I've just reviewed the comments (not the MS too), because when I finished reading the comments it was clear that CH has not changed the analysis substantively or properly grasped my criticisms. His replies suggest an odd perspective on things - for example, he claims that a tendency for authors to round off some p-values but not others to make the results look nicer is not p-hacking (it clearly is!). He also thinks the literature review section in Head et al, which clears up many of the issues we are arguing about here (e.g. it does not contain rounded p values), is about how p-hacking affects meta-analysis (which it's not, really). CH also links to a neat paper by Krawczyk, which independently replicated Head's results, and solved the whole issue of biased rounding by re-calculating p-values from the test statistics - maybe CH could have a crack at that? CH also slightly tweaked the model I wrote, but I think the improvements are cosmetic and do not change the conclusion that Head's test is more 'sensitive' (by which I mean, less likely to return a false negative).

Response to reviewer 1, comment 2:

My original comment was:

We reasoned that researchers might be more or less likely to round off in different parts of the p-curve, which could add important biases to our analysis. For example, one might reasonably assume that researchers are less likely to round off p values in the vicinity of the "magic threshold", p = 0.05, because of the importance attached to it by researchers and reviewers. For example, rounding p = 0.0451 up to p = 0.05 would be a bad idea for researchers, because they will fall afoul of reviewers that think their result is "only just" significant, when actually p = 0.0451 is (widely but wrongly) considered to be good evidence against the null. By contrast, many scientists would not think twice about rounding p = 0.0351 up to p = 0.04, because it looks significant either way. Thus, the rounding would be biased, and the p=0.04 spike would be bigger than the p=0.05 spike (as we observe).

Re-reading it now, I would also add that an unscrupulous researcher might round 0.044 down to 0.04, making it look 'more significant' to people conditioned to regard p-values this way. However rounding down from 0.054 down to 0.05 is risky, because reviewers might ask to present the unrounded number, since it's so close to the magical p=0.05 and they'd want to know on what side of the imaginary line it falls. Alternatively, maybe lots of people round down 0.054 to 0.05 in order to get it over the line.

CH agrees that biased rounding occurs, but maintains that this is no reason not to include this potentially tainted data in the analysis:

I do not disagree with in essence --- I only disagree with the extent that it unequivocally shows there is only one way the choices that can be made in the data analysis strategy. I will repeat it here again: I am not trying to say my approach is correct and the reviewer's is false; I am merely stating there can be justifiable changes to the analysis and that this severely affects the results.

I find this a bit perplexing. CH's paper seemed to say very strongly that our approach was not correct, both via its content and the negative tone of voice. Also, if CH agrees with me that one needs to treat the rounded-off numbers with a lot of suspicion (because of evidence of widespread, biased rounding practices, which is clearly a form of p-hacking), then doesn't it follow that CH agrees that his analysis is less correct than Head et al's one? I feel that I presented several reasons why CH's new analysis is not justifiable, and I don't think his rebuttals of them were sufficient (as explained throughout this letter).

Incidentally, I thank CH for pointing me to that Krawczyk 2015 paper - not sure why I've never seen it before. Krawczyk also observed lots of suspicious activity near p=0.05, replicating Head et al nicely. By doing extra legwork compared to Head et al, Krawczyk showed that the suspicious 'p-value bump' just under 0.05 is partly due to biased reporting practices (namely, differentially picking a "<" sign vs a "=" sign in different parts of the p-curve, in a way that makes stuff look 'more significant'), but also other forms of p-hacking (compare figures 1 and 4 in Krawczyk 2015). Kudos.

Importantly though, I think CH has misread the Krawczyk 2015 paper. CH says that Figure 5 in Krawczyk allows one to estimate the rates at which authors round off to 0.04 vs rounding off to 0.05, which is a key parameter need for the models that CH and I did, as well as my general idea that rounding is biased and can stuff up CH's test. But that's not what Figure 5 looks at - Figure 5 is about reporting p-values with an "=" sign versus a "<" sign, and cannot tell us anything about rounding (CH's mistake is understable because the writing in Krawczyk is not very clear). However, Krawczyk 2015's Figure 1 (when compared to Figure 4) does show that people are about 3 times more likely to write "p = 0.05" than they are to write "p = 0.04", due to rounding. This is the opposite of what was found in Head et al - I wonder why! The only thing I can think of is that Head et al contained all sorts of science, while Krawczyk's paper is about psychology only, and there might be 'cultural' differences between fields in the way that people try to cheat by rounding p-values (maybe psychologists love to round down to p=0.05 and hope nobody notices). Nevertheless, Krawczyk's paper confirms my verbal argument (plus the data in Head et al) that people display very different rounding practices near p = 0.04 vs p = 0.05, which means that all tests for p-hacking should either A) exclude these two values (as in Head et al), or B) recalculate all the p-values from the test statistics to avoid rounding-related biases (as in Krawczyk's paper). This is one of the main reasons why I believe that CH's new test is not as trustworthy as ours.

