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. Briefly, summarizing our response:

1. We did not explicitly interpret null-findings (*p* > .05) as evidence that the original authors the findings as support for the null. We now add statements like “results that are generally interpreted as null-findings”, without reference to how the original authors interpreted them.
2. We are sympathetic to the RP:P project (in fact, two of the authors were involved with most of the analyses and writing of the [Supplementary](http://science.sciencemag.org/content/sci/suppl/2015/08/26/349.6251.aac4716.DC1/Aarts-SM.pdf) report on statistics [pages 11-26]). We elaborated more on the RPP results, also using other re-analyses of the RPP data, putting our results of the Fisher test in perspective. We would like to disclose that the analysis presented in Application 3 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and actually conducted this analysis for a reply to an RP:P comment as part of the RP:P team (this comment remained unpublished, but can be found on the OSF website of the RP:P, <https://osf.io/smjge/>).
3. We agree with you that the use of *p*-hacking “solves” the power problem, in the sense that with *p*-hacking it is likely to obtain a statistically significant result when power is low (such that *p*-hacking increases power). We added a paragraph on *p*-hacking

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, as we showed with an additional analysis and references. We also carefully rephrased interpretations of our analyses, in line with the reviewer’s comments.

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.

*Thank you for your positive evaluation of our manuscript.*

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.

*We address all your and the reviewer’s comments in italics.*

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).

*We did not explicitly interpret null-findings (p > .05) as evidence that the original authors regard the findings as support for the null. More precisely, we do not regard the interpretation of the results throughout Application 1 at all, 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. 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 do interpret them as evidence for the null, they should be careful because our results show that false negatives may be lurking*

*We added this more explicitly in the introduction to Application 1, with the following: “Note that this application only investigates the evidence of false negatives in articles, not how authors might interpret these findings (i.e., we do not assume all these nonsignificant results are interpreted as evidence for the null).”*

*We also 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 2,009 papers that have no nonsignificant results when the alpha is shifted to .10. The change from p > .05 to p > .10 of course lowers the power of the Fisher test to detect FN because it lowers the number of “nonsignificant” p-values per paper. We do not present the results on p > .10 in the revised manuscript, but include the code in our OSF project to do so. We now refer to this in the section on “Evidence of false negatives in articles.” with the statement: “Results did not substantially differ if nonsignificance is determined based on $\alpha=.10$ (the analyses can be rerun with any set of $p$-values larger than a certain value based on the code provided on OSF; \url{https://osf.io/qpfnw}).”*

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.

*We do not feel that we criticize the RP:P, at all. We just attempt to objectively interpret the RP:P findings.*

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.).

*We now elaborate more on the RP:P findings using other reanalyses of these findings as well. After presenting the results of the Fisher test and our interpretation of them, we conclude that the Fisher test is apparently not powerful to detect at least one FN in a set of results that may contain many zero or (very) small true effects – as opposed to detect stand-alone FN findings, which our section on the power of the Fisher test demonstrates (Table 2 in our manuscript). Then, we refer to four recent re-analyses of the RP:P findings with new, advanced statistical techniques, putting our analyses and RP:P findings into a new light. We briefly summarize the findings of these re-analyses, before coming to our final conclusion. This conclusion reads: “All in all, conclusions of our analyses using the Fisher are in line with other statistical papers re-analyzing the RPP data (with the exception of Johnson et al.) suggesting that studies in psychology are typically not powerful enough to distinguish zero from nonzero true findings.”*

*See our discussion section of Application 3 for a strongly revised section on the RP:P findings.*

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.

*It was not our intention to hold the RP:P to a very high standard, but just to provide an objective analysis of what the RP:P findings tell us about true effects in psychology. As noted previously, we expanded our discussion of the RP:P findings by relating our findings to those of four other statistical papers on the RP:P effect sizes. In our extended discussion we indeed conclude that studies in psychology are not powerful enough to distinguish zero from nonzero findings.*

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?

*Thank you for noting this. We deleted “irresponsible”, but kept the other parts of the sentence because it precisely reflects the spirit of the RP:P paper; be careful with drawing strong or final conclusions on an individual effect if its replication failed.*

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.

*See our revised section on the RP:P, where we address re-analyses of individual results. More particularly, we state on the basis of a Bayesian analysis of individual effects that takes publication bias into account: “assuming equally likely zero, small, medium, large true effects, they conclude that only 13.4% of individual effects contain substantial evidence (Bayes factor > 3) of a true zero effect.”*

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.

*Thank you for your suggestion, with which we wholeheartedly agree. We added a few sentences on p-hacking and its relation to power and its (lack of) change over time in the general discussion section. We include “… use p-hacking to artificially increase statistical power …” and added:*

*“The effects of p-hacking are likely to be the most pervasive, with many people admitting to using such behaviors at some point [28] and publication bias pushing researchers to find statistically significant results. As such, the problems of false positives, publication bias, and false negatives are intertwined and mutually reinforcing.”*

*Note, however, that analyzing the specific relation between p-hacking and power goes beyond the scope of the paper, given that there are many parameters that have different effects (see also Hartgerink et al. 2016 for difficulties of modeling p-hacking behaviors and their effects on p-value distributions).*

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

*Yes. It is characteristic of effect size estimation based on ɳ2. We made this clearer by adapting our sentence to: “Because effect sizes and their distribution typically overestimate population effect size based on effect size ɳ2, particularly when sample size is small \cite{Voelkle 2007-at,Hedges1981-og}, we also compared the observed and expected adjusted nonsignificant effect sizes that correct for such overestimation of effect sizes (right panel of Figure 3; see Appendix B).”*

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

*These numbers refer to the correlation effect size. We altered “mean effect r” to “mean correlation effect r” in the revised manuscript to clarify this.*

-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.

*The results in Figure 4 only consider papers that report nonsignificant results, so for parsimony of interpretation we prefer to retain this format.*

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

*For the large majority of tests, degrees of freedom is equal to the sample size minus a constant (usually the number of parameters, e.g. the number of groups in ANOVA/t-test), so we are confident in using this as a proxy. We added a similar description in the running text “degrees of freedom is a direct proxy of sample size resulting from the sample size minus the number of parameters in the model”*

-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?

*Indeed, it indicates there is evidence for >= 1 true effect in the significant gender results, just as the application to the nonsignificant results indicates if there is evidence for >=1 true effect in the nonsignificant results.*

*Is this interesting? We think so, and are sure many researchers agree with us. Note that the “test of evidential value” in p-curve is very similar (same logic, but the test uses a different statistical test) to the Fisher test applied to significant results. P-curve/evidential value tests have recently gained considerable popularity, which we interpret as evidence that researchers find it interesting to examine evidence of at least one true nonzero effect in a set of statistically significant findings.*

-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?

*In Applications 2 and 3 this is not possible. In Application 1 our unit of analysis is papers, with possibly multiple nonsignificant p-values per paper. In Application 2 and Application 3 we analyze single p-values, i.e. our unit of analysis is p-values.*

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.

*Thanks, no problem.*

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.

*We would like to take the opportunity here in our response to underline the importance of preventing or limiting false negatives. First, false negatives may hamper theory development. For instance, assume an interaction effect exists in the population such that the effect exists for men but not for women. Because of lack of power, only the main effect is observed and not the interaction. Because of this, the conclusion in the paper on the mechanism underlying the finding is fundamentally wrong. Second, false negatives may hamper human health or even cause death. In medicine, the positive effect of some medicine may depend on patient characteristics. A good example is the effect of antidepressants. Although they may have a positive effect on adults (which, by the way, is not sure at all), they seem to have awful consequences for young adults, such as suicide. This interaction was and can not be observed in small N samples (i.e., is a false negative). Particularly when sample size is low, finding true moderators is hardly possible (without p-hacking).*

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).

*Yes, but please do not forget “q-hacking”, which is researchers ignoring or omitting unwanted statistically significant findings, or changing test results to make unwanted significant results statistically nonsignificant. The reviewer refers to “q-hacking” as well, resulting in “clean results”.*

*Anyway, we agree that our wording in the paper was a bit too strong; in the abstract we adjusted  
 “This is unwarranted” to “This might be unwarranted”, “Neglecting effects can lead to ...” to “False negatives can lead to ...”. In the general discussion we adjusted “importance of focusing on false negatives as well” to “importance of paying attention to false negatives alongside false positives”.*

*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 already caused so many problems for science.*

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.

*Thank you!*

# 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?

*Thank you for your comment. We respond to your comment in two ways, first by re-analysis and second based on theoretical or simulation results. First, you are right that the ICC is typically used for normally distributed variables. However, transforming the non-significant p-values to a more normally distributed variable by a logodds transformation (log[p/(1-p)], with p equal to the transformed p-value and excluding p=1 for computational reasons), does not change the ICC value much; its value changes to 0.00175.*

*Second, it is well-known that p-values can be all over the place, see for instance Cumming’s “dance of the p-values” (e.g.,* [*here*](https://www.youtube.com/watch?v=5OL1RqHrZQ8)*). This is particularly so when focusing on nonsignificant p-values. On the other hand, for nonzero true effect size in combination with very large sample sizes most p-values will be <.001, but significant p-values are not examined in our paper.*

*We decided to add the value of the ICC after logit-transforming the p-values to our paper, with the following sentence: “the ICC of the log odds transformed p-values was similar, with ICC = 0.00175 after excluding p-values equal to 1 for computational reasons”.*

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.

*We agree with the reviewer that reverse p-hacking (or what we call q-hacking) may be rather frequent. This was one of the reasons why we started Application 2: in order to inspect whether evidential value was dependent on the expectation of the result and its significance. However, given the small number of observations in some cells we were unable to actually say anything about q-hacking.*

*We do not feel confident to state anything about whether nonsignificant results may have been focal or not. In Application 2, we directly experienced how difficult it is to determine whether a result is expected or not, and determining whether a result is a focal result or not is likely to be just as difficult or opaque. This was also a problem in the RP;P project for many of the 100 findings that were replicated.*

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.

*Good point. When k=1, the Fisher method is another way of testing whether the result deviates from a null effect conditional on being not statistically significant. The row for k=1 provides the power of that test. In the paper we now compare this power to the power of a regular t-test to test the null-hypothesis, with the following paragraph in the discussion of the statistical properties of the Fisher test:*

*“To put the power of the Fisher test into perspective, we can compare its power to reject the null based on one statistically nonsignificant result (k = 1) with the power of a regular t-test to reject the null. If η = .1, the power of a regular t-test equals 0.17, 0.255, 0.467 for sample sizes of 33, 62, 119, respectively; if η = .25, power values equal 0.813, 0.998, 1 for these sample sizes. The power values of the regular t-test are higher than that of the Fisher test, because the Fisher test does not make use of the more informative statistically significant findings.”*

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).

*We do not see the later rows of Table 4 and their interpretation as misleading or pessimistic, but as factual and realistic. Interpreting these rows is subtle. If exactly one nonsignificant is generated by a true nonzero effect, the power of the Fisher test* ***decreases*** *as a function of the number of p-values. If* ***all*** *nonsignificant p-values are generated the same underlying true nonzero effect size, power increases as a function of the number of p-values (see Table 2). The fact that the probability of rejecting the null-hypothesis of “no true nonzero effect” generally increases as function of the number of p-values in a paper suggests that multiple findings with nonzero true effects may exist in these papers.*

*We added a brief explanation to the text explaining that a higher percentage for articles with more results does not necessarily mean something ‘bad’, with the following sentence: “Consequently, we observe that journals with articles containing a higher number of nonsignificant results, such as JPSP, have a higher proportion of articles with evidence of false negatives. This is the result of higher power of the Fisher method when there are more nonsignificant results, and does not necessarily reflect that a nonsignificant p-value in e.g. JPSP has a higher probability of being a false negative than one in another journal.”*

*Finally, note that 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 some of these results are false negatives (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.

*We understand, since we originally had the same feeling as you seem to have. However, research on interpretations of NHST findings show that these findings are interpreted dichotomously and as representative of the true effect (see Hoekstra et al, 2006). 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. We include a paragraph in the introduction to highlight this:*

*“Unfortunately, NHST has led to many misconceptions and misinterpretations (e.g., [6, 7]). The most serious mistake relevant to our paper is that many researchers accept the null-hypothesis and claim no effect in case of a statistically nonsignificant effect (about 60%, see [8]). Hence, most researchers overlook that the outcome of hypothesis testing is probabilistic (if the null-hypothesis is true, or the alternative hypothesis is true and power is less than 1) and interpret outcomes of hypothesis testing as reflecting the absolute truth. At least partly because of mistakes like this, many researchers ignore the possibility of”*

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.