Reanalyzing Head et al. (2015): Investigating the robustness of widespread *p*-hacking

Chris H Hartgerink Corresp. [1]

[1] Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands

Corresponding Author: Chris H Hartgerink
Email address: c.h.j.hartgerink@tilburguniversity.edu

Head et al. (2015b) provided a large collection of *p*-values that, from their analytic perspective, indicates widespread statistical significance seeking (i.e., *p*-hacking). This paper inspects this result for robustness. They correctly argue that an aggregate *p*-value distribution could show a bump below .05 when left-skew *p*-hacking occurs frequently. Theoretically, the *p*-value distribution should be a smooth, decreasing function, but the distribution of reported *p*-values shows systematically more reported *p*-values for .01, .02, .03, .04, and .05. Moreover, the elimination of $p = .045$ and $p = .05$, as done in the original paper, is debatable. Given that systematically more *p*-values are reported to two decimal places and the disputable selection of the bins $.04 < p < .045$ versus $.045 < p < .05$, I did not exclude $p = .045$ and $p = .05$, and I adjusted the bin selection to $.03875 < p \leq .04$ versus $.04875 < p \leq .05$. Results of the reanalysis indicate that no evidence for left-skew *p*-hacking remains when we take into account a second-decimal reporting tendency. Taking into account reporting tendencies is especially important because this dataset does not allow for the recalculation of the *p*-values. Moreover, given the weight of the findings by Head et al. (2015b), it is important that these findings are robust to choices that can be debated, if the conclusion is to be considered unequivocal. Although no evidence of widespread left-skew *p*-hacking is found in this reanalysis, this does not mean that there is no *p*-hacking at all. These results nuance the conclusion by Head et al. (2015b), indicating that the results are not robust and that the evidence for widespread left-skew *p*-hacking is ambiguous at best.

Reanalyzing Head et al. (2015): Investigating the robustness of widespread p-hacking

Chris HJ Hartgerink[1]

[1]**Department of Methodology and Statistics, Tilburg University, the Netherlands**

ABSTRACT

Head et al. (2015b) provided a large collection of p-values that, from their analytic perspective, indicates widespread statistical significance seeking (i.e., p-hacking). This paper inspects this result for robustness. They correctly argue that an aggregate p-value distribution could show a bump below .05 when left-skew p-hacking occurs frequently. Theoretically, the p-value distribution should be a smooth, decreasing function, but the distribution of reported p-values shows systematically more reported p-values for .01, .02, .03, .04, and .05. Moreover, the elimination of $p = .045$ and $p = .05$, as done in the original paper, is debatable. Given that systematically more p-values are reported to two decimal places and the disputable selection of the bins $.04 < p < .045$ versus $.045 < p < .05$, I did not exclude $p = .045$ and $p = .05$, and I adjusted the bin selection to $.03875 < p \leq .04$ versus $.04875 < p \leq .05$. Results of the reanalysis indicate that no evidence for left-skew p-hacking remains when we take into account a second-decimal reporting tendency. Taking into account reporting tendencies is especially important because this dataset does not allow for the recalculation of the p-values. Moreover, given the weight of the findings by Head et al. (2015b), it is important that these findings are robust to choices that can be debated, if the conclusion is to be considered unequivocal. Although no evidence of widespread left-skew p-hacking is found in this reanalysis, this does not mean that there is no p-hacking at all. These results nuance the conclusion by Head et al. (2015b), indicating that the results are not robust and that the evidence for widespread left-skew p-hacking is ambiguous at best.

Keywords: nhst, p-hacking, qrps, reanalysis

INTRODUCTION

Head et al. (2015b) provided a large collection of p-values that, from their analytic perspective, indicates widespread statistical significance seeking (i.e., p-hacking) throughout the sciences. This result has been questioned from an epistemological perpective, where analyzing all reported p-values in research articles answers the supposedly inappropriate question of evidential value across all results (Simonsohn et al., 2015). Adjacent to epistemological concerns, the robustness of widespread p-hacking in these data can be questioned. Head et al. (2015b) had to make several analytic decisions, which might have affected the results. In this paper I evaluate the analytic strategy with which Head et al. (2015b) found widespread p-hacking and propose that this effect is not robust to justifiable changes in the analytic strategy.