Response to reviewer 1, comment 3:

Here CH comments on my new model, and says:

More specifically, I think three things are missing from the provided model: (1) a realistic p-value simulation, (2) how effect sizes alter the results, and (3) the P1 and P2 applied are unrealistic.

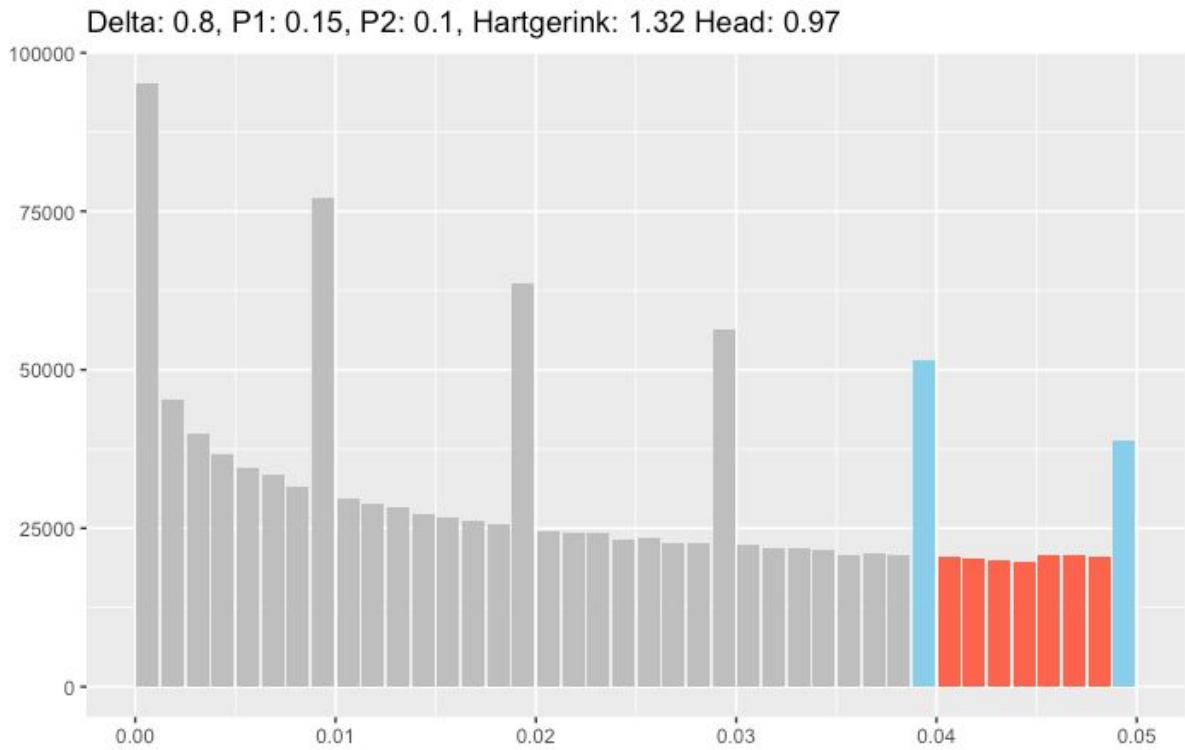
I agree with 1, but it does not invalidate my point. I just used an exponential distribution originally so that I could get the model done quickly (this is only a review after all!), and graphically illustrate my point that Hartgerink's method is more conservative assuming that people display biased rounding practices of the type that I hypothesised and modelled. Also, if I were to nitpick I'd say that CH's method for generating the p-curve is also not realistic (e.g. because it assumes every study has exactly the same effect size, when really there is a distribution of true effects in Head et al that probably has some sort of right-skewed distribution). But I don't think realism is necessary - the important thing for my toy model is that the curve is right-skewed and one can control the degree of skew using a parameter in the model - any curve that looks roughly like this is fine to illustrate my point that CH's test is biased under my assumption of biased rounding near 0.04 and 0.05, and the test gets more biased the stronger the right skew.

