PLOS Biology

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

--Manuscript Draft--

Manuscript Number:	PBIOLOGY-D-15-01367R1
Full Title:	Reanalyzing Head et al. (2015): No widespread p-hacking after all?
Article Type:	Formal Comment
Corresponding Author:	Chris H.J. Hartgerink, MSc. Tilburg University Tilburg, NETHERLANDS
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Tilburg University
Corresponding Author's Secondary Institution:	
First Author:	Chris H.J. Hartgerink, MSc.
First Author Secondary Information:	
Order of Authors:	Chris H.J. Hartgerink, MSc.
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Competing Interest For yourself and on behalf of all the authors of this manuscript, please declare below any competing interests as described in the "PLoS Policy on Declaration and Evaluation of Competing Interests." You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work, acknowledging all financial support and any other relevant financial or competing interests. If no competing interests exist, enter: "The authors have declared that no competing interests exist." If you have competing interests to declare, please fill out the text box completing the following statement: "I have read the journal's policy and have the following conflicts"	The authors have declared that no competing interests exist.

* typeset	
Ethics Statement All research involving human participants must have been approved by the authors' institutional review board or equivalent committee(s) and that board must be named by the authors in the manuscript. For research involving human participants, informed consent must have been obtained (or the reason for lack of consent explained, e.g. the data were analyzed anonymously) and all clinical investigation must have been conducted according to the principles expressed in the Declaration of Helsinki. Authors should submit a statement from their ethics committee or institutional review board indicating the approval of the research. We also encourage authors to submit a sample of a patient consent form and may require submission of completed forms on particular occasions. All animal work must have been conducted according to relevant national and international guidelines. In accordance with the recommendations of the Weatherall report, "The use of non-human primates in research" we specifically require authors to include details of animal welfare and steps taken to ameliorate suffering in all work involving non-human primates. The relevant guidelines followed and the committee that approved the study should be identified in the ethics statement. Please enter your ethics statement below and place the same text at the beginning of the Methods section of your manuscript (with the subheading Ethics Statement). Enter "N/A" if you do not require an ethics statement.	N/A

July 08, 2015

Dear dr. Kousta,

With pleasure, I hereby resubmit my manuscript ‘Reanalyzing Head et al. (2015): No widespread p-hacking after all?’ for consideration in *PLOS: Biology*. Thank you for finding four reviewers and I thank the reviewers for their extensive comments and their transparency.

I feel that the reviews were highly constructive and that the comments were greatly helpful in improving the manuscript. All revisions are explicated in the manuscript file with tracked changes. Revisions in direct response to the reviewers are explicated in the response to the reviewers.

In your decision letter two aspects of the reviewers were highlighted, which are both extensively addressed in the responses and summarized here. In short, the two aspects regarded the invalid use of the p-curve method and the suggestion for sensitivity analyses across binwidths. My response to the former is that all p-curve analyses investigate p-value distributions, but not all p-value distributions are investigated with p-curve analyses. The current paper investigates a p-value distribution without applying the p-curve method, hence cannot improperly apply the p-curve method. The suggested sensitivity analyses were conducted and the results were consistent for different binwidths.

I look forward to your reply and hope you will find my revised manuscript an improvement.

Kind regards,

Chris H.J. Hartgerink

Tilburg University
Warandelaan 2, 5037AB, Tilburg
c.h.j.hartgerink@tilburguniversity.edu

Response to reviewers, *Reanalyzing Head et al. (2015) by CHJ Hartgerink*

Reviewer Notes:

Reviewer #1

Signed review: Uri Simonsohn

[Reviewer 1 remark #1]
This paper provides a critique of Head et al analyses of the distribution of all p-values from a large number of papers. The paper by Head et al. is helplessly flawed, so I am favorably inclined to a critique of it. But unfortunately the most important flaws in Head et al are not only not addressed in the critique, but worse, they also invalidate the critique's results.

Let me start by giving a sense of my reservations with Head et al. by copy pasting from a new paper we are writing about p-curve, the quote comes from a section covering how researchers may misuse p-curve.

