Thank you for submitting your formal comment on Head et al.'s recently published paper. We have now heard from three independent reviewers (please see below), as well as from the authors (please see attached). [Note that Reviewers #1 and #4 have opted to waive anonymity: Uri Simonsohn and Eric-Jan Wagenmakers.]

Although we cannot accept your Formal Comment in its present form, we would like to invite you to submit a much revised version for further consideration and re-review. In your revision, you must take into account and address all comments with modification of your manuscript, additional analyses and/or rebuttal in a detailed 'response to reviewers' document. I would particularly like to highlight the following points for your attention:

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.

We would be grateful if your revised manuscript by Jul 20 2015 11:59PM. Please email us ([plosbiology@plos.org](mailto:plosbiology@plos.org)) to discuss this if you have any questions or concerns. At this stage, your manuscript remains formally under active consideration at our journal. Therefore, please notify us by email if you do not wish to revise your manuscript for PLOS Biology and instead wish to pursue publication elsewhere, so that we may end consideration of the manuscript at PLOS Biology.

If you do still intend to submit a revised version of your manuscript, please go to <http://www.editorialmanager.com/pbiology/> and log in as an Author. Click the link labelled 'Submissions Needing Revision'. You will find your submission record there.

Thank you again for your submission to our journal. We hope that our editorial process has been constructive thus far, and we welcome your feedback at any time. Please don't hesitate to contact us if you have any questions or comments.

Sincerely,

Stavroula Kousta

Stavroula Kousta, Ph.D

Senior Editor

PLOS Biology

Carlyle House

Cambridge, CB4 3DN

4401223446969

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

Reviewer Notes:

Reviewer #1: Signed review: Uri Simonsohn

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.

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))

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.

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.

#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.

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.

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.

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