**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).

**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