QUOTE BEGINS<:
In Simonsohn, Nelson, and Simmons (2014) we explained how p-values should be selected from studies in order for p-curve analysis to be valid, e.g., writing "Most studies report multiple p values, but not all p values should be included in p-curve. Included p values must meet three criteria: (a) test the hypothesis of interest, (b) have a uniform distribution under the null, and (c) be statistically independent of other p values" (p.540). […]

The most extreme violation consists of selecting all p-values in an article. One example is by Head et al. (2015), who p-curved all p-values published in Open Access journals, simultaneously violating the three principles outlined above. It addition to asking an arguably meaningless question, "what's the evidential value of all tests, whether relevant or irrelevant, whether supportive or contradictive of the hypotheses of interest?" the analyses provide a statistically invalid answer. The error is surely not sinister in intent. Rather, p-curve, like any meta-analytic tool, needs to be carefully applied in order to be correctly interpreted.'
>QUOTE ENDS.

Probably the most serious problem with Head et al is that most p-values reported in papers are not of interest.

Some of these involve relationships that are obviously true and are reported to describe the data or as covariates. For example, studies examining the effect of something on life expectancy, will report the effect of age and gender on life expectancy. Both massively true effects with right-skewed p-curves but which are not of intrinsic interest to the researchers.

Some of these irrelevant p-values involve inexistent effects, for example when testing covariates across treatment vs control to ensure randomization worked. Because the true effect is zero, the expected p-curve is flat.

Most papers, then report a few relevant p-values (that could be p-hacked), but mostly irrelevant ones that are right-skewed on-average. This means that the p-curve of all p-values is not only meaningless,

and statistically invalid, but it is also biased away from p-hacking, because for p-hacking to detectable it would be need to be very intense, probably impossibly so.

While I was not, sadly, asked to review Head et al., all the limitation above apply to the new analyses in the critique as well.

The critique under review does not address the issues above, focusing instead on two narrow operationalization issues.

*****

[Response to reviewer 1 remark #1]

I thank the reviewer for this extensive analysis of p-value distributions. I agree that p-curve estimation as a meta-analytic technique of a specific effect should be done on systematically selected p-values, to prevent the selection of p-values irrelevant to the effect of interest. All p-curve estimation methods inspect p-value distributions, but not all analyses of p-value distributions are p-curve estimations, however.

This paper (and the original by Head et al.) investigates aggregate p-value distributions and not one specific effect. Hence, the p-curve procedure itself is not applied here whilst the objections pertain specifically to the p-curve method. An aggregate, heterogeneous p-value distribution is a mixture distribution of the uniform distribution under H_0 and a right-skew distribution under H_1. P-hacking behaviors could affect the p-value distribution, of which one example, optional stopping, creates a peak of p-values below .05 under the null (see also Lakens' 2014 reanalysis of Masicampo and Lalande). If, as Head et al. argue, optional stopping under the null or behaviors with similar effects occur on a large scale, the p-value distribution at the aggregate level could show such a peak below .05. In other words, if these highly specific behaviors occur very often in specific circumstances, such a peaked p-value distribution could be found. If this effect is found in such an aggregate p-value distribution, this is sufficient evidence for the presence of such specific behaviors at a large scale. As a consequence of these remarks by reviewer #1, I decided to add several paragraphs to these, and other, objections in the introduction, mirroring my response to the reviewer. This addition is made in the introduction, pages 1-2 lines 15-6 (across pages).

> *### ADDITION START ###*
> *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.*
> *### ADDITION END ###*

*****

[Reviewer 1 remark #2]
First, it shows that if instead of dropping results that in original research are imprecisely reported as "p<.05," as Head et al did, we replace them with "p=.05" then we see a vast excess mass at .05 which we could (incorrectly) interpret as evidence of p-hacking. Incorrectly because many of those p<.05 are, of course, not p=.05 so the replacement leads to an over-estimation of p=.05s. That a statistic known to be biased upwards has a high estimate is not diagnostic.

To be clear, dropping p<.05s, what Head et al did, is not a solution either, because many of those p<.05 may be p-hacked findings that are being dropped. The only solution is to re-compute p-values based on reported test statistics, as we do in our p-curve papers and online p-curve app (www.p-curve.com/app3)
*****
[Response to reviewer 1 remark #2]
The reviewer seems to indicate that p < .05 was replaced with p = .05 in the submitted MS; this was not the case and my apologies if this was unclear. All inexactly reported p-values (i.e., all p < ... or p > ...) were dropped, including p < .05. In other words, only exactly reported p-values were retained (i.e. p = ...). Head and colleagues used "p < .05" as selection criterion across only such exactly reported p-values; I merely extended the selection to "p ≤ .05". I agree that recalculating p-values is most optimal, but the original dataset did not contain sufficient information to do this. To ensure clarity in the manuscript, I rewrote the selection as follows on page 3 lines 21-23:

