that journal policies and research practices are changing to accommodate open science.

Data Accessibility Statement

All the materials, data, and analysis scripts can be found on this paper's project page on the Open Science Framework: https://osf.io/538bc/.

Additional Files

The additional files for this article can be found as follows:

- **Detailed Analyses Study 3.** The full details on the analyses of hypotheses 1 through 6 and the ad-hoc solutions to the convergence problems can be found in the Supplemental Information at https://osf.io/4gx53/ and in the R code at https://osf.io/8e3gr/. DOI: https://doi.org/10.1525/collabra.102.s1
- **The Prevalence of Reporting Inconsistencies over Time.** More details about the general trends in (gross) inconsistencies over time can be found in the Supplemental Materials at https://osf.io/5j6tc/. DOI: https://doi.org/10.1525/collabra.102.s2

Notes

[1] The data sharing recommendation of JDM states: *"We encourage the submission of raw data at the time of review, and we include the data of accepted articles with the articles (unless this is for some reason difficult). We will also include stimuli, questionnaires, and code, when these are necessary to understand exactly what was done (again, unless this is difficult for some reason).",* (http://journal.sjdm.org/).

[2] At the time of writing, JBDM actually did implement a data sharing policy: "Journal of Behavioral Decision Making *encourages authors to share the data and other artefacts supporting the results in the paper by archiving it in an appropriate public repository. Authors should include a data accessibility statement,*

including a link to the repository they have used, in order that this statement can be published alongside their paper." (retrieved from http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1099-0771/homepage/ForAuthors.html, October 2017). We emailed JBDM's editorial office to ask when they changed their data policy and if it had stayed the same from 2006 to 2014, but unfortunately they did not reply. Based on information from web archives, we can see that in July 2017 this data policy was not yet part of the author guidelines and therefore does not affect our conclusions (information retrieved from https://web.archive.org/web/20170713015402/http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1099-0771/homepage/ForAuthors.html, October 2017).

[3] In the conversion from PDF to plain text, "=" signs were often translated to "¼". We adapted statcheck such that it would also recognize these cases. Furthermore, the downloaded articles contained a non-standardly reported test results that statcheck wrongly recognized as chi-square tests. This we also fixed in this adapted version of statcheck.

[4] See Uri Simonsohn's post on Data Colada: http://datacolada.org/2013/09/17/just_posting_it_works/ and the analysis of the case by Retraction Watch: http://retractionwatch.com/2013/09/10/real-problems-with-retracted-shame-and-money-paper-revealed/#more-15597.

[5] FP's data policy: *"To comply with best practice in their field of research, authors must also make certain types of data available to readers at time of publication in stable, community-supported repositories such as those listed below, unless in case of serious confidentiality concerns (for example, research involving human subjects). Although not mandatory, authors may also consider the deposition of additional data-types (see below)."* FP's editorial office let us know via email that they supported the TOP guidelines since 2015: *"Frontiers supports the Transparency and Openness Promotion (TOP) guidelines, which state that materials, data, and code described in published works should be made available, without undue reservation, to any qualified researcher, to expedite work that builds on previous findings and enhance the reproducibility of the scientific record."* Both quotes retrieved from http://home.frontiersin.org/about/author-guidelines, Materials and Data Policies, May 17, 2017.

[6] We could use systematic text searches because all research articles in FP have a standard header indicating the type of article. We included articles with the header "Original Research ARTICLE", "Clinical Trial ARTICLE", "Methods ARTICLE", and "Clinical Case Study ARTICLE", which resulted in 2,693 articles. We also wanted to extract whether the articles were received before or after PLOS' data sharing policy that came into effect March 1st 2014. In FP, this is also systematically indicated at the bottom of the article (e.g., "Received: 22 October 2010; Paper Pending Published: 10 November 2010; Accepted: 01 December 2010; Published online: 14 December 2010"). Because these dates were always reported in the same place and in the same way, we could use systematic text searches in R again to extract when the articles were received and published.