Regarding point 2: Some of the panels in Hartgerink's new graph look at cases where the effect across all studies is low or zero, which is not realistic and does not reproduce a curve that looks like the one we observe in Head et al (or the independent one from Krawczyk 2015). So, it doesn't matter if the kind of biases I talk about are weaker for hypothetical distributions of effect sizes that do not exist in the real data we are studying.

Point 3 is true: I only chose such big differences between P1 and P2 to highlight my point. I agree that having P1 being only be a little bigger than P2 is probably the most realistic, which still results in CH's method being more biased than our method (but not *hugely* more so), both in my model and in his revised version which has a better method for making the fake p-values.

My other concern about CH's new model is that he has chosen values of P1 and P2 based on his misreading of the Krawczyk paper (see my comments above). One should get P1 and P2 by looking at Krawczyk Figure 1: to my eyes, it looks like a 2.5-fold difference, not a 1.25-fold difference as CH's new model assumes. Actually, it would be better to get this information from the Head dataset, since that's what we are studying here. I would guess the difference is about 1.5-fold, from eyeballing the spikes in Figure 1 of CH's submission.

So, let's run CH's new model with a 1.5-fold difference in rounding, and a 'delta' parameter that makes the curve look like the real-world curve:



The output tells us that Hartgerink's test is very conservative: even under a no-p-hacking scenario (except for the biased rounding practices), the lower bin contains 1.32x more p-values, so like I said before the odds are stacked against CH's test detecting p-hacking. For the right-hand bin to be larger than the left one, you'd need really strong and prevalent p-hacking. Here, Head's test just about picks up the bias rounding practice (note the little red bump), because a lot of the p-values between 0.04 and 0.045 have moved left into the left blue bin. Head's test would also be comparatively more sensitive to p-hacking of the normal kind because the upper bin is not intrinsically higher than the lower one, as in CH's test.

Surprisingly, CH's review says:

Note that these models do NOT include p-hacking behaviors and are simply p-value distributions with rounding behaviors.

Reading this was a surprise to me - biasing whether or not you round off a number based on whether it is near or far from p=0.05, in order to make it look "more significant", is absolutely a form of p-hacking! I think anyone would agree with that. One main reason that some journals say "authors must give all p-values to 3 decimal places" is to stop people from misleading readers like this.

CH then writes: *"In other words, I disagree with the reviewer that my method is less sensitive, because these models cannot test sensitivity, only specificity."* CH also wrote *"any conclusions about the sensitivity of either test cannot be made based on these models, because p-hacking is not simulated in these models."*

The models CAN test sensitivity, as CH's model and mine clearly show. Even under the null hypothesis of no p-hacking, the p=0.04 peak is always bigger than the p=0.05 one if the

effect size is non-zero, and can get even bigger when people round off differentially near these two peaks. So, “Caliper tests” which test the null hypothesis that the bins are of equal size are highly conservative, therefore they are insensitive (i.e. prone to false negatives), because even without p-hacking we don’t expect the bins to be equal.

Response to Reviewer 1 comment 6

CH writes: *I am merely trying to provide an alternative way of looking at the data; given the large implications of the findings from the Head paper I wanted to be sure the findings were robust. I am not trying to say that the Head paper is correct or incorrect or that my approach is better, but simply that it is not as unequivocal as it was presented. It is up to the readers to decide what they find more convincing.*

I’m surprised to hear you say you are not trying to say Head et al was incorrect, because you seem to spend a lot of time saying just that. Also, while it is great for readers to be able to think about new analyses, those analyses should hold up to scrutiny too - I think I have clearly shown that the new one is substantially flawed, and your new model backs up my point. It’s also up to PeerJ to decide what they consider a sufficient advance for publication. Given that the new test is demonstrably worse than the original, is this paper enough of an advance to be worthwhile? The reanalysis makes up essentially all the content in the paper, so I don’t think it adds a lot (and indeed it contains considerable misinformation - I feel like I am forever adding small corrections).

