**Reanalyzing Head et al. (2015): No widespread p-hacking after all?**

C.H.J. Hartgerink

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

\*Corresponding author

E-mail: [c.h.j.hartgerink@tilburguniversity.edu](mailto:c.h.j.hartgerink@tilburguniversity.edu)

Megan Head and colleagues [1] provide a large collection of p-values that, from their perspective, indicates widespread statistical significance seeking (i.e., p-hacking) throughout the sciences. The analyses that form the basis of their conclusions operate on the tenet that p-hacked papers show p-value distributions that are left skew below .05 [2]. In this paper I evaluate their selection choices and analytic strategy and show that these affect the results substantially. The version control of this paper is available at <https://osf.io/sxafg/>.

The p-value distribution of a set of heterogeneous results, as collected by Head et al., should be a mixture distribution of only the uniform p-value distribution under the null hypothesis *H0* and right-skew p-value distributions under the alternative hypothesis *H1*. Questionable, p-hacking behaviors affect the p-value distribution. An example is optional stopping, which causes a peak of p-values just below .05 only if the null hypothesis is true [3].

Head et al. correctly argue that an aggregate p-value distribution could show a peak below .05 if optional stopping under the null, or other behaviors with similar effects, occurs frequently. Consequently, a peak below .05 (i.e., left-skew), is a sufficient condition for the presence of specific forms of p-hacking. However, this peak below .05 is not a necessary condition, because other types of p-hacking do not cause such a peak. For example, one might use optional stopping when there is a true effect [3] or conduct multiple analyses, but only report that which yielded the smallest p-value. Therefore, if no peak is found, this does not exclude that p-hacking occurs at a large scale.

This paper is structured into three parts: (i) explaining the data analytic strategy of the reanalysis, (ii) reevaluating the evidence for left-skew p-hacking based on the reanalysis, and (iii) discussing the findings in light of the literature.

## Reanalytic strategy

Head and colleagues their data analytic strategy focused on comparing frequencies in the last and penultimate bins from .05 at a binwidth of .005. Based on the tenet that p-hacking introduces a left-skew p-distribution [2], evidence for p-hacking is present if the last bin has a sufficiently higher frequency than the penultimate one in a binomial test. Applying the binomial test to two frequency bins has previously been used in publication bias research and is typically called a Caliper test [4, 5], applied here specifically to test for left-skew p-hacking.

The two panels in Fig 1 describe the selection of p-values in the original and current paper. The top panel shows the selection made by Head et al..04 < *p* < .045 versus .045 < *p* < .05), where the right bin shows a slightly higher frequency than the left bin. This is the evidence Head et al. found for p-hacking. However, if we expand the range and look at the entire distribution, we see that this is an unrepresentative part of the distribution of significant p-values.

Histogram of p-values as selected in Head et al. (.04 < *p* < .045 versus .045 < *p* < .05; top) and the full *p*-value distribution ≤ .05 (binwidth = .00125; bottom).

The bottom panel in Fig 1 indicates there is a reporting tendency at the second decimal for p-values larger than or equal to .01. If no reporting tendencies existed, the distribution would show a reasonably smooth distribution, resembling the distribution between 0 and .01. However, the depicted distribution violates this, where p-value frequencies drastically increase at each second decimal place in the distribution. A post-hoc explanation for this is that three decimal reporting of p-values has only been prescribed since 2010 in psychology [8], where it previously prescribed two decimal reporting [7, 8]. Because reporting has occurred to the second decimal place for a long time and can be seen to have a substantial effect on the distribution, I think it is important to take this into account in the bin selection.

Head et al. selected the bins as indicated in the top panel in Fig 1, removing the second decimal. For their tests of p-hacking, they compared the bin frequency of the adjacent bins .04 < *p* < .045 versus .045 < *p* < .05. The original authors “suspect that many authors do not regard *p* = .05 as significant” [1], which is why they eliminate the second decimal from their analyses by using the selection criterion < .05. Previous investigation of p-values reported as exactly .05 revealed that 94.3% of 236 cases interpret this as statistically significant [9].

This contradicts the premise that most researchers do not interpret *p* = .05 as significant, which removes the reason for eliminating the second decimal. Consequently, only exactly reported p-values smaller than or equal to .05 were retained for the reanalyses, whereas Head et al. retained only exactly reported p-values smaller than .05. Moreover, because of reporting tendencies and the inclusion of the second decimal, the analyses need to compare the frequencies below .04 and .05 (e.g., .03875 < *p* < .04 versus .04875 < *p* < .05 for binwidth .00125). This corresponds to the two bins shown in the bottom panel of Fig 1 at .04 and .05.

In this paper, binomial proportion tests for left-skew p-hacking were conducted in both the frequentist and Bayesian framework, where *H0: Prop. ≤ .5*. The frequentist p-value gives the probability of the data if the null hypothesis is true, but does not quantify the probability of the null and alternative hypotheses. A Bayes Factor (*BF*) quantifies these latter probabilities, either as *BF10*, the alternative hypothesis versus the null hypothesis, or vice versa, *BF01*. A *BF* of 1 indicates that both hypotheses are equally probable, given the data. In this specific instance, *BF10* is computed and values > 1 can be interpreted, for our purposes, as: the data are more likely under left-skew p-hacking than under no left-skew p-hacking. *BF10* 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 one hypothesis, which were assumed to be equally likely in the prior distribution. For the current analyses, equal priors were assumed.

## Reanalysis results

Results of the reanalysis

indicate that no evidence for left-skew p-hacking remains when we take into account a second-decimal reporting bias. Initial sensitivity analyses using the original analysis script strengthened original results after eliminating DOI selection and using *p* ≤ .05 as selection criterion instead of *p* < .05. However, as explained above, this result is confounded due to not taking into account the second decimal. Reanalyses across all disciplines showed no evidence for left-skew p-hacking, *Prop. =* .417, *p* > .999, *BF10 <* .001 for the Results sections and *Prop. =* .358, *p* > .999, *BF10 <* .001 for the Abstract sections. These results are not dependent on binwidth .00125, as is seen in Table 1 where results for alternate binwidths are shown. Separated per discipline, no binomial test for left-skew p-hacking is statistically significant in either the Results- or Abstract sections (see S1 File). This indicates that the effect found originally by Head and colleagues does not hold when we take into account that reported p-values show reporting bias at the second decimal.

**Table 1. Results of reanalysis across various binwidths (i.e., .00125, .005, .01).**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Abstracts | Results |
| Binwidth = .00125 | (.03875−.04) | 4597 | 26047 |
|  | (.04875−.05) | 2565 | 18664 |
|  | *Prop*. | 0.358 | 0.417 |
|  | *p* | >.999 | >.999 |
|  | *BF10* | <.001 | <.001 |
| Binwidth = .005 | (.035−.04) | 6641 | 38537 |
|  | (.045−.05) | 4485 | 30406 |
|  | *Prop*. | 0.403 | 0.441 |
|  | *p* | >.999 | >.999 |
|  | *BF10* | <.001 | <.001 |
| Binwidth = .01 | (.03−.04) | 9885 | 58809 |
|  | (.04−.05) | 7250 | 47755 |
|  | *Prop*. | 0.423 | 0.448 |
|  | *p* | >.999 | >.999 |
|  | *BF10* | <.001 | <.001 |

## Discussion

The current reanalysis thus finds no evidence for widespread left-skew p-hacking. This might seem inconsistent with previous findings, such as the low replication rates in psychology [10] or high self-admission rate of p-hacking [11]. However, these results are not necessarily inconsistent because they are not mutually exclusive, as explained below.

Low replication rates could be caused by widespread p-hacking, but can also occur under systemic low power [12, 13]. Previous research has indicated low power levels in, for example, psychology [14, 15] and randomized clinical trials [16]. As a consequence of low power it is often argued that there is a high prevalence of false positives [17], which would result in low replication rates.