[7] The exact date at which the open data policy at PLOS was implemented is not entirely clear. In the editorial announcing the policy it was stated the policy was implemented at March 1st 2014 (Bloom et al., 2014), but at the data availability web page, it was stated that the starting date was March 3rd (http://journals.plos.org/plosone/s/data-availability). For our study we retained March 1st.

[8] Similar to articles in FP, PLOS articles also have a standard header indicating the type of article. Again, we used systematic text searches in R to identify the research articles, but for this it was not enough to only search for "Research Article", since this phrase could also just occur in the full text of the manuscript. We therefore also specified the context in which the phrase "Research Article" should occur. We included either the phrase "Open Access Peer-Reviewed Research Article" or "Browse Topics Research Article", rendering 7,700 articles from before the data sharing policy, and 1,515 articles from after the policy.

[9] Note that this probability is smaller than one would expect based on the general inconsistency prevalence in Nuijten et al. (2016). This is due to the estimation method in the power analysis, which takes into account the random intercept, resulting in a lower probability of an inconsistency than observed directly in the data.

[10] It is possible that our sample contained articles that contained reporting inconsistencies because those inconsistencies were the topic of investigation (Bakker & Wicherts, 2014; Veldkamp et al., 2014; Wicherts et al., 2011). However, this sample is so small that it is unlikely to affect our general conclusions.

[11] Discrepancies in coding data availability statements mainly arose in PLOS articles before they introduced the standardized data availability statements. There were also a few instances in which coders disagreed whether a statement such as "all *relevant* data are available" could be counted as a data availability statement.

[12] Discrepancies in coding data availability mainly arose in cases where the shared data deviated from "standard" experimental data (e.g., in a meta-analysis or in genetic research), or when data about the stimuli were confused with collected data.

[13] Technically, we should call this variable "Journal/Publisher", since the results from PLOS did not all come from a single article. However, for the sake of readability and consistency with the preprint, we will call this variable "Journal".

[14] See https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ for details.

[15] We note that Kidwell et al. (2016) found that some articles that did share data did not receive an Open Data Badge, but it was unclear why. Conversely, there were also articles with an Open Data Badge that did not have available data. Even though these cases were rare, they indicate that having one of the Open

Practice Badges is not necessarily a perfect indicator of open practice.

[16] "Please note: Psychological Science uses StatCheck, an R program written by Sacha Epskamp and Michele B. Nuijten that is designed to detect inconsistencies between different components of inferential statistics (e.g., t value, df, and p). StatCheck is not designed to detect fraud, but rather to catch typographical errors (which occur often in psychology; see https://mbnuijten.com/statcheck/). We run StatCheck only on manuscripts that are sent out for extended review and not immediately rejected after extended review. Authors are informed if StatCheck detects any inconsistencies. Authors are welcome to run StatCheck before submitting a manuscript (http://statcheck.io/)." Retrieved from http://www.psychologicalscience.org/publications/psychological_science/ps-submissions#OPEN, October 2017.

[17] Note that this probability is lower than one would expect based on the general inconsistency prevalence of roughly .10 in PS (Nuijten et al., 2016). This is due to the estimation of the regression coefficients, which takes into account the random intercept, resulting in a lower probability of an inconsistency than observed directly in the data.

[18] In the preregistration we stated that we had 2,307 articles in total, but this seems to have been a mistake.

[19] In the preregistration we stated that we had 576 articles in total, but this seems to have been a mistake.

[20] We thank Julia Rohrer for pointing this out to us in her review.

[21] Note that this does not add up to the total sample size of 79,784 extracted APA reported NHST results (Study 1: N = 6,482; Study 2: N = 52,376; Study 3: N = 20,926). This is because results reported as $p < .07$ could not be classified as significant or not significant, and these results were not included in this analysis.

Acknowledgements

We would like to thank Sofie Swaans and Elise Crompvoets for their assistance in coding the articles for open data. We would also like to thank the reviewers Charlotte Hartwright, Julia Rohrer, and an anonymous reviewer for their comments. Their input has considerably improved this manuscript.

