June XX, 2015

Dear PLOS staff,

I hereby resubmit my manuscript ‘Reanalyzing Head et al. (2015): No widespread p-hacking after all?’ I thank the reviewers for their extensive comments and I would like to thank reviewers #1 and #4 for their transparency.

I feel that the reviews were highly constructive and their comments were greatly helpful in improving the manuscript. I hereby submit a revised version of the manuscript for consideration in the *PLOS: Biology* journal. The revisions are explicated in the file with tracked changes.

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

1) Reviewer #1 raises a novel criticism, both of your comment and the original paper: the suggestion that the p-curve analysis is invalid, because p-values were not appropriately selected (note that, on the basis of this, there should not have been any evidence for p-hacking in the original paper -- a rounding bias could be an explanation for that). Please consider carefully the implications of this point for the analyses you pursue here.

2) As Reviewer #4 points out, the first half of your analysis is almost irrelevant and should be removed in order to focus on the key issue you have identified (2nd digit rounding bias). With respect to the latter, as this reviewer recommends, it is essential to conduct a sensitivity analysis across bin-widths.

------------------------------------------------------------------------

Reviewer Notes:

Reviewer #1: Signed review: Uri Simonsohn

*[Reviewer 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.

*[Answer 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.) explicitly 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 this method. Such an aggregate, heterogeneous p-value distribution is a mixture distribution of the uniform distribution under H0 and a right-skew distribution under H­1. Questionable behaviors could affect the p-value distribution, of which one example, optional stopping, creates a peak of p-values around .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, only if these highly specific behaviors occur very often, such a peaked p-value distribution would 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 a paragraph pertaining to these objections.*

*[Reviewer 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](http://www.p-curve.com/app3))

*[Answer remark #2]*

*The reviewer seems to indicate that p < .05 was replaced with p = .05 in the 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. 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 extended this selection to “p ≤ .05”. I agree that recalculating p-values is the best solution, but the original dataset did not contain sufficient information to do this.*

*To ensure clarity in the paper, I \_\_\_*

*[Reviewer 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.

*[Answer remark #3]*

*See answer to reviewer #1; remark #1.*

*[Reviewer 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.

*[Answer remark #4]*

*I fully agree with the analysis provided by the reviewer. I am unfamiliar with the extent to which such ‘bragging’ occurs though. I do agree 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) raises the question of what we actually know when we find no left skew. However, when left skew is found, this is a sufficient condition to conclude these behaviors occur. Because Head et al. initially indicate they did find this sufficient condition, the results of this reanalysis would invalidate the results. For the size of the claims made (i.e., p-hacking throughout the sciences) this connotation is, in my view, important.*

*In order to clarify this point further, I added*

#R CODE computing share of .04s to .05s with 90% power #Critical t-values for "strong reanalaysis"

t.03875=abs(qt(.03875/2,df=38))

t.04 =abs(qt(.04/2, df=38))

t.04875=abs(qt(.04875/2,df=38))

t.05 =abs(qt(.05/2, df=38))

#Power of true studies

power=.9

#Effect size giving 90% power

library(pwr)

d=pwr.t.test(n=20,power=power)$d

#Noncentrality parameter for desired power

ncp=d\*sqrt(10)

#Expected ratio of (.03875-.04) vs (.04875-.05)

#Probability that p<.04

prop1=1-pt(t.04,df=38,ncp=ncp)

#probability that p<.03875

prop2=1-pt(t.03875,df=38,ncp=ncp)

#Probability that p<.05

prop3=1-pt(t.05,df=38,ncp=ncp)

#probability that p<.04875

prop4=1-pt(t.04875,df=38,ncp=ncp)

#Share of p-values in bin .03875-.04

bin.04=prop1-prop2

bin.05=prop3-prop4

#Ratio of .04 to .05

bin.04/bin.05

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.

**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.

[]

**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]*

*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; the different bin selection only starts in the “strong reanalysis”.*

**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

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 kind words and the compliment on how the paper is written. 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. I have rephrased “strong reanalysis into “full reanalysis” for clarity purposes, which I hope removes potential misinterpretation of the label.*

Reviewer #4:

[]

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.

[]

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).

[]

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]*

*Upon the reviewer’s remark, I agree that the notation is confusing. I altered the “P” to “\_\_” for clarity. 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.*

[]

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

[]

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.

[]

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