ADDITION START
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
ADDITION END

***

[Reviewer 1 remark #3]
Second, the critique proposes a new test comparing the proportion of .03875<p<.04, henceforth ".04s," to the proportion of .04875<p<.05, henceforth, ".05s." The result of this analysis is that papers report more .04s that .05s; the author goes on to conclude that "no evidence for p-hacking remains"

The problem alluded to earlier, that most p-values in papers aren't of interest, extends to this new analysis as well.
***
[Response to reviewer 1 remark #3]
See response to reviewer 1 remark #1, where I indicate that a peak below .05 is a sufficient, but not necessary condition for detecting p-hacking.
***

[Reviewer 1 remark #4]
To give a more concrete sense, a test powered to 90% will have 40% more .04s than 05s (see R Code below). Because many papers include p-values for many obvious relatinoshipos (say powered at 90%), even if all key p-values were p-hacked in all papers, the high proportion of .04s coming from the other tests would make it undetectable with the tests proposed in this paper.

There are other problems. If a study obtains p<<.05 for the key result, say p=.000001 it is common for researchers to show robustness (to brag) and show that controlling for X,Y and Z the effect is still p<.05. This will tend to further bias p-curve away from p-hacking if we include "all" reported p-values, because low p-values get counted multiple times. There is also the problem that, as we hose in our 2014 paper, p-hacking of true effects does not lead to left skew.

I will stop here, the whole enterprise seems hopeless to me.
*****
[Response remark #4]
I agree with the analysis provided by the reviewer. I am unfamiliar with the extent to which such 'bragging' behaviors occur though. I do concur that left skew (i.e., peak below .05) does not readily occur even when p-hacking is present (very specific conditions required, e.g., under the null hypothesis for optional stopping). This raises the question of what we actually know when we find no left skew. When left skew is found, this is a sufficient condition to conclude some specific type of p-hacking occurs. Because Head et al. initially indicate they did find this sufficient condition, the results of this reanalysis would invalidate that particular result. For the size of the claims made (i.e., p-hacking throughout the sciences) this connotation is, in my view, of value. I also refer to my response to remark #1 where I added a paragraph in the submitted MS on exactly this.
*****
[skip R code of Reviewer 1]

Reviewer #2

Megan Head et al. Please see comments in attachment.

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

The author raises some valid points and I think their comment is well thought through. I commend their effort to check the robustness of our findings and offer alternative analyses. Here I will detail some of the reasoning behind our methods and comment on why we believe our method is preferable to the "strong reanalysis" that the Dr. Hartgerink suggests.

[Reviewer 2 remark #1]

Our data selection criteria

The author states that we use four data selection criteria that require more justification. We do that here:

i) using papers with one DOI: we neglected to state this in our manuscript, but the reason we did this was that when we inspected papers that had more than one DOI they appeared to be collections of conference papers or abstracts, since these were not research articles we decided to exclude them. We did not systematically check that this was the case for a large number of papers, because we had no reason to believe that this exclusion criteria would bias our results in any particular direction, but rather would just reduce our sample size. Since statistical power was not an issue with our very large sample sizes this was not a major concern, and we thought it better to restrict our data to papers with only one

DOI. This supposition is supported by Dr. Hartgerink's reanalysis: the measured effect size is similar, but the p value associated with it is smaller due to the increase sample size.

ii) papers with non-zero authors: the reason we excluded these papers is exactly that described in the comment, that is, that they tend not to be original research papers.

iii) not including papers with p = 0.05: We reasoned that not all papers reporting p=0.05 regarded this result as significant, and given this we preferred to err on the side of being conservative. We had not seen the paper, Nuijten et al (2015), which shows that ~95% of 236 cases reporting p as exactly 0.05 as statistically significant. This result is good justification for including p = 0.05 in our dataset. However, the bins used in our analysis did not include 0.05, for an additional reason, namely the problems caused by authors rounding their p values to 2 decimal places (as mentioned by Hartgerink later in his comment).

iv) retaining only exact p-values: the reason for excluding p-values presented as p<0.05, was that it is impossible to know what they really were without recalculating them from test statistics, which is clearly impossible for the very large dataset obtained using our text mining approach. Dr Hartgerink seems to concur that this criterion was justified.

