Dear Editor Vazire,

We would like to kindly thank you for organizing the review of our manuscript and your own extensive comments. We appreciate the constructive comments and provide a point-by-point reply to both below. We hereby submit our revised manuscript.

The main points of your comments seem to revolve around (1) the interpretation of (non)significance, (2) an overstatement of the RP:P reanalysis' findings, and (3) the effect of p-hacking in relation to false negatives. First, we do not regard the interpretation of significance explicitly in this paper, given that it has been inspected before by Hoekstra et al. (2008) and goes beyond the specific scope of inspecting evidence for false negatives in nonsignificant results. Second, the RP:P reanalysis indeed should be presented more nuanced, which we attempt. Third,

The main points of the reviewer's comments revolved around (1) the use of ICC for non-normal data, (2) some questions and remarks on how the results can and should be interpreted. The interdependency of p-values seems not to be a problem for p-values both theoretically and based on the data. We replied to the questions, and have made some revisions to explicate several of the underlying ideas in the revised manuscript.

Yours sincerely,

Chris Hartgerink, also on behalf of Jelte Wicherts and Marcel van Assen

# Editor comments

Dear Mr Chris H.J. Hartgerink,

Thank you for your submission to Collabra. I was able to get one reviewer with expertise in quantitative methods, and this reviewer did an outstanding job. I also independently read the manuscript before consulting this review. Both of us found much to like in your manuscript, and identified some issues that need to be addressed. Thus, I invite you to submit a revised version of this paper for further consideration at Collabra.

The reviewer did an outstanding job of articulating their concerns with the paper and I will not summarize all of their points in my letter. I would like you to address all of their concerns in the letter of response and if possible in the revised paper. Below I will add a few points of my own.

First, your arguments rest on the assumption that authors and/or readers are interpreting the non-significant results in the papers as evidence for the null. There are several alternative possibilities. One is that some of these null results may actually be interpreted as evidence against the null, particularly those that are marginally significant. Second, authors and readers may sometimes acknowledge that their non-significant results are inconclusive and may explicitly resist treating them as evidence for the null. Third, they may be ignored altogether (though I admit this is less likely since you are not using non-significant results that are only reported in tables, and only using ones that are reported in APA style in the text). This is not necessarily a problem for all of your points, but it does matter for the more conceptual argument about the potential harm that false negatives do to our field. Thus, I would caution you against assuming that non-significant results are always interpreted as evidence for the null. You may also consider excluding results with p-values between .05 and .10 from your analyses, since these are especially likely to have been interpreted as evidence against the null. I don’t have a strong feeling about this last suggestion – what’s more important to me is that you avoid making a strong assumption about how authors and readers are interpreting these non-significant findings (except in cases where you coded this, as perhaps in application 2).

## Reply to Editor Comment 1

As we noted in Application 2, determining the interpretation of a finding is rather difficult. Given previous investigation, it appears to be quite common that nonsignificant results are interpreted as that the null is true (see Hoekstra et al. 2008 reference in the manuscript). However, we do not regard the interpretation of the results themselves explicitly throughout Application 1, given that this falls outside the research question of "is there evidence for false negatives in nonsignificant results?". We tried to make this more explicit in the revised manuscript.

As such, we do not assume "that non-significant results are always interpreted as evidence for the null", but want to show to the reader that if they identify with doing so, they should beware because of these results.

We reran the analyses with alpha = .10 as suggested, and concluded that albeit different, the notion that there is evidence for >= 1 FN in the papers remains; the 66.7% changes to 42.7% after excluding 2009 papers that have no nonsignificant results when the alpha is shifted to .10. This is of course also the effect of lower power to detect FN by decreasing the amount of nonsignificant p-values analyzed. We do not present these results in the revised manuscript, but include the code in our OSF project to do so.

Regarding the analyses of RP:P, you make an important point about the fact that many of the non-significant results are inconclusive, and you also acknowledge that the authors of the RP:P themselves point this out. However, I think it is somewhat of a straw man to criticize the RP:P for not having enough power to differentiate between a null result and a true effect of r = .10. I don’t think many people think that the RP:P was intended to have that kind of power/precision. Moreover, your own test has very low precision for testing this (since the confidence interval goes all the way from 0 to 100% of non-significant findings being potential false negatives). Thus, I would suggest focusing more on the analyses examining the possibility that some of the RP:P non-significant findings are actually false negatives of medium or large effects. Here, you can rule out more possibilities (i.e., you can say that your analyses suggest there are fewer than 21 medium effects that were non-significant in the RP:P replications, etc.). Perhaps this is my bias showing through, but I also thought that a few sentences in this section were overstatements. Specifically, you write that the RP:P results “say hardly anything about whether there are truly no effects” (p. 19) – I agree that this statement is correct, but I think it places a lot of emphasis on distinguishing between null and very small effects, so holds the RP:P to a very high standard. That’s fair enough, but it may not be clear to readers that this is the standard you are using. Similarly, you also write “any conclusions on the validity of individual effects based on “failed” replications, as determined by statistical significance, is unwarranted and irresponsible.” First, the word “irresponsible” implies that this is a common interpretation (i.e., that people are doing this irresponsible thing) – is it? Second, I am not familiar with every single result in the RP:P, but is it really the case that none of the individual results were strong enough to warrant some kind of conclusion on the validity of that individual effect? I could imagine that some of the studies had sufficient precision to draw some conclusion based on a non-significant result. I am nitpicking here, and I will admit that I am probably more sensitive than the average reader to statements that could easily be misinterpreted, so feel free to take these suggestions or leave them.

## Reply to Editor Comment 2

Before responding, we would like to disclose that this analysis was done by CHJH and MvA, who both participated in the aggregate analyses of RP:P and actually conducted this analysis for a reply to an RP:P comment as part of the RP:P team (this comment remained unpublished).

We agree that the power of the Fisher test to detect false negatives in specific studies is low and our analysis only wants to provide a way of viewing the RP:P data a bit more nuanced than how many people have interpreted them ("only 36% of psych studies replicate!"). RP:P might also have false negatives, and this analysis shows it can be any combination of zero, small, medium, and large effects in the population that give rise to this rate of reproducibility.

Nonetheless, we agree that our analysis might be confounded with assessing individual studies. We have incorporated several references in the revised manuscript that do try to tackle this to clearly distance the analysis in this paper.

On page 21, you discuss “potential explanations” for the lack of change in power/sample size in the field over time. One piece missing from your explanation, in my opinion, is p-hacking. P-hacking lets us get significant results at a high rate even when our sample sizes suggest that our studies should be underpowered. Likewise, in the paragraph beginning “Reducing the emphasis on binary”, you also do not mention reducing p-hacking as another route to improving the situation. Transparency/disclosure requirements would go a long way to preventing people from being able to p-hack their way to significant results using small samples. If they were no longer able to do that, they would be required to get larger samples in order to detect the phenomena they’re studying. Again, feel free to take or leave this suggestion, but it stood out to me as an important missing piece of the puzzle in this section.

## Reply to Editor Comment 3

We agree with the reasoning of the editor that power goes towards 1 when researchers start to p-hack. However, to analyze the relation between p-hacking and power goes beyond the scope of the paper, given that there are many parameters that can vary (see also Hartgerink et al. 2016 for difficulties of modeling p-hacking behaviors and their effects on p-value distributions).

We tried to incorporate a constructive framing of the problem of false negatives, and we regard p-hacking primarily as the result of the emphasis on binary decisions instead of the other way around (just like publication bias). We incorporated a few sentences on this in the revision, given that this is one potential explanation as well.

Smaller points: -On page 10, you write “because effect size […] are typically overestimated population effect sizes” – is this true even for non-focal tests?

## Reply to Editor Comment 4

Yes. It is similar to how Hedges' g corrects for small sample bias in Cohen's d, regardless of whether it is focal or not (one can also think of omega-squared in ANOVA's).

-On page 13, you write the mean effect is r = .257 in 1985 and r = .187 in 2013 – what are these numbers referring to?

## Reply to Editor Comment 5

These numbers refer to the correlation effect size.

-Figure 4 – I wonder if this would be more useful as a percentage of all papers, not as a percentage of papers that report a non-significant result? Both are interesting, I’m not sure which one readers will find more interesting.

## Reply to Editor Comment 6

The results in Figure 4 only talk about the papers that report nonsignificant results, so for parsimony of interpretation we prefer to retain this format (we discussed this during the drafting as well).

-Figure 5 – how confident are you that degrees of freedom are a good proxy for N?

## Reply to Editor Comment 7

Degrees of freedom is equal to the sample size minus a constant (e.g., sample size minus the number of groups in ANOVA/t-tests), so we are confident in using this as a proxy.

-What does the Fisher test mean for the significant results in the gender analysis? Does the result just mean that at least one of those is a true non-null effect? Is this interesting?

## Reply to Editor Comment 8

It indicates there is evidence for >= 1 true effect in the significant gender results. This is interesting, because apparently there is little variance in whether there is evidential value for significant and nonsignificant results, yet there are differences in how these results are interpreted (in Application 2 we did actually code this).

-It wasn’t clear to me why you didn’t report, in applications 2 and 3, the percent of non-significant results that your analyses suggest are false negatives, as you did in application 1. In application 3, I’m guessing this is because the confidence interval around this point estimate is so wide that the point estimate would be misleading – is this correct? Is this also the explanation for why you didn’t report this statistic in application 2?

## Reply to Editor Comment 9

In Applications 2 and 3 this is not possible. In Application 1 we analyze 6,951 sets of nonsignificant values (i.e., 6,951 papers), but in Applications 2 and 3 only one set of nonsignificant results.

I’ll end with one last general point that is more a reflection than a suggestion for you. Please feel free to ignore it completely – I don’t like it when editors act like they are a co-author on the paper, and this is definitely a case where my disagreement is a matter of opinion rather than fact, so you have no obligation whatsoever to take my advice. Instead, consider it a sample of N = 1 of how some readers may interpret your argument.

I found myself disagreeing with your overall conclusion that the emphasis on false positives is “unwarranted” (abstract, p. 3). In my view, your data do not show that false negatives are a big problem, because your data do not speak to how non-significant results are interpreted in the papers themselves, nor what impact they have on subsequent research (i.e., do they deter others from pursuing the same question?). Many of the arguments that we should pay more attention to false negatives rest on these assumptions. Moreover, as you state yourself, the vast majority of focal tests in published psychology papers are statistically significant, which suggests that the negative results are likely non-focal tests. This puts more burden on those who claim we should be paying more attention to false negatives to demonstrate that these negative results actually have an impact on research decisions. Moreover, it raises the possibility that researchers are not “neglecting effects due to a lack of statistical significance” (abstract), but neglecting them because they don’t care about them. If they did care about those effects, they may very well have invested more effort into getting them to reach the threshold for significance (either through increasing sample size, reducing measurement error, or p-hacking).

## Reply to Editor Comment 10

Our wording was a bit too strong; we meant to say that there is an unbalanced emphasis on false positives (adjusted in revisions). We do think that regarding this as a problem in the published literature is too narrow; after all, the unpublished literature suffers from worse power problems most likely. However, we cannot inspect the problem of false negatives in the unpublished literature reliably (for Applications 1 and 2).

We try to show in this paper that there is a need to regard the potential of false negatives alongside (!) false positives, and provide a way to look for them. We hope that these findings will contribute to a shift away from binary decision making in one study, which has caused so many problems for the literature already.

In summary, the reviewer and I found much to like about your paper, and also had suggestions for improving your manuscript. I look forward to receiving your revision.

# Reviewer B

1. General comments and summary of recommendation Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:

* Are the methodologies used appropriate?
* Are any methodological weaknesses addressed?
* Is all statistical analysis sound?
* Does the conclusion (if present) reflect the argument, is it supported by data/facts?
* Is the article logically structured, succinct, and does the argument flow coherently?
* Are the references adequate and appropriate?: This article attempts to estimate the proportion of False Negative (FN) results in the published psychological literature. Overall, this is a well written article. A few minor comments/thoughts about the first analysis (since the methods in the other two analyses were similar/the same).

In the first analysis, the authors examined a very large number of non-significant results from eight psychology journals. The results were automatically (using an R package statcheck) extracted from over 14 thousand articles. There were multiple results extracted per paper; in fact, it seems that on average 3.5 non-significant results per paper were extracted (Table 3). The assumption of independence required for the application of the Fisher test (which assumes the p-values values are uniformly distributed) is therefore potentially violated. The authors deal with this by computing the ICC for non-significant results, and report it to be .001, which they suggest indicates independence of p-values within a paper. I would like to see a bit more discussion here as to whether this test is sufficient to dismiss the violation of the independence assumption as non-consequential in most/many of these articles. Are there are other references that have applied the ICC to p-values? I’m unfamiliar with this application. For instance, data on which ICC is computed are typically normally distributed. From a more practical standpoint, if one looked at a few articles at random, does the assumption that the same data are not used for the non-focal test seem reasonable? If an average number of studies per paper is 3, and each study reports a gender analysis that comes out non-significant, I believe in the independence of the resulting p-values. But I can imagine in many articles other scenarios are at play and this is not the case. What would be the impact on the analyses?

## Reply to Reviewer Comment 1

Due to the reviewer's comment, we went back to the data and inspected the ICC based on the log-odds of the transformed p-values, which are more normally distributed than the transformed p-values. The resulting ICC is 0.00002. We are unaware of any specific ICC applications based on p-values.

Anecdotally: our recollection is that in the RP:P project there was severe disagreement about experts trying to assess what the focal effects in a paper were. Additionally, how large would the articles inspected have to be to come to a reliable assessment of whether the results throughout the entire dataset are in fact independent? If only 10 papers out of 6,951 are inspected, making inferences for the entire set is still uncertain and does not provide any more certainty for the results than not manually inspecting them.

We also think that the impact of dependency is not a major problem due to the uniform distribution of p-values under H0 (independence) and the major variation in p-values that results even if H1 is true (see also Cumming's dance of the p-values to that respect).

The argument is that non-significant results reported in papers tend to be non-focal, and therefore are not p-hacked. The focal results in each paper, however, are either p-hacked or otherwise selected for significance (publication bias). This may be a bit far fetched but I wondered about whether selecting on main results can affect the distribution of the non-focal results in some way. It may be an interesting thought exercise. Related, is there any way in which publication bias/p-hacking can select for non-significant non-focal results? For instance, perhaps some journals are more likely to publish a “clean” set of results, where the main effects are found but the interactions with gender etc. are not significant.

## Reply to Reviewer Comment 2

We discussed reverse p-hacking (or what we call q-hacking) as the reviewer describes. This was one of the reasons why we started Application 2. However, given the small cell sizes we were unable to actually say anything about this.

We do not feel confident to state anything about the focality of the nonsignificant results, because this requires direct interpretation and is rather unclear (see also reply above).

I need a clarification on the meaning of the k=1 line in Table 4. I can only understand the test if more than one result per paper exists. Possibly I missed something. If the results for k=1 are indeed meaningful, can these values be related to an estimate of power to detect an average non-focal result? (i.e., use a different denominator to get the percentage). It may help the reader to place them in context. A comparison between average power for non-focal effects and what we know about average power for focal effects (e.g., Cohen) could be informative. A priori, as we expect many non-focal effects to be small or zero (so I’m using the word “power” here more loosely to refer to rates of non-rejection of H0), this number should be way lower than the average power for focal effects, which is already dangerously low in psychology, as we know.

## Reply to Reviewer Comment 3

When k=1, the Fisher method is another way of testing whether the result deviates from a null effect but with the prior information that we already know the result is nonsignificant. This gives more evidential weight to values just above alpha, given that if there is truly no effect the likelihood of them occurring is higher under H1 than H0.

The later rows in Table 4, as well as the average row, are a bit misleading because it is unfair to expect a statistical test to perform perfectly—when k>10, is it really reasonable to expect that not a single non-significant p-value in any article would not be a false negative? It makes it look like the problem with FNs is very bad (e.g., for JPSP, the rate is 94% with k>20). I would caution the reader against a pessimistic interpretation of the numbers in the high k rows, or in the average row (which is almost too misleading to be computed).

## Reply to Reviewer Comment 4

When more nonsignificant results are reported and these are not uniformly distributed, the likelihood does increase that there is >= 1 false negative. Of course, the test itself can have meta-decision errors, but with larger k the power to detect also becomes higher. Combined with dichotomous interpretation of p-values, we consider this a realistic interpretation and not a pessimistic one. A paper reporting only significant results can be just as unbelievable as one reporting only nonsignificant results. Similarly, if (e.g.) a JPSP paper reports more than 20 nonsignificant results, it is very likely that there is a false negative (the opposite of Francis' 'too good to be true').

One could also ask, somewhat cynically, whether we really learn anything from a paper that says that, on average when we examine false negative results reported in a paper (with the average number of them being 3.5), that at least one of them is wrong. This seems like something we know a priori. What would be your response to this? Disclaimer: I do find the paper worthwhile; I just wonder if the first analysis is over-stated.

## Reply to Reviewer Comment 5

We had this same feeling sometime from a theoretical perspective, but apparently this is not apparent when NHST is applied and results are interpreted dichotomously and as representative of the true effect (see Hoekstra et al, 2008). This (once again) indicates that we should not trust human understanding of probabilities to assess probabilistic processes such as inferences based on statistics, and we hope to contribute to this realization by providing these results and an easy-to-use method to analyze (nonsignificant) statistical results.

1. Figures/tables/data availability: Please comment on the author’s use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).: Yes
2. Ethical approval: If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared. If not, please detail where you think a further ethics approval/statement/follow-up is required.: Human ethics not relevant.
3. Language: Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.: The article is well written. I would not call it "jargon free", and it could be written more plainly. I'm not the best person to comment on this.