The p-value distribution of a set of true- and null results without p-hacking should be a mixture distribution of only the uniform p-value distribution under the null hypothesis H_0 and right-skew p-value distributions under the alternative hypothesis H_1. Questionable, p-hacking behaviors affect the distribution of statistically significant p-values, potentially resulting in left-skew (i.e., a bump) below .05, but not necessarily so (Hartgerink et al., 2016; Lakens, 2014; Bishop and Thompson, 2016). An example of a questionable behavior that can result in left-skew is optional stopping (i.e., data peeking) if the null hypothesis is true (Lakens, 2014).

Consequently, Head et al. (2015b) correctly argue that an aggregate p-value distribution could show a bump below .05 when left-skew p-hacking occurs frequently. Questionable behaviors seeking just statistically significant results, such as (but not limited to) the aforementioned optional stopping under H_0, could result in bump below .05. Hence, a bump below .05 is a sufficient condition for the presence of specific forms of p-hacking. However, this bump below .05 is not a necessary condition, because other types of p-hacking can still occur without a bump below .05 presenting itself (Hartgerink et al., 2016; Lakens, 2014; Bishop and Thompson, 2016). For example, one might use optional stopping when there is a true effect or conduct multiple analyses, but only report that statistical test which yielded the smallest p-value. Therefore, if no bump of statistically significant p-values is found, this does not exclude that p-hacking occurs at a large scale.

In the current paper, the conclusion from Head et al. (2015b) is inspected for robustness. Their conclusion is that the data fullfill the sufficient condition for p-hacking (i.e., show a bump below .05), hence, provides evidence for the presence of specific forms of p-hacking. The robustness of this conclusion is inspected in three steps: (i) explaining the data and analytic strategies (original and reanalysis), (ii) reevaluating the evidence for a bump below .05 (i.e., the sufficient condition) based on the reanalysis, and (iii) discussing whether this means that there is no widespread p-hacking in the literature.

DATA AND METHODS

In the original paper, over two million reported p-values were mined from the Open Access subset of PubMed central. PubMed central indexes the biomedical and life sciences and permits bulk downloading of full-text Open Access articles (`https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`). By mining these full-text articles for p-values, Head et al. (2015b) extracted more than two million p-values in total and analyzed a subset of statistically significant p-values ($\alpha = .05$). Their mining procedure included all reported p-values, including those that were reported without an accompanying test statistic. For example, the p-value from the result $t(59) = 1.75, p > .05$ was included, but also a lone $p < .05$.

Head et al. (2015b) their data analytic strategy focused on comparing frequencies in the last and penultimate bins from .05 at a binwidth of .005 (i.e., $.04 < p < .045$ versus $.045 < p < .05$). Based on the tenet that a sufficient condition for p-hacking is a bump of p-values below .05 (Simonsohn et al., 2014), sufficient evidence for p-hacking is present if the last bin has a significantly higher frequency than the penultimate bin in a binomial test. Applying the binomial test to two frequency bins has previously been used in publication bias research (Caliper test; Gerber et al., 2010; Kühberger et al., 2014), applied here specifically to test for p-hacking behaviors that result in a bump below .05. The binwidth of .005 and the bins $.04 < p < .045$ and $.045 < p < .05$ were chosen by Head et al. (2015b) because they expected the signal of this form of p-hacking to be strongest in this part of the distribution. They excluded $p = .05$ "because [they] suspect[ed] that many authors do not regard $p = 0.05$ as significant" (p.4).

Figure 1 shows the selection of p-values in Head et al. (2015b) in two ways: in green, which shows the results as analysed by Head et al. (i.e., $.04 < p < .045$ versus $.045 < p < .05$), and in grey, which shows the entire distribution of significant p-values available to Head et al. after eliminating those results depicted in black. The two green bins (i.e., the sum of the grey bins in the same range) show a bump below .05, which indicates p-hacking. The grey histogram in Figure 1 shows a more fine-grained depiction of the p-value distribution and does not clearly show a bump below .05, because it is dependent on which bins are compared. However, the grey histogram clearly indicates that results around the second decimal tend to be reported more frequently when $p \geq .01$.