Funding Information

The preparation of this article was supported by The Innovational Research Incentives Scheme Vidi (no. 452-11-004) from the Netherlands Organization for Scientific Research.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

- Contributed to conception and design: MN, JW, MVA, JB
- Contributed to acquisition of data: MN, JB, CV, LDA, MVA, JW
- Contributed to analysis and interpretation of data: MN, MVA, JW
- Drafted and/or revised the article: MN, MVA, JW
- Approved the submitted version for publication: MN, JB, CV, LDA, MVA, JW

References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & **Cubelli, R.** (2017). Questionable research practices among Italian research psychologists. *PLoS One, 12*(3). DOI: https://doi.org/10.1371/journal.pone.0172792

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & **Ioannidis, J. P. A.** (2011). Public availability of published research data in high-impact journals. *PLoS One, 6*(9), e24357. DOI: https://doi.org/10.1371/journal.pone.0024357

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

Anagnostou, P., Capocasa, M., Milia, N., Sanna, E., Battaggia, C., Luzi, D., & **Bisol, G. D.** (2015). When Data Sharing Gets Close to 100%: What Human Paleogenetics Can Teach the Open Science Movement. *PLoS One, 10*(3). DOI: https://doi.org/10.1371/journal.pone.0121409

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

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

Barber, L. K. (2017). Meticulous manuscripts, messy results: Working together for robust science reporting. *Stress and Health, 33*(2), 89–91. DOI: https://doi.org/10.1002/smi.2756

Baron, J. (2011). Acknowledgements and report for the year 2010. *Judgment and Decision Making, 6*(2), 1–3.

Berle, D., & **Starcevic, V.** (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research, 16*(4), 202–207. DOI: https://doi.org/10.1002/mpr.225

Bloom, T., Ganley, E., & **Winker, M.** (2014). Data access for the open access literature: PLOS's data policy. *PLoS biology, 12*(2), e1001797. DOI: https://doi.org/10.1371/journal.pbio.1001797

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

Ceci, S. J. (1988). Scientists Attitudes toward Data Sharing. *Science Technology & Human Values, 13*(1–2), 45–52.

Chamberlain, S., Boettiger, C., & **Ram, K.** (2014). rplos: Interface to PLoS Journals search API. R package version 0.4.0. http://CRAN.R-project.org/package=rplos.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press. DOI: https://doi.org/10.1515/9781400884940

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

Eich, E. (2014). Business not as usual. *Psychological science, 25*(1), 3–6. DOI: https://doi.org/10.1177/0956797613512465

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

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

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

Finkel, E. J., Eastwick, P. W., & **Reis, H. T.** (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology, 108*(2), 275–297. DOI: https://doi.org/10.1037/pspi0000007

Garcia-Berthou, E., & **Alcaraz, C.** (2004). Incongruence between test statistics and P values in medical papers. *Bmc Medical Research Methodology, 4*(1), 13. DOI: https://doi.org/10.1186/1471-2288-4-13

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., & **Tressoldi, P.** (2017). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PLoS One, 12*(4), e0175583. DOI: https://doi.org/10.1371/journal.pone.0175583

Hedrick, T. E. (1985). Justifications for and obstacles to data sharing. *Sharing research data*, 123–147.

Hubbard, R., & **Ryan, P. A.** (2000). The historical growth of statistical significance testing in psychology–and its future prospects. *Educational and Psychological Measurement, 60*, 661–681. DOI: https://doi.org/10.1177/0013164400605001

Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & **Goodman, S. N.** (2015). Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLoS biology, 13*(10), e1002264–e1002264. DOI: https://doi.org/10.1371/journal.pbio.1002264

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

Kidwell, M. C., Lazarevic, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Nosek, B. A., et al. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLoS biology*, 1–15. DOI: https://doi.org/10.1371/journal.pbio.1002456

Krawczyk, M., & **Reuben, E.** (2012). (Un)Available upon Request: Field Experiment on Researchers' Willingness to Share Supplementary Materials. *Accountability in Research: Policies and Quality Assurance, 19*, 175–186. DOI: https://doi.org/10.1080/08989621.2012.678688

Lindsay, D. S. (2017). Sharing Data and Materials in Psychological Science. *Psychological science, 28*(6), 699–702. DOI: https://doi.org/10.1177/0956797617704015