Additionally, high self-admission rates of p-hacking [11] pertain to such behaviors occurring at least once. Even if there is widespread occurrence of p-hacking across researchers, this does not necessitate frequent occurrence. In other words, a researcher might admit to having p-hacked sometime during his career, but this does not necessitate that it occurred frequently. Moreover, as noted in the introduction, not all p-hacking behaviors lead to left-skew in the p-value distribution. The method used to detect p-hacking in this paper is sensitive to only left-skew p-hacking and it is therefore possible that other types of p-hacking occur, but are not detected.

In this reanalysis two minor 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 detecting left-skew p-hacking. In light of this limitation I ran the original analysis from Head et al., but included the second decimal, which resulted in the comparison of .04 ≤ *p* < .045 versus .045 < *p* ≤ .05. This analysis also yielded no evidence for left-skew p-hacking, *Prop. =* .457, *p* > .999, *BF10 <* .001. Second, the selection of only exactly reported p-values might have distorted the p-value distribution due to minor rounding biases. Previous research has indicated that p-values are somewhat more likely to be rounded to .05 rather than to .04 [18]. Therefore, selecting only exactly reported p-values might cause an underrepresentation of .05 values, because p-values are more frequently rounded and reported as < .05 instead of exactly (e.g., *p* = .046). This limitation also applies to the original paper by Head et al. and is therefore a general, albeit minor, limitation to analyzing p-value distributions.

## Conclusion

Based on the results of this reanalysis, it can be concluded that the original evidence for widespread evidence of left-skew p-hacking [1] does not hold. Additionally, absence of evidence for left-skew p-hacking should not be interpreted as evidence for the absence of left-skew p-hacking. In other words, even though no evidence for left-skew p-hacking was found, this does not mean it does not occur at all — it only indicates that it does not occur so frequently such that the aggregate distribution of significant p-values in science becomes left-skewed.

## Acknowledgments

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

References

[1] Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of p-hacking in science. PLoS Biol 13: e1002106.

[2] Simonsohn U, Nelson LD, Simmons JP (2014) P-curve: A key to the file-drawer.

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

[4] Gerber A, Malhotra N, Dowling C, Doherty D (2010) Publication bias in two political behavior literatures. American Politics Research 38: 591-613.

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

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

[7] APA (1983) Publication manual of the American Psychological Association. Washington, DC: American Psychological Association, 3rd edition.

[8] APA (2001) Publication manual of the American Psychological Association. Washington, DC: American Psychological Association, 5th edition.

[9

[10] Baker M (2015) First results from psychology’s largest reproducibility test. Nature News.

[11] John LK, Loewenstein G, Prelec D (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological science 23: 524–532.

[12] Bakker M (2014) Flawed intuitions about power in psychological research. In: Good science, bad science: Questioning research practices in psychological research. pp. 109–120.

[13] Bakker M, van Dijk A, Wicherts JM (2012) The Rules of the Game Called Psychological Science. Perspectives on Psychological Science 7: 543–554.

[14] Cohen J (1962) The statistical power of abnormal social psychological research: A review. Journal of Abnormal and Social Psychology 65: 145–153.

[15] Sedlmeier P, Gigerenzer G (1989) Do studies of statistical power have an effect on the power of studies? Psychological Bulletin 105: 309–316.

[16] Moher D, Dulberg CS, Wells GA (1994) Statistical power, sample size, and their reporting in randomized controlled trials. JAMA: the journal of the American Medical Association 272: 122–124.

[17] Ioannidis JPA (2005) Why most published research findings are false. PLoS medicine 2: e124.

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

## Supporting Information

## [S1 File. Full reanalysis results per discipline.](https://github.com/chartgerink/2015head/raw/master/submission/round\%201\%20review/S1\%20Reanalysis\%20results\%20per\%20discipline.xlsx)