***

[Response to reviewer 2 remark #1]

I thank the reviewers for clarifying their data analytic choices. On points (ii) and (iv) the reviewers correctly state I agree with their judgment. On point (i) there seems to be a minor difference of opinion with respect to the DOI selection. Considering that, as the reviewers note, the results are similar, the conclusion seems that it does not really matter. I retained the selection change of point (i) in the text (see page 4 line 18-20 of the revised MS), to show that it does not matter for the sensitivity reanalysis and actually strengthens the case of the original authors. On point (iii) the reviewers seem to agree with the selection change (but hedge to disagreeing with the later analyses, which I respond to in remark #3).

***

[Reviewer 2 remark #2]

Sensitivity reanalysis

The author states that this reanalysis only changed the data selection criteria and not the actual analysis. When looking at the code provided on OSF it appears that the author has also altered the bins. This is an important issue and I think the author should make this clear in their comment. One of the reasons we chose the bins we did was to avoid problems arising when researchers round results to two decimal places (we go into this more below in response to Dr Hartgerink's strong analysis).

***

[Response to reviewer 2 remark #2]

I thank the reviewers for taking the time to inspect my analysis code. They state that I altered the bin selection in the sensitivity reanalysis, but I cannot find any such alteration in my code. I used the original code for the sensitivity reanalysis (lines 59-73 of the original `analysis.r` file as available on Dryad, which correspond to lines 84-98 in the `chjh analyses.r` file). The only changes made are data selection changes

as recognized in reviewer 2 remark #1, but these do not pertain to the bin selection. The bin selection changed only for the "strong reanalysis" of the previously submitted MS. This remark therefore seems incorrect.

*****

[Reviewer 2 remark #3]

Data analytic strategy and Strong reanalysis

We agree with Dr. Hartgerink that careful selection of the bins that are compared is of vital importance and we put a lot of thought into the appropriate bins to use before beginning our analyses. The issue of how to deal with inexact reporting (i.e. p< rather than p=) and rounding were major considerations when selecting our bins. Another major consideration was being able to detect p-hacking in the presence of strong evidential value (i.e. if most p-0values document tests where the true effect size is non-zero, the distribution of p values will show right skew, hindering our power to detect p-hacking, which tends to add left skew).

To avoid issues associated with two decimal reporting, the edges of our bins did not contain p values that could be exactly expressed in a number given to two decimal places; that is, the bins excluded numbers like 0.04. The bin ranges were: $0.04 < p < 0.045$ (lower bin), and $0.045 < p < 0.05$ (upper bin). Our choice of bins does mean excluding values of p=0.05 however, which the authors comment suggests makes our analysis more conservative.

In order to enable inclusion of p-values equal to 0.05 Dr. Hartgerink instead chooses to compare bins that both include p-values reported to two decimal places (lower bin: 0.03875-0.04, upper bin 0.04875-0.05). The choice of these bins raises two important issues: 1) It assumes that studies are equally likely to round to 0.04 as they are to 0.05. Given that 0.05 is the threshold of significance and 0.04 is not, we think it is reasonable to believe that rounding rules may be applied differently around 0.04 and 0.05, and thus that it is unwise to include the numbers 0.04 and 0.05 in a test for p hacking. For instance, p-values that are just under 0.05 may be more likely to be reported as <0.05 than rounded up to 0.05 (and thus disappear from our dataset), whereas p-values that are just under 0.04 are more likely to be rounded up than reported as <0.04. This bias in reporting practice would cause a dearth of p values in the upper bin, and hence mask evidence of p-hacking (this is likely one of the reasons that Dr. Hartgerink's test did not produce the same evidence for p hacking as ours). 2) The use of bins that are not directly next to each other makes it more difficult to detect p-hacking when there is evidential value (i.e. data in which there is a true effect). Evidential value leads to a strong right skew in the distribution of p-values. P-hacking leads to a left skew in p-values just below 0.05. When these two distributions are combined strong evidential value can mask p-hacking even if it is prevalent. To be sensitive to p-hacking in the face of strong evidential value, a test must include bins as close to 0.05 as possible. While this doesn't make the authors choice of bins wrong, it does make them less sensitive to p-hacking and it is not surprising that he did not find p-hacking using these bins.

Given the issues with assigning upper and lower bins for comparison outlined above, we believe our analysis is a better way to detect p-hacking than the one outlined in the comment.

Megan Head, Luke Holman, Rob Lanfear & Michael Jennions

***

[Response to reviewer 2 remark #3]
I thank the reviewers for their analysis of their own considerations and constructive comments on my analyses. The selection of .04 < p < .045 and .045 < p < .05 does indeed remove the second decimal from all analyses, but this analytic choice seems dependent on researchers interpreting only p < .05 as significant. The reviewers seemed to agree with changing the selection criterion to p ≤ .05 (see remark #1 by the reviewers), but apparently stick to their analyses choices under the p < .05 framework.

Their critique of more rounding occurring at .05 than at .04 seems plausible and I inspected a recent paper by Krawczyk (2015) in PLOS ONE to see whether this notion is supported by data. It seems there is some evidence (Fig 5 of Krawczyk, 2015; see here), but this difference is not large. Hence, this is a valid point which I added as a limitation, but I note this also affects the original paper by Head et al by eliminating those values which, if not rounded, would be included in their selection. This limitation reads "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." This addition is found on pages 6-7 lines 19-3 (across pages).

Additionally, the reviewers postulate the difficulty of detecting p-hacking when there is evidential value. This is more a remark of the method of detecting p-hacking via a peak of p-values than a remark of my bin selection. P-hacking only creates left skew when specific behaviors occur (e.g., optional stopping) in specific circumstances (e.g., under the null hypothesis). This was also the main comment of reviewer #1 to which I refer for my adjustments. I fully agree and note this also applies to the original paper.

Finally, the reviewers postulate that the inability to detect p-hacking is due to using non-adjacent bins in light of evidential value. In response to this, I ran the original analysis but included the second decimal to take into account reporting bias, i.e.,.04 ≤ p < .045 vs .045 < p ≤ .05. This also indicates no evidence for p-hacking, despite using adjacent bins (proportion final bin: .457). I added this connotation as a limitation on page 6 lines 14-19, which reads as follows:

"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."

***

Reviewer #3

The author presents an interesting reanalysis of the data from the paper previously published in PLoS Biology by Head et al (2015). I think the MS is well written and what the author presents as alternative analyses are worth publishing (especially noting about the issue about rounding. But I think the current conclusion seem to be too strong.

The author conducted what he calls "strong analysis" comparing different two bins from the bins used in Head et al. But I am not sure whether it is 'stronger' analysis without any further justification or evidence. The author is making an assumption here about the comparison of his bins being better than the original comparison. To me, this analysis is just an alternative way of testing p-hacking. So it is not really 'strong' enough to totally discount the original finding. Thus, I think the author will need to make his conclusion more moderate.

To me, both ways of testing p-hacking seem valid with some limitations (all depends on which assumptions are more right). Probably this new analysis is too conservative and the original too liberal. Without further data or evidence, I cannot really tell.
*****
[Overall response to reviewer #3]
I thank the reviewer for the compliments on my submission. It seems as if the term "strong reanalysis" is interpreted as meaning a stronger type of analysis. My apologies if this seemed as such; I intended it as a label of the type of reanalysis, not the value of the analysis itself. Because of the changes reviewer #4 suggested this differentiation is no longer made in the MS, leaving no room for potential misinterpretation.

Reviewer #4

[Reviewer 4 remark #1]
This submission points out a crucial flaw in the initial analysis from Head et al. However, the author should bring this issue out more clearly, by cutting irrelevant material and by emphasizing the main point. For instance, the crucial issue is that p-values suffer from "a reporting tendency at the second decimal place." At this stage the reader needs to be taken through Figure 1, and be told exactly why the bin closest to .05 is not representable. If the reader digests only the text, the main point can hardly be understood.
Before we get to the crucial flaw, the author confuses the readership by including largely irrelevant sections on "data analytic choices" and "sensitivity analysis". These two sections should be summarized in a footnote: the bone of contention is the re-analysis acknowledging the reporting tendency.

***
[Response to reviewer 4 remark #1]
I thank the reviewer for his critical point of view of the submitted MS. I note that PLOS BIO does not accept footnotes and I have therefore added a sentence in the running text on this, and have deleted the sections on 'data-analytic choices' and 'sensitivity analysis'. This sentence reads "Initial sensitivity analyses using the original analysis script strengthened original results after eliminating DOI selection and using $p \leq .05$ as selection criterion instead of $p < .05$" on page 4 lines 18-20. I very much like the idea of using Figure 1 to guide the reader through the thought-process and have now added the paragraphs below (pages 2-4, lines 17-3 across pages) and extended Figure 1 with an additional panel:

ADDITION START###

> *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. (i.e., $.04 < p < .045$ 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.
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.

ADDITION END###

*****

[Reviewer 4 remark #2]
So I completely agree that the reporting tendencies confound the analysis of Head et al. in a major way. We then move on to an alternative analysis. The alternative analysis focuses on bins near .05 versus .04. I am largely OK with that, even though, in the presence of a true effect and without any p-hacking, values near .04 should be more frequent than those near .05 (should this be acknowledged or corrected for?). The author than goes on and states "the binwidth is adjusted from .005 to .00125 for more precision and comparability with previous research". At this point, every reader will wonder – "oh, that's weird, I wonder what result the original binwidth would give?" In my opinion, the results need to be reported across a range of binwidths. Any single choice will lead to discussion and hide useful information. For instance, Head et al. might well argue that narrowing down the interval has lead to a decrease in power (which seems an unlikely interpretation given the data, but still).
***
[Response to reviewer 4 remark #2]
I agree with the reviewer that sensitivity analyses are a welcome addition and have now added these in the paper in Table 1. Reviewer #4 makes a similar note as reviewer #1 on how the p-value distribution, in case of a true effect and no data-peeking, would show more .04s than .05s. I fully agree and refer to my addition made in response to reviewer #1 remark #1 as a reply to this remark.
*****

[Reviewer 4 remark #3]
Let's move to the results of the re-analysis. First, it is highly confusing to denote the proportion by P, and test for significance using p. The notation should be changed. Second, I recommend a one-sided Bayesian proportion test in order to quantify evidence against the null.
***
[Response to reviewer 4 remark #3]
I altered the "P" to "Prop." to avoid this confusion. I very much welcome the suggestion for the Bayesian proportion test, and have incorporated these analyses alongside the frequentist proportion test to appeal to both audiences. I have also added a section on how to interpret the Bayes Factor:

ADDITION START###

> *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.*

ADDITION END###

***

[Reviewer 4 remark #4]
So, my major commendations are:

1. Cut the sections on "data analytic choices" and "sensitivity analysis";

2. Point out the existence of the reporting-tendency more clearly, and explicitly discuss its ramifications using a concrete example.

3. In the re-analysis, conduct a sensitivity analysis across bin-widths
***
[Response to reviewer 4 remark #4]
For point 1 and 2, see response to remark #1 (i.e., adjusted this; took up the concrete example as well); for point 3 see response to remark #2 (i.e., adjusted).
***

[Reviewer 4 remark #5]
Finally, I also recommend that the author discuss, briefly, the consequences of this result. For instance, both the paper from Head and the current reply hinge on the fact that p-hacking expresses itself in left-skewed distributions of p-values. I don't think that this is necessarily true. Some forms of p-hacking (conducting multiple analyses and reporting the most significant one) will masquerade as a true effect. This realization challenges the conclusion from Head et al.
***

[Response to reviewer 4 remark #5]
I agree with the reviewer and have tried to incorporate this severe connotation. See also reviewer #1 response to remark #4, where the same point was raised.
***

[Reviewer remark #6]
Also, the author's main conclusion is inconsistent with self-reports of p-hacking (John et al., 2012). It is also inconsistent with the low replication rates that are now observed across the board. This challenges both the conclusions from Head et al. and the one from the current reply. These inconsistencies need to be addressed explicitly.
From my perspective, this just shows that p-hacking is often indistinguishable from a real effect, making it impossible to detect by statistical means alone.

Eric-Jan Wagenmakers
***

[Response to reviewer 4 remark #6]
I very much enjoy the link to the John et al. paper and the low replication rates, but I disagree with the reviewer about it being inconsistent because these findings are not mutually exclusive (e.g., low replication rates are also possible through systemic low power, and admission rates pertain to whether researcher ever showed this behavior and not how often). I do however agree that it is important to point out and incorporate the following section in the MS on pages 5-6, lines 11-13 (across pages).

ADDITION START###

> *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.*

###ADDITION END###
***

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

C.H.J. Hartgerink[1]

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

*Corresponding author

E-mail: 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 H_0 and right-skew p-value distributions under the alternative hypothesis H_1. 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. (i.e., $.04 < p < .045$ 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.

Fig. 1. Histogram of p-values as selected in Head et al. ($.04 < p < .045$ versus $.045 < p < .05$; top) and the full p-value distribution $\leq .05$ (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 H_0: *Prop.* $\leq .5$. 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 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 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 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 $p \leq .05$ 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.* $= .417$, $p > .999$, $BF_{10} < .001$

for the Results sections and *Prop.* = .358, $p > .999$, $BF_{10} < .001$ 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
	BF_{10}	<.001	<.001
Binwidth = .005	(.035−.04)	6641	38537
	(.045−.05)	4485	30406
	Prop.	0.403	0.441
	p	>.999	>.999
	BF_{10}	<.001	<.001
Binwidth = .01	(.03−.04)	9885	58809
	(.04−.05)	7250	47755
	Prop.	0.423	0.448
	p	>.999	>.999
	BF_{10}	<.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 $.04 \leq p < .045$ versus $.045 < p \leq .05$. This analysis also yielded no evidence for left-skew p-hacking, $Prop. = .457$, $p > .999$, $BF_{10} < .001$. 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. Journal of Experimental Psychology: General 143: 534-47.

[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] Nuijten MB, Hartgerink CHJ, Van Assen MALM, Epskamp S, Wicherts JM (2015). The Prevalence of Statistical Reporting Errors in Psychology (1985-2013). URL `https://osf.io/e9qbp/`.

[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.

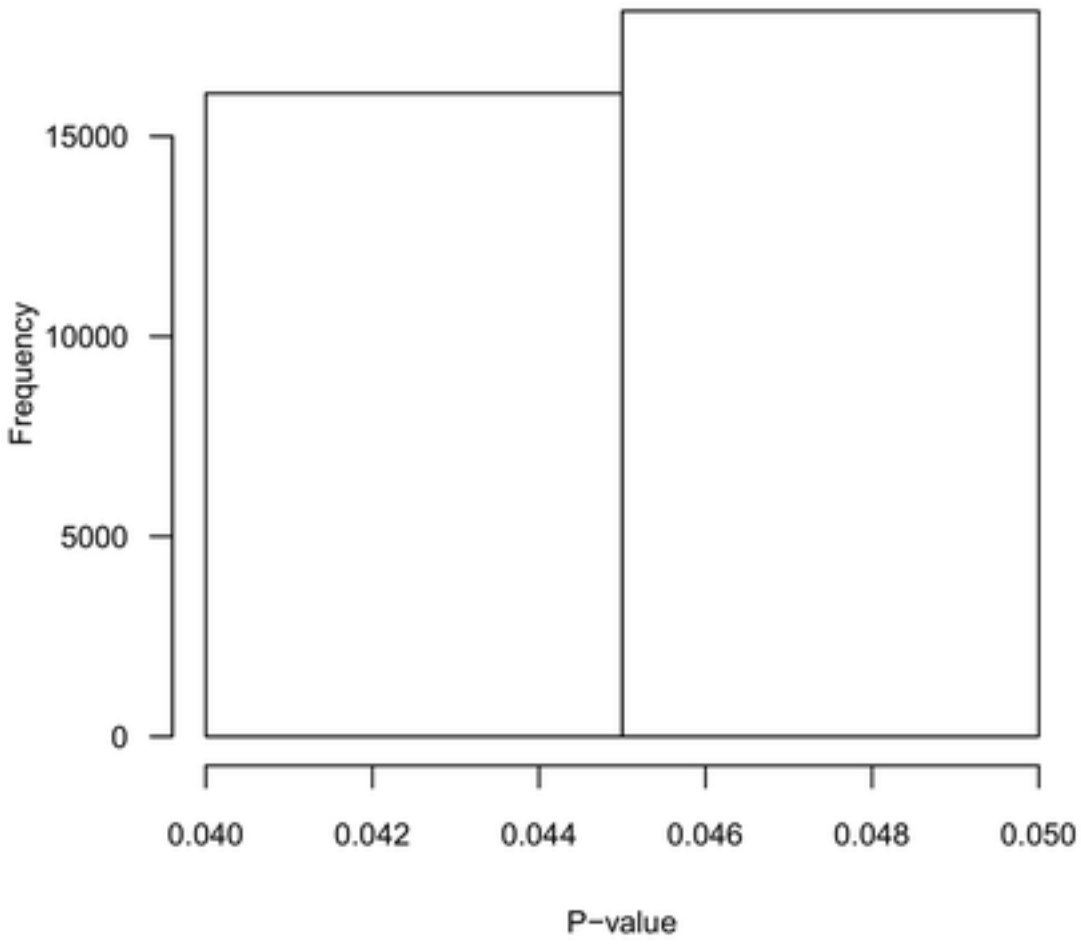
[18] Krawczyk M (2015) The search for significance: A few peculiarities in the distribution of P values in experimental psychology literature. PloS one 10: e0127872.

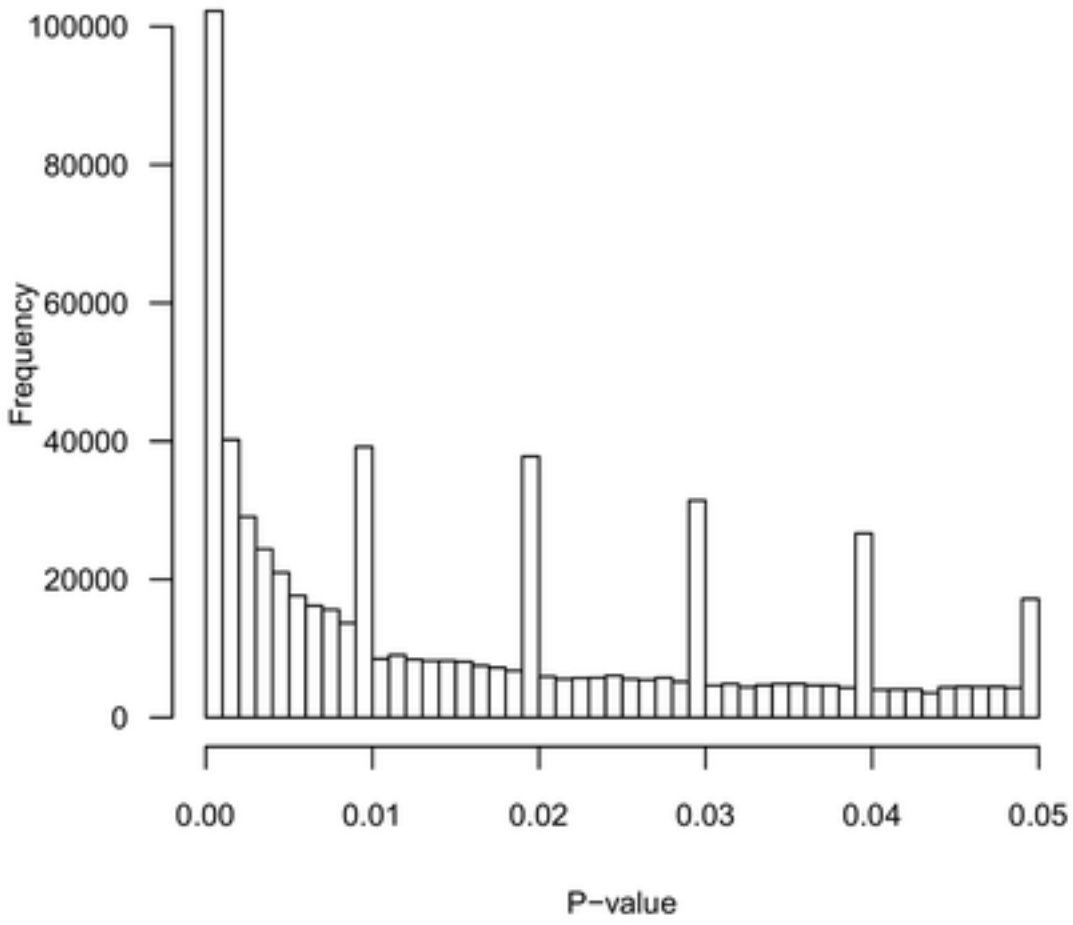
Supporting Information

S1 File. Full reanalysis results per discipline.

Fig 1

Click here to download Figure: Fig1.tif





S1 File

Click here to download Supporting Information: S1 Reanalysis results per discipline.xlsx

Tracked changes revised MS

Click here to download Other: 20150708 tracked changes revision.docx

Data Review URL
https://osf.io/sxafg/