Theoretically, the p-value distribution should be a smooth, decreasing function, but the grey distribution shows systematically more reported p-values for .01, .02, .03, .04 (and .05 when the black histogram is included). As such, there seems to be a tendency to report p-values to two decimal places, instead of three. For example, $p = .041$ might be correctly rounded to $p = .04$. A potential post-hoc explanation is that three decimal reporting of p-values is a relatively recent standard, if a standard at all. For example, it has only been prescribed since 2010 in psychology (APA, 2010), where it previously prescribed two decimal reporting (APA, 1983, 2001). Given the results, it seems reasonable to assume that other fields might also report to two decimal places instead of three, most of the time.

Moreover, the analytic strategy by Head et al. (2015b) eliminates $p = .045$ without justification and $p = .05$ based on a potentially invalid assumption of when researchers regard results as statistically significant. $P = .045$ is not included in the bins selected ($.04 < p < .045$ versus $.045 < p < .05$), while seriously affecting the results. If $p = .045$ is included, no evidence of a bump below .05 is found (the left black bin in Figure 1 is then included; frequency $.04 < p \leq .045 = 20114$ versus $.045 < p < .05 = 18132$). Moreover, upon inspecting the original code to test for a bump below .05 (Head et al., 2015a), the inclusion or exclusion of the endpoints of the bins is not consistent. The endpoints are excluded when comparing $.04 < p < .045$ versus $.045 < p < .05$, but the lower end is included when comparing $.03 \leq p < .04$ versus $.04 \leq p < .05$. $P = .05$ was consistently excluded because Head et al. (2015b) assumed researchers

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2439v1 | CC BY 4.0 Open Access | rec: 12 Sep 2016, publ: 12 Sep 2016

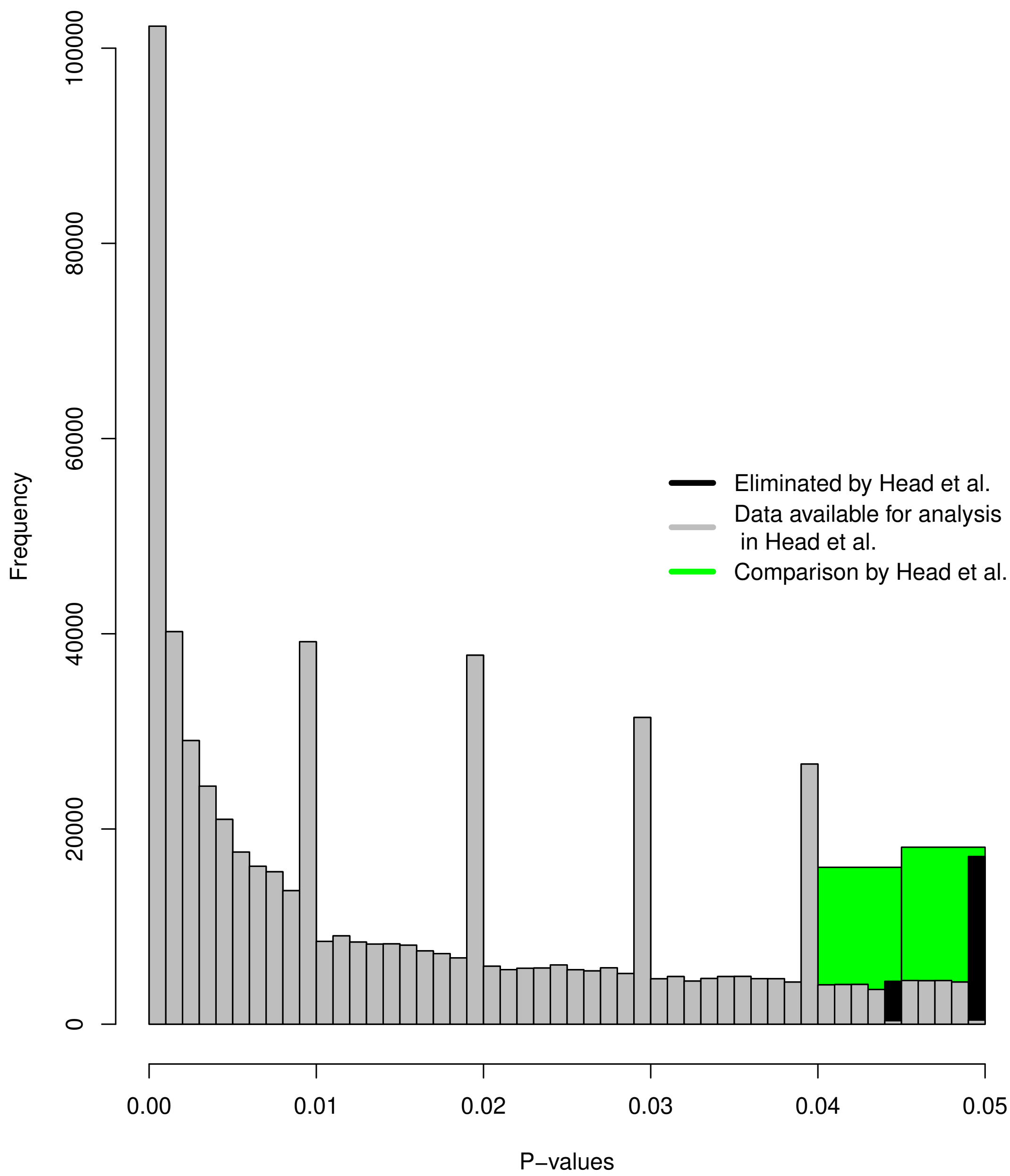

Figure 1. Histograms of p-values as selected in Head et al. (in green; $.04 < p < .045$ versus $.045 < p < .05$), the significant p-value distribution as selected in Head et al. (in grey; binwidth = .00125). The green and grey histograms exclude $p = .045$ and $p = .05$; the black histogram shows the frequencies of results that are omitted because of this.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2439v1 | CC BY 4.0 Open Access | rec: 12 Sep 2016, publ: 12 Sep 2016

did not interpret this as statistically significant. Researchers interpret $p = .05$ as statistically significant more frequently than they thought: 94% of 236 cases investigated by Nuijten et al. (2015) interpreted $p = .05$ as statistically significant, indicating this assumption might not be valid.

Given that systematically more p-values are reported to two decimal places and the disputable selection of the bins $.04 < p < .045$ versus $.045 < p < .05$, I did not exclude $p = .045$ and $p = .05$, and I adjusted the bin selection to $.03875 < p \leq .04$ versus $.04875 < p \leq .05$. Visually, the newly selected data are the grey and black bins from Figure 1 combined, where the rightmost black bin (i.e., $.04875 < p \leq .05$) is compared with the large grey bin at .04 (i.e., $.03875 < p \leq .04$). The bins $.03875 < p \leq .04$ and $.04875 < p \leq .05$ were selected to take into account that the data show systematically more p-values reported to two decimal places, which might indicate a reporting tendency. This altered bin selection takes such a reporting tendency into account and consequently includes the information available in these data.

The reanalytic strategy for the bins $.03875 < p \leq .04$ and $.04875 < p \leq .05$ is similar to Head et al. (2015b) and applies the Caliper test to detect a bump below .05, with the addition of Bayesian Caliper tests. The Caliper test investigates whether the bins are equally distributed or that the penultimate bin (i.e., $.03875 < p \leq .04$) contains more results than the ultimate bin (i.e., $.04875 < p \leq .05$; $H_0 : Proportion \leq .5$). Sensitivity analyses were also conducted, altering the binwidth from .00125 to .005 and .01. Moreover, the analyses were conducted for both the p-values extracted from the abstracts- and the results sections separately.

The results from the Bayesian Caliper test and the traditional, frequentist Caliper test give results with different interpretations. The p-value of the Caliper test gives the probability of more extreme results if the null hypothesis is true, but does not quantify the probability of the null- and alternative hypothesis. The Bayes Factor (BF) quantifies the probabilities of the hypotheses in the model and creates a ratio, either as BF_{10}, the alternative hypothesis versus the null hypothesis, or vice versa, BF_{01}. A BF of 1 indicates that both hypotheses are equally probable, given the data. In this specific instance, BF_{10} is computed and values > 1 can be interpreted, for our purposes, as: the data are more likely under p-hacking that results in a bump below .05 (i.e., left-skew p-hacking) than under no left-skew p-hacking. BF_{10} values < 1 indicate that the data are more likely under no left-skew p-hacking than under left-skew p-hacking. The further removed from 1, the more evidence in the direction of either hypothesis is available. For the current analyses, the prior belief of presence or absence of p-hacking was assumed to be equal.

REANALYSIS RESULTS

Results of the reanalysis indicate that no evidence for a bump below .05 remains when we take into account a second-decimal reporting tendency. Reanalyses showed no evidence for left-skew p-hacking, $Proportion = .417, p > .999, BF_{10} < .001$ for the Results sections and $Proportion = .358, p > .999, BF_{10} < .001$ for the Abstract sections. Table 1 summarizes these results for alternate binwidths (.00125, .005, and .01) and shows results are consistent across different binwidths. Separated per discipline, no binomial test for left-skew p-hacking is statistically significant in either the Results- or Abstract sections (see the Supplemental File). This indicates that the evidence for p-hacking that results in a bump below .05, as presented by Head et al. (2015b), seems to not be robust to minor analytic changes such as taking into account the tendency to report p-values to two decimal places.

DISCUSSION

Head et al. (2015b) collected p-values from full-text articles and analyzed these for p-hacking, concluding that p-hacking is widespread throughout the sciences. Given the weight of such a finding, I inspected whether evidence for widespread p-hacking was robust to some substantively justified changes in the data selection. After taking into account systematically more p-values that are reported to the second decimal and including $p = .05$, the results indicate that evidence for widespread p-hacking, as presented by Head et al. (2015b) is not robust to these analytic changes. The conclusion drawn by Head et al. (2015b) might still be correct, but the data do not undisputably show so. Moreover, even if there is no p-hacking that results in a bump of p-values below .05, other forms of p-hacking that do not cause such a bump can still be present and prevalent (Hartgerink et al., 2016; Lakens, 2014; Bishop and Thompson, 2016).

Taking into account reporting tendencies is especially important because this dataset does not allow for the recalculation of the p-values. Previous research has indicated that when the recalculated p-value distribution is inspected, the theoretically expected smooth distribution does occur even when the reported

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2439v1 | CC BY 4.0 Open Access | rec: 12 Sep 2016, publ: 12 Sep 2016

		Abstracts	Results
Binwidth = .00125	(.03875 – .04)	4597	26047
	(.04875 – .05)	2565	18664
	Proportion	0.358	0.417
	p	$>.999$	$>.999$
	BF_{10}	$<.001$	$<.001$
Binwidth = .005	(.035 – .04)	6641	38537
	(.045 – .05)	4485	30406
	Proportion	0.403	0.441
	p	$>.999$	$>.999$
	BF_{10}	$<.001$	$<.001$
Binwidth = .01	(.03 – .04)	9885	58809
	(.04 – .05)	7250	47755
	Proportion	0.423	0.448
	p	$>.999$	$>.999$
	BF_{10}	$<.001$	$<.001$

Table 1. Results of the reanalysis across various binwidths (i.e., .00125, .005, .01) and different sections of the paper.

p-value distribution shows reporting tendencies (Hartgerink et al., 2016). Given that the text-mining procedure implemented by Head et al. (2015b) does not allow for recalculation of p-values, the effect of reporting tendencies needs to mitigated by altering the analytic strategy.

Even after mitigating the effect of reporting tendencies, these analyses were all conducted on a set of aggregated p-values, which can either detect p-hacking that results in a bump of p-values below .05 if it is widespread, but not prove that no p-hacking is going on in any of the individual papers. Firstly, there is the risk of an ecological fallacy. These analyses take place at the aggregate level, but there might still be research papers that show a bump below .05 at the paper level. Secondly, some forms of p-hacking also result in right-skew, which is not picked up by the Caliper test and is difficult to detect in a set of heterogeneous results (we attempted to detect this in Hartgerink et al., 2016). As such, if any detection of p-hacking is attempted, this should be done at the paper level and after careful scrutiny of which results are included (Simonsohn et al., 2015; Bishop and Thompson, 2016).

LIMITATIONS AND CONCLUSION

In this reanalysis two limitations remain with respect to the data analysis. First, selecting the bins just below .04 and .05 results in selecting non-adjacent bins. Hence, the test might be less sensitive to detect a bump below .05. In light of this limitation I ran the original analysis from Head et al. (2015b), but included the second decimal (i.e., $.04 \leq p < .045$ versus $.045 < p \leq .05$). This analysis also yielded no evidence for a bump of p-values below .05, $Proportion = .431, p > .999, BF_{10} < .001$. Second, the selection of only exactly reported p-values might have distorted the p-value distribution due to reporting tendencies in rounding. For example, a researcher with a p-value of .047 might be more likely to report $p < .05$ than a researcher with a p-value of .037 reporting $p < .04$. Given that these analyses exclude all values reported as $p < X$, this could have affected the results. There is some indication that this rounding tendency is a bit stronger around .05 than around .04 (a factor of 1.25 approximately based on the original Figure 5; Krawczyk, 2015), which might result in an underrepresentation of p-values around .05.

Given the weight of the findings by Head et al. (2015b), it is important that these findings are robust to choices that are not unequivocal. In this paper, I explained why a different analytic strategy can be justified, and as a result no evidence of widespread p-hacking that results in a bump of p-values below .05 is found. Although this does not mean that there no p-hacking occurs at all, the conclusion by Head et al. (2015b) should not be taken at face value considering that the results are not robust to (minor) choices in the data analytic strategy. As such, the evidence for widespread left-skew p-hacking is ambiguous at best.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2439v1 | CC BY 4.0 Open Access | rec: 12 Sep 2016, publ: 12 Sep 2016

SUPPORTING INFORMATION

S1 File. Full reanalysis results per discipline: `https://osf.io/aby85/`.

All files (e.g., manuscript, analysis scripts) for this article: `http://dx.doi.org/10.5281/zenodo.61398`

ACKNOWLEDGMENTS

Joost de Winter, Marcel van Assen, Robbie van Aert, Michèle Nuijten, and Jelte Wicherts provided fruitful discussion or feedback on the ideas presented in this paper. The end result is the author's sole responsibility.

REFERENCES

APA (1983). *Publication manual of the American Psychological Association*. American Psychological Association, Washington, DC, 3rd edition.

APA (2001). *Publication manual of the American Psychological Association*. American Psychological Association, Washington, DC, 5th edition.

APA (2010). *Publication manual of the American Psychological Association*. American Psychological Association, Washington, DC, 6th edition.

Bishop, D. V. M. and Thompson, P. A. (2016). Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ*, 4:e1715.

Gerber, A., Malhotra, N., Dowling, C., and Doherty, D. (2010). Publication bias in two political behavior literatures. *American Politics Research*, 38:591–613.

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., and van Assen, M. A. L. M. (2016). Distributions of p-values smaller than .05 in psychology: what is going on? *PeerJ*, 4:e1935.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015a). Data from: The extent and consequences of p-hacking in science.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015b). The extent and consequences of p-hacking in science. *PLOS Biology*, 13:e1002106.

Krawczyk, M. (2015). The search for significance: A few peculiarities in the distribution of P values in experimental psychology literature. *PloS one*, 10(6):e0127872.

Kühberger, A., Fritz, A., and Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PloS one*, 9:e105825.

Lakens, D. (2014). What p -hacking really looks like: A comment on Masicampo and LaLande (2012). *The Quarterly Journal of Experimental Psychology*, 68(4):829–832.

Nuijten, M. B., Hartgerink, C. H. J., Van Assen, M. A. L. M., Epskamp, S., and Wicherts, J. M. (2015). The Prevalence of Statistical Reporting Errors in Psychology (1985-2013). *Behavior Research Methods*.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143:534–47.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to Ulrich and Miller (2015). *Journal of experimental psychology. General*, 144(6):1146–1152.