Response to Reviewer 1 comment 12

When I pointed out that CH ignores our non-text-mined data, which also found evidence of p-hacking, CH wrote: *The other part of the Head paper pertains to how p-hacking affects meta-analyses and therefore is different from the extent (as is clearly delineated in the paper headings). Given that the reanalysis pertains only to the extent of p-hacking, I do not include this in the reanalysis.*

This isn’t really accurate. The non-text-mined data in Head et al served to address one big shortcoming of our text mining data. Simonsohn et al’s original p-hacking work was clear that one should build a p-curve using tests from papers’ main hypotheses, and not include “junk” tests (e.g. a K-S test to see if the data are normal), or pool tests from totally different fields of research, etc, because this would make tests for p-hacking more conservative (since nobody would bother to p-hack the secondary results). So, we collected p-values from the literature by hand - the meta-analyses were just used as a guide, for various reasons that Head et al explained. That section is not about “*how p-hacking affects meta-analyses”* - it is an independent p-value collection that was hand-collected, that finds a bump just as in the main dataset, and it also avoids the issues to do with rounding that is central to CH’s present submission. I think it’s totally worth pointing out in your MS, as it undermines your paper’s main message, which seems to be “Head et al didn’t really find any p-hacking”.

Response to Reviewer 1 comment 13

Regarding the critique of Head et al by Simonsohn and colleagues that you quoted: *"The most extreme violation consists of selecting all p-values in an article. One example is by Head, Holman, Lanfear, Kahn, and Jennions (2015), who p-curved all p- values published in Open Access journals. The paper asks an arguably meaningless question— "**What is the evidential value of all tests**, whether relevant or irrelevant, whether supportive or unsupportive of the hypotheses of interest?" and provides a statistically invalid answer."*

I have now found and read that paper, and actually the part you quote is about the evidential value section of Head et al, not the p-hacking section (note the bold I added). I concur more or less with Simonsohn et al's critique in this quote, but it does not dampen the p-hacking finding of our paper, which I believe CH would agree with. The fact that we included "irrelevant" tests in our p-curve just makes it less sensitive to detect p-hacking, but it could not falsely generate the bump that we see. Right?

Note also that this quote harkens back to our meta-analysis - we did the meta-analysis in anticipation of this criticism, because the meta-analysis part does not chuck unrelated p-values into the same stats blender like the text mining does (which to be fair, we did spell out in Head et al).

I fail to understand what part of these lines make no sense, as the reviewer does not specify his issues with it.

I think the original version of your paper had a non-grammatical, confusing statement about epistemology. I fail to understand why you failed to understand what I didn't understand :)

Response to Reviewer 1 comment 24

In this comment, I wrote:

This is the crux of why your re-analysis is wrong. You later say that "This altered bin selection takes such a reporting tendency into account and consequently includes the information available in these data." I would argue that your test does not take anything into account - you just throw the tainted data into the test, and let them confound the results. If I am wrong, you need to explain what is 'being taken into account' at least once in your paper.

CH's reply was: *See Reviewer 1 comment 20.*

That comment hardly addresses mine! Your revision still involved throwing data into your analysis which you and I and Krawczyk have all shown cannot be trusted, because it is rounded off in a way that masks p-hacking, as demonstrated by my model (which I only did because CH was not getting it when I spelled it out verbally). Comment 20 also does not explain how CH's analysis "Accounts for" anything. What do you think it accounts for, and how?

Response to Reviewer 1 comment 26

Here, CH writes: *Additionally, the reviewer mentions "The odds are thus stacked against finding a significant result in a test for p- hacking. When you find one (as we did, in both our text mining and literature review), it implies p-hacking is pretty darn common and/or strong." It is a logical fallacy to consider that if evidence is improbable and you find evidence, that the evidence then must be strong. As shown in my response to Reviewer 1 comment 4, it might very well be false positive*

Actually I was claiming that Head's test is very likely to return a false negative because it uses an incorrect null hypothesis (since the true null cannot be known), unless p-hacking is strong, in which case the chance of a false negative is low. Therefore, given that we observed an effect, the more likely (but not the only) explanation is that the effect of p-hacking is strong. Right? I don't think I'm making a logical fallacy here.

Also, you didn't show that our result "might very well be a false positive". You showed that if people do biased rounding of their p-values to make their results look nicer, then you get a bump just like the one we observed. That sort of practice is definitely p-hacking, by anyone's definition, so it wouldn't be a false positive.