Longo, D. L., & **Drazen, J. M.** (2016). Data sharing. *The New England Journal of Medicine, 374*, 276–277. DOI: https://doi.org/10.1056/NEJMe1516564

Morris, P. E., & **Fritz, C. O.** (2017). Meeting the challenge of the Psychonomic Society's 2012 Guidelines on Statistical Issues: Some success and some room for improvement. *Psychonomic Bulletin & Review*, 1–7. DOI: https://doi.org/10.3758/s13423-017-1267-y

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., Ioannidis, J. P., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021. DOI: https://doi.org/10.1038/s41562-016-0021

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Yarkoni, T., et al. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425. DOI: https://doi.org/10.1126/science.aab2374

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

Nuijten, M. B., Van Assen, M. A. L. M., Hartgerink, C. H. J., Epskamp, S., & **Wicherts, J. M.** (2017). The validity of the tool "statcheck" in discovering statistical reporting inconsistencies. Preprint retrieved from: https://psyarxiv.com/tcxaj/.

Piwowar, H. A., Day, R. S., & **Fridsma, D. B.** (2007). Sharing detailed research data is associated with increased citation rate. *PLoS One, 2*(3), e308. DOI: https://doi.org/10.1371/journal.pone.0000308

Sakaluk, J., Williams, A., & **Biernat, M.** (2014). Analytic Review as a Solution to the Misreporting of Statistical Results in Psychological Science. *Perspectives on Psychological Science, 9*(6), 652–660. DOI: https://doi.org/10.1177/1745691614549257

Schönbrodt, F. D. (2015). p-checker: One-for-all *p*-value analyzer. Retrieved from: http://shinyapps.org/apps/p-checker/.

Shrout, P. E., & **Rodgers, J. L.** (2017). Psychology, Science, and Knowledge Construction: Broadening

Perspectives from the Replication Crisis. *Annual Review of Psychology, 69*(1).

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological science, 24*(10), 1875–1888. DOI: https://doi.org/10.1177/0956797613480366

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

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

Sterling, T. D., & **Weinkam, J. J.** (1990). Sharing Scientific-Data. *Communications of the Acm, 33*(8), 112–119. DOI: https://doi.org/10.1145/79173.791822

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & **Storms, G.** (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra, 1*(1), 1–5. DOI: https://doi.org/10.1525/collabra.13

Veldkamp, C. L. S., Hartgerink, C. H. J., Van Assen, M. A. L. M., & **Wicherts, J. M.** (2017). *Shared responsibility for statistical analyses and statistical Reporting errors in psychology articles published in PLOS ONE (2003–2016).* Retrieved from: https://psyarxiv.com/g8cjq.

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

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Rennison, D. J., et al. (2014). The availability of research data declines rapidly with article age. *Current biology, 24*(1), 94–97. DOI: https://doi.org/10.1016/j.cub.2013.11.014

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & **Kievit, R. A.** (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632–638. DOI: https://doi.org/10.1177/1745691612463078

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature, 480*, 7. DOI: https://doi.org/10.1038/480007a

Wicherts, J. M. (2013). Science revolves around the data. *Journal of Open Psychology Data, 1*(1), e1. DOI: https://doi.org/10.5334/jopd.e1

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

Wicherts, J. M., Borsboom, D., Kats, J., & **Molenaar, D.** (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726–728. DOI: https://doi.org/10.1037/0003-066X.61.7.726

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Bourne, P. E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*, 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.102.pr

How to cite this article: Nuijten, M. B., Borghuis, J., Veldkamp, L. S. C., Dominguez-Alvarez, L., Van Assen, A. L. M. M., & Wicherts, J. M. (2017). Journal Data Sharing Policies and Statistical Reporting Inconsistencies in Psychology. *Collabra: Psychology,* 3(1): 31, pp. 1–22. DOI: https://doi.org/10.1525/collabra.102

Senior Editor: Simine Vazire

Editor: Chris Chambers

Submitted: 13 July 2017 **Accepted:** 15 November 2017 **Published:** 15 December 2017

UNIVERSITY of CALIFORNIA PRESS | Collabra

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS