Anne Scheel

Department of Human-Technology Interaction
School of Innovation Sciences
Eindhoven University of Technology

Faculty of IE&IS, ATLAS 9.417
Postbus 513, 5600 MB Eindhoven, The Netherlands
a.m.scheel@tue.nl

18th February, 2021

Resubmission of Manuscript ID AMPPS-20-0081

Dear Dr Tullett,

Thank you very much for the thoughtful and constructive assessment of our manuscript. We apologise sincerely for the delayed resubmission. We have carefully considered each point raised by you and the reviewers and revised the paper accordingly.

The first concern you highlighted was raised by the third reviewer, Dr Fanelli, who believed our comparison of positive results in standard reports and Registered Reports to be invalid because he assumed that different measures had been applied to the two groups. In our attached point-by-point response to the reviewers we clarify that this was not the case: The sole difference between the groups was how they were sampled, but our coding procedure was identical for standard reports and Registered Reports (i.e., we did not compare verbal hypotheses in the one case to statistical hypotheses in the other, as Dr Fanelli assumed we did). To remove any potential ambiguity about our method for other readers, we have added a sentence to the Method section which explicitly asserts the identical treatment of papers from both groups.

The second point you highlighted was raised by the first reviewer, Dr Nuijten, who expressed concern that we might have overstated the causal interpretation of our observational finding and not paid sufficient attention to potential confounds. To address this point, we revised the introduction of our study to better reflect the limitations of our observational approach, and added a discussion of the potential confound of editorial-policy differences between journals that do vs don't offer Registered Reports in the Limitations section.

However, we chose not to follow Dr Nuijten's suggestion to change the title of our manuscript. As we elaborate in our point-by-point response to the reviewers, we do not agree that the current title has a strong causal implication: 'An excess of positive results' is a description of the standard literature in Psychology, which we and others have interpreted as 'excessively' positive based on a-priori considerations of prior probability and statistical power. Our study investigates this excess by testing

whether it occurs in the new population of Registered Reports. We further believe that a more plainly descriptive title, such as 'Substantially fewer positive results published in Registered Reports than in the standard Psychology literature' might do more harm than good by exacerbating the risk of unwarranted causal interpretations among readers. A final reason to keep the current title is that we would like to avoid confusion among the few dozen colleagues who have already cited our preprint. However, we are open to changing the title if you believe that this would be more appropriate.

We hope that the changes we made to the manuscript and the clarifications we provide in our response to the reviewers serve to address your main concerns with the paper.

Kind regards,

Anne Scheel, Mitchell Schijen, and Daniël Lakens
Eindhoven University of Technology

Response to reviewers

Reviewer: 1

Comment
In this study, the authors compared the prevalence of supported hypotheses in Standard Reports (SR) and Registered Reports (RR). Part of the study could be considered a direct replication of Fanelli's highly cited 2012 paper. Furthermore, the authors provide a systematic overview of the type of language used in RRs to introduce hypotheses – a resource that may prove very useful for future meta-scientific studies.

I think this is an interesting and well-written paper with solid and transparent methods. Specifically, I think the open science practices displayed within this paper are exemplary: preregistered with deviations from the preregistration explicitly mentioned, data and code shared in a very organized way, and a list of the software used, to name some.

I only have a few remarks that may potentially improve the manuscript, listed in order of importance.

Response
Thank you for the encouraging and constructive comments.

Comment
1. My main point of critique is that I think the authors are sometimes too strong in their conclusions. They rightly point out that their study is not an experiment, and several confounds are imaginable that could explain the lower rate of supported hypotheses in RRs. I agree with the authors that it seems plausible that the difference is caused by the higher rigor and transparency in RRs as compared to SRs, but this remains speculative as the methods do not justify a causal conclusion. In the abstract and the limitations section the authors provide nuanced explanations of their results, but in some parts of the paper the conclusions are too strong. Specifically, I think that the following two sections may need some rephrasing:

- *The title is not justified by the findings. By stating in the title that there is an "excess of positive results" you already strongly imply a causal interpretation of the results. I would suggest rephrasing the title to be more descriptive.*

Response
We have tried to be very cautious in our causal inferences without disregarding a certain level of abductive reasoning that we think is justified and necessary to interpret our findings. We explicitly do not wish to 'smuggle in' statements that misrepresent the limitations of our study design, and we are grateful for being alerted to any statements in our manuscript that might be unintentionally misleading in this regard.

However, we respectfully disagree about the causal content of our current title and the notion that a more descriptive title would be less misleading. 'An excess of positive results' is a statement about the standard literature in psychology, which is justified by previous empirical studies of the percentage of positive results combined with a-priori considerations of statistical power and prior probability (as well as empirical assessments of power in the literature). As we discuss on page 3-4, a positive result rate of over 90% simply cannot be reconciled with a 'healthy', bias-free literature, and can thus be considered 'excessive'. As such, the current title accurately reflects the structure of our paper: We begin with a discussion of the (known) excess of positive results in the standard literature, and then proceed to investigate it through a comparison with Registered Reports (and find that the RR literature is less 'excessive' and more compatible with outcomes that would be expected in a 'healthier' publication system). We do not claim that there is a single specific factor that fully explains the difference between Registered Reports and standard reports, and no such statement is implied in the title.

Although we would prefer to keep the current title (not least because a new title may lead to confusion among colleagues who have already cited our preprint), we are generally open to adopting a more plainly descriptive one, such as 'Substantially fewer positive results published in Registered Reports than in the standard Psychology literature'. However, we genuinely believe that this would lead to an increase of misinterpretations rather than a decrease, since readers will likely draw spontaneous inferences about the causes of our result from such a purely descriptive title.

Comment

- *In line 105-115, the authors already make the argument that publication bias and QRPs would be the most likely explanation for a lower positive result rate in RRs. I think this already needs some nuance and/or an introduction of potential alternative explanations.*

Response

We have changed the paragraph in question, removed the inference to publication bias and QRPs, and now refer to the possibility of differences in prior probability (line 113-121):

> "Because the standards for research quality in Registered Reports are at least equal to ordinary peer review, and the statistical power requirements may exceed those in the standard literature (Maxwell, 2004; Szucs & Ioannidis, 2017), such a difference would be unlikely to be due to 'failed' studies or false negatives. Barring large confounds, such as substantial differences in the prior probability of hypotheses tested in Registered Reports versus the standard literature, a much lower positive result rate in Registered Reports might then indicate that publication bias is not a desirable filter for poorly conducted studies, and that we ought to worry about high-quality negative results we are missing because of it."

Comment

2. I did not see a mention of whether the coders were blinded to the type of article and/or the hypothesis. If not, this may have affected the outcomes: if coders expected fewer positive results in RRs, it could have biased the coding. This may need to be included as a potential limitation.

Response

Thank you for noting this oversight. We added a sentence to the Method section (line 209-212) and now also mention this in the Limitation section (line 362-363).

> Line 209-212:
> 'Because removing all indicators that could have identified Registered Reports as such from their full texts would have been practically impossible, coding was not blind to publication format (Registered Report vs standard report).'
>
> Line 362-363:
> 'Since coders could not be blinded to an article's publication format, their judgment may have been biased.'

Comment

3. In their limitations section, the authors point out potential confounding variables. One possible confound that wasn't mentioned but that may be of interest is a systematic difference in the journals included in each of the groups. There may be something systematically different about the journals that accept RRs that may affect the rate of positive results. It may be worth checking if there are any obvious differences in the journals (e.g., whether journals that publish RRs are more often open access, whether they have additional open science policies in place, whether they are more often interdisciplinary, etc.

Response

A systematic difference in editorial policies between journals offering RRs and those not offering RRs might indeed affect the positive result rate (although the rather modest effects of many journal policy innovations such as author checklists make us sceptical that this could explain a substantial portion of the very large difference we observed). We have added this potential confound to the Limitation section (line 372-384) and also conducted a cursory analysis to address the concern (the calculation of the new numbers reported here has been added to our open analysis code):

> 'As a third potential confound, journals that offer Registered Reports may have more progressive editorial policies which aim to reduce publication bias and type-I error inflation for all empirical articles they publish. This could lead to less bias in the Registered-Reports literature even if the format's safeguards against certain QRPs were actually ineffective. Additional research, ideally with prospective and experimental or quasi-experimental study designs, is needed to further investigate the influence of such factors. However, a cursory look at the three journals which contributed both standard reports and Registered Reports to our dataset (*Attention, Perception, and Psychophysics*, *Cognition and Emotion*, and *Frontiers in Psychology*)

> suggests that the pattern observed in our main analysis may hold for within-journal comparisons, which would speak against a strong influence of an editorial-policy confound: In these three journals, 11/13 (84.62%; 95% CI [54.55, 98.08]) standard reports had positive results, compared to only 7/14 (50.00%; 95% CI [23.04, 76.96]) Registered Reports.'

Comment

4. In line 127-134, the authors argue that an advantage of this study compared to Allen and Mehler is the use of an "established" method. However, the fact that Fanelli has used this method twice, does not inherently make this a better method than the one by Allen and Mehler. This is especially not a very strong argument given that you did not in the end retain Fanelli's categories, and switched to a binary classification. Furthermore, you updated Fanelli's method by extending the sentences to search for, which is great. Therefore, I would remove this argument and simply state that there was a parallel effort you were unaware of.

Response

We agree with the reviewer that the mere fact that a method was used in previous publications says little about its quality. We have therefore removed the words 'using a previously established method (Fanelli, 2010, 2012)'. The respective sentence now reads (line 137-139):

> 'A major advantage of our study, which was planned around the same time (we were unaware of Allen and Mehler's parallel efforts), is the ability to directly compare Registered Reports with the standard literature.'

Our original phrasing was intended to convey the advantage of a reproducible method which has been adequately described in previous studies over the informal and scarcely documented survey provided in Allen & Mehler's opinion piece.

We would like to point out that we did not use a different method than Fanelli (2010). We retained his coding categories: Like Fanelli, we first coded hypotheses in three categories ('support', 'partial support', 'no support' - this information is available in our open dataset) and then aggregated 'support' and 'partial support' into one category for the eventual binary analysis. This is identical to the analysis by Fanelli, who also coded the articles in 3 categories, but aggregated two categories into one for the final binary analysis.

Comment

5. Please also report the degrees of freedom in your chi2 result (e.g., line 249, but also in other cases).

Response

Thank you very much for noting this oversight. We have added the degrees of freedom to all reported X^2-values in the manuscript.

Comment
6. In your disclosures you state that you report how you determined your sample size. I may have missed it, but I don't remember seeing a justification for the 150 SRs?

Response
Thank you again for pointing out this inconsistency. We added the following sentence to the Method section (line 165-168):

> 'The sample size of standard reports was pre-specified to replicate the one used by Fanelli (2010), n= 150, since it matched the maximum number of Registered Reports available at the time (n= 151, see below) and piloting indicated that the required coding time would just fit our resource constraints.'

Signed,
Michèle Nuijten

Reviewer: 2

Comment
In "An excess of positive results: Comparing the standard Psychology literature with Registered Reports," the authors attempt to "test if Registered Reports in Psychology have a lower positive result rate than articles published in the traditional way and to estimate the size of this potential difference." The reported study achieves that goal, plus a few others. My comments below mostly fall into two categories: 1) Minor recommendations for clarification or 2) Explanations for why possible confounds are adequately addressed by the authors.

The most important limitation of the study, which is adequately addressed by the authors, is the sampling strategy to focus on the 1st hypothesis reported in the standard reports. This decision is justified in the methods and the discussion further emphasizes how slightly different strategies would result in similar findings.

My overall recommendation is to be highly supportive of the work, and for the authors to consider some of the minor points mentioned below.

Minor points:
1) Line 62: "and QRPs have been admitted by scientists in several survey studies." Considering adding evidence of this practice in ecology, evolution, and education fields: Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. PLOS ONE, 13(7), e0200303.
https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10%2Fgdtmg2&data=04%7C01%7C%7C9e95c427a8754e01750208d87cd6a7c0%7Ccc7df24760ce4a0f9d75704cf60efc64%7C1%7C1%7C637396608662357221%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%

3D%7C3000&sdata=Y%2FoisnS1puS566n1aJE9YTKbPjSBQvgD5P5mcEGr%2BG8%3D&reserved=0

Makel, M. C., Hodges, J., Cook, B. G., & Plucker, J. (2019). Questionable and Open Research Practices in Education Research [Preprint]. EdArXiv. https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.35542%2Fosf.io%2Ff7srb&data=04%7C01%7C%7C9e95c427a8754e01750208d87cd6a7c0%7Ccc7df24760ce4a0f9d75704cf60efc64%7C1%7C1%7C637396608662357221%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000&sdata=KjNSNTM%2FOrVrouHMfOopRRonKyI9yYSUmAoPAGT2R1Q%3D&reserved=0

Response

Thank you for these useful suggestions. We have added the two references to the sentence in question (line 66-69):

> '(…) and QRPs have been admitted by scientists in several survey studies (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; Fiedler & Schwarz, 2016; Fraser, Parker, Nakagawa, Barnett, & Fidler, 2018; John, Loewenstein, & Prelec, 2012; Makel, Hodges, Cook, & Plucker, 2019).'

Comment

2) Line 70: "How many positive and negative results would such an unbiased literature contain in reality? We investigated this question by comparing the rate of positive results in the Psychology literature to studies published in a new format designed to minimise publication bias and QRPs: Registered Reports." I am convinced by this argument, but would a skeptic be?

Response

In the section following the sentence in question ('Methods to mitigate bias', pp. 5-6), we expand on this argument and describe the mechanisms of the Registered Reports format that protect against bias as well as those intended to safeguard other aspects of research quality (such as statistical power). We hope that these elaborations will convince even a sceptic, but we are open to additional suggestions to reinforce our argument (or discuss potential weaknesses we might have missed).

Comment

3) Line 109 "statistical power requirements may exceed those in the standard literature (Maxwell, 2004; Szucs & Ioannidis, 2017)," I recommend that the authors note that Szucs & Ioannidis, 2017 reinforces the presumption made in line 39-41, that the known rate of positive results is unlikely to be caused by a preponderance of highly powered studies. (Also repeated in line 329)

Response

Thank you for this helpful suggestion. We have added a sentence to the respective section in the Introduction (now on line 45-47) and an insertion to the respective sentence in the Discussion (now on line 345-347):

> Line 45-47:
> 'These two assumptions appear highly implausible a priori, and available

evidence on average statistical power in the literature shows that at least one does not hold (e.g., Szucs & Ioannidis, 2017).'

Line 346-347:
'Because this is highly implausible and contradicted by available evidence (e.g., Szucs & Ioannidis, 2017), the standard literature is unlikely to reflect reality.'

Comment

4) Line 145: "After conducting a pilot to test the planned procedure, we preregistered our study (https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fosf.io%2Fs8e97&data=04%7C01%7C%7C9e95c427a8754e01750208d87cd6a7c0%7Ccc7df24760ce4a0f9d75704cf60efc64%7C1%7C1%7C637396608662357221%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000&sdata=TCgkjmSl8R4isqdiTNTlpA6cF7Aau64UfUJdLcXM1UU%3D&reserved=0)." I recommend linking directly to the authors' preregistration document: https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fosf.io%2Fsy927%2F&data=04%7C01%7C%7C9e95c427a8754e01750208d87cd6a7c0%7Ccc7df24760ce4a0f9d75704cf60efc64%7C1%7C1%7C637396608662357221%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000&sdata=Fz%2FXzTb7rFrWfNRA3dg4qOZoRcyoa3aFxkskmBdYRqY%3D&reserved=0 to help the reader find their precise measures more easily.

Response

We have followed this suggestion and changed the link accordingly.

Comment

5) Line 162: "Using the same sample size as Fanelli (2010), we randomly selected 150 papers from the 1919 search results." Slightly confusing phrasing, I recommend "Using the same sample size as Fanelli (2010), we randomly selected 150 papers from the 1919 papers that resulted from that search."

Response

We address this in the next point.

Comment

6) Figure 1: I did not see an explanation for the "1 additional SR sampled" study added, though I did see an explanation for the resampling for the 8 additional SRs. This is extremely minor and might be addressed through rephrasing line 163 "Excluded papers were replaced by resampling…" to "Excluded papers were replaced by resampling twice…."

Response

Thank you for noting these suboptimal wordings. Regarding comment 6, we resampled an additional standard report because one of the original 8 resampled papers had to be excluded, which is displayed in Figure 1. We have changed the

paragraph in question to integrate comment 5 and 6 with a comment from Reviewer 1. It now reads (line 165 - 174):

> "The sample size of standard reports was pre-specified to replicate the one used by Fanelli (2010), n = 150, since it matched the maximum number of Registered Reports available at the time (n = 151, see below) and piloting indicated that the required coding time would just fit our resource constraints. Standard reports were selected by searching the 633 journals listed under 'Psychiatry/Psychology' in the Essential Science Indicators database for papers published between 2013 and 2018 that contained the phrase 'test* the hypothes*' in title, abstract, or keywords. We then randomly selected 150 papers from the 1919 papers that resulted from this search. Excluded papers were replaced by resampling twice (this decision was not preregistered), which led to accidental oversampling and a final sample size of 152 (see Fig. 1)."

Comment

7) Line 223: The determination of the "smallest effect size of interest" was the difference between the positivity rate of psych to social science research (lowest observed in the "soft" sciences). Though ultimately arbitrary, this decision was made explicitly, made ahead of time, and overall more justifiable that any alternative effect size. Therefore, I would recommend that the journal agree with the authors' decision in this case.

8) There is a clear description of pre registered results, which are unambiguously stated.

9) The exploratory findings do not substantially alter the main results. The database of hypothesis identifying phrases may be useful to other researchers in the future and, by itself, would be a used contribution to the literature.

10) Limitations: Not experimental, and the authors of RRs could simply be more carefully and assertive in reporting null results. These limitations are accurate and are appropriately described by the authors.
Another limitation: Sampling bias with Faneli's method is addressed appropriately- namely by the literature that sees similar results with other methods. Essentially, this is appears to be a highly generalizable finding

11) Deviations from preregistration are clearly indicated by the authors and do not alter the interpretability of my main findings:
"Excluded samples were replaced by resampling,"
"We overturned the preregistered plan that AS would additionally code a random subset of both groups, because the number of double-coded papers seemed sufficient after double-coding only the difficult cases."

12) Reviewer Disclosures: I feel that it is important to disclose my relationship to this body of research. I am partially responsible for maintaining the databases described by the authors, such as the RR website, the document that describes individual journals' policies, and the database of known RRs. I am a proponent of the RR model and work to implement it in various outlets. I do not feel that any of that diminished my review, but that

information should be shared with authors and, if reviews can be published (please do so), with later readers.

Response
Thank you for the constructive and careful assessment.

Reviewer: 3

Comment
I will start by disclosing my status as Daniele Fanelli, and I was pleased to see that my old results were replicated. As I will explain below in my comments, I think that this aspect of the study, and its ability to replicate similar surveys on registered reports (RR), is an interesting contribution. What appears to be the central test of the study, however, is fundamentally flawed. The authors used one method to sample the psychological literature, and used what is a completely different method to sample and measure positive results in RR. the two kinds of positive results being measured are of a fundamentally different kind. The first method, accurately replicated by the authors, will draw a sample of articles that, in the abstract, declare to have "tested a hypothesis" (or similar string) and will then count how many studies concluded the hypothesis is supported, regardless of how many experiments and statistical tests this implied - and it could indeed be no statistics at all! The second method takes articles that registered a series of statistical hypotheses, and sees whether the first test rejected the null. It seems clear from the start that these samples are not comparable. They are not only sampled in different ways (minor problem) but they have different measures taken on them. One is an omnibus assessment of what an overall study concluded, the other is an assessment of rejection of null hypothesis. I agree that the two concepts are not completely separated, but they cannot be confused either. I also appreciate that, as per original methods, when multiple hypotheses were tested the authors looked in depth in the paper at the first hypothesis only. But, going by memory, this kind of multiple hypothesis testing was rare in psychology and would still not entail that a single P value determined the conclusion. Multiple hypotheses were more common in Economics, and interestingly there was some indication that, when they did, then the first hypotheses was supported less frequently (possibly because the second or third hypothesis was supported instead).

Response
Thank you for the openness and for the comprehensive review.
Regarding this first and central concern, there seems to be some confusion about the method we employed. Contrary to what the reviewer seems to assume, we did not apply different measures to standard reports and Registered Reports: We did not evaluate verbal hypotheses in the one case and statistical hypotheses in the other. Rather, we applied the method described in Fanelli (2010) to all papers in our dataset, as we write on page 10:

> ‘The main dependent variable was whether the first hypothesis was supported or not, as reported by the authors. We tried to follow Fanelli’s coding procedure as closely as possible:

> By examining the abstract and/or full- text, it was determined whether the authors of each paper had concluded to have found a positive (full or partial) or negative (null or negative) support. If more than one hypothesis was being tested, only the first one to appear in the text was considered. We excluded meeting abstracts and papers that either did not test a hypothesis or for which we lacked sufficient information to determine the outcome. (Fanelli, 2010, p. 8)'

To reiterate: For both standard reports and Registered Reports, we identified the first verbal hypothesis in the abstract (or in the full text, if the hypothesis was not sufficiently clear from the abstract) and evaluated whether the authors had concluded support (full or partial) for the hypothesis or not. Indeed, as the reviewer points out, this approach ignores 'how many experiments and statistical tests this implied - and it could indeed be no statistics at all!'

After carefully revisiting the manuscript, we remain unsure about which statements might have given the reviewer the impression that we used a different coding procedure for Registered Reports. However, to minimise the risk of such misunderstandings for other readers, we added the following sentence on line 200-201:

> 'The coding procedure was identical for both article formats in all other respects.'

Comment

I don't particularly dispute the essence of this study's findings. There would be all sorts of reasons why RR and a random sample of studies would be different, which might include the removal of biases as well as many other factors, which however the study would not control for. The study's conclusions would be weaker than they currently appear, even if the measure was exactly the same in the two groups. But the fact that the measurement is not the same makes drawing any conclusions from the analysis even more problematic.

Response

Again, as we explain above, we used the same measure for both article formats. We tried to be very mindful of and transparent about the limitations of our descriptive, non-experimental study design, and are open to discussing additional limitations that we might have missed.

Comment

Here below I include a few specific comments and suggestions (notes I made whilst reading).

p3 l37: I wouldn't use the words "registered", and would clarify that the measure in question expresses the OVERALL conclusions, i.e. not necessarily a single statistical test, but possibly the combination of multiple lines of evidence. Moreover, statistics were not necessarily involved in reaching such conclusions.

Response

We were struggling to address this comment in full since the paragraph mentioned here does not contain the word 'registered'. To convey more detail of the method used in Fanelli (2010), we changed the sentence starting on line 36 in the following way:

> 'Similarly, a seminal study by Fanelli (2010) analysed authors' verbal conclusions in hypothesis-testing papers sampled from the literatures of 20 disciplines and found that 91.5 % of papers published in Psychology claimed support for their first hypothesis — the highest estimate of all disciplines in the study.'

Comment

p3 l55: I would notice two things: First, that the hypotheses in question are not mutually exclusive: so selection bias is not a "more plausible" explanation, but a complementary one.

Response

In the section of the manuscript highlighted here, we reject the hypothesis that the positive result rate in the psychological literature accurately reflects the research psychological researchers conduct. Our argument is that for this hypothesis to be true, 'both statistical power and the proportion of true hypotheses (i.e., the prior probability that the null hypothesis is false) that are tested must exceed 90%' (p. 3, line 41-42), which we deem highly implausible. Contrary to the reviewer's claim, the alternative hypothesis we introduce next — 'a selection bias towards statistically significant results in the published literature' (p. 4, line 49-50) — is not complementary to the first hypothesis: The literature cannot simultaneously be completely accurate (i.e., unbiased) and biased.

Comment

Second: we don't know how much of a "prediction" is based on prior evidence, either published or collected "tacitly" by the researchers. In other words, the priors' that researchers had about these hypotheses need to be high, which is not such an impossible scenario.

Response

If we understand the reviewer's argument correctly, he suggests that a prior probability of >90% is not implausible since authors may build on established knowledge either from the literature or from their own unpublished studies. Our response is twofold:

First, if the prior knowledge comes from the literature or unpublished studies, then extremely high priors are a new achievement that was made possible by less well-informed earlier work, which thus had much lower priors. The assumption that earlier *published* work had much lower priors is at odds with e.g. Sterling (1959), who reported 97% significant results in the literature more than 70 years ago. The assumption that authors' priors are based on *unpublished* work with lower priors, however, is actually in line with our own argument, namely that the published literature does not accurately reflect the research that is being conducted.

Second, even if it were possible to find a plausible explanation for such extremely high priors in the literature, these extremely likely hypotheses would still need to be tested with more than 90% statistical power. As we discuss in the manuscript, there is plenty of evidence suggesting that average power in psychological research is very far below this number.

Comment

p4 l45 more recent literature could be cited here, surrounding the "cluttered office problem". vs "file drawer"

Response

Thank you for this comment. We have changed the sentence in question and added two more recent references. The sentence now reads (line 70-72):

> 'Some have argued that negative results are often uninformative or the result of low-quality research and should not be published at the same rate as positive results to avoid cluttering the literature (e.g., Cleophas & Cleophas, 1999; Baumeister, 2016; Mitchell, 2014).'

Comment

p7 l29 again, I don't believe one should draw statistical conclusions about P values from this method.

Response

As we explain above, there seems to be a misunderstanding about our method. We did not analyse or draw conclusions about *p*-values in Registered Reports. We hope that our clarification noted above sufficiently addresses this point.

Comment

p10 l12 Something to notice here is that I had to look up the specific hypotheses often in Fields like Economics, where multiple statistical hypotheses were tested, often in a single regression function. If I remember correctly, there was at least suggestive evidence that the first hypothesis tested might often be less likely to be supported. I think this might be relevant since the RR sample used a fundamentally different method from the 2010 paper, which could reflect more closely this multiple-testing condition.

Response

Again, the concern raised here does not actually apply to our method, since the process of identifying and evaluating the (verbal) hypotheses we coded was identical for standard reports and Registered Reports. If the reviewer should remain unconvinced, we would like to invite him to consult our open dataset (https://osf.io/aqr2s/) and compare the coded hypothesis quotes (variable name: 'hypothesis_quote') and result quotes (variable name: 'result_quote') for standard reports and Registered Reports.

Comment

p11 l 55 Setting a threshold of 6% seems flawed to me, because percentages are not a linear metric. For example if the original values had been 95% and 89%, there would be no room at the top to test for a 6% difference.

Response

We would like to point out that our hypothesis was directional: We predicted that the positive result rate of Registered Reports would be lower than the positive result rate of standard reports. Insufficient 'room at the top to test for a 6% difference' thus would not have affected our test.

However, the reviewer's comment alerted us to the fact that the language we used to describe our test criteria is suboptimal in this regard, in that although we specified a directional prediction, we did report symmetric equivalence bounds, which was not necessary: The statement 'We would accept our hypothesis … if the observed difference … was significantly smaller than 0 *and* not statistically equivalent to a range from −6% to +6% (both at $\alpha = 5\%$)' (lines 251-254) is in fact equivalent to the statement 'We would accept our hypothesis … if the observed difference … was significantly smaller than 0 *and* not statistically equivalent to a range from -6% to 0%'. This is because for the first condition to hold, the 90% CI of the estimate would have to end below 0. A situation where the effect was significantly smaller than 0 yet its CI would somehow extend to the region between 0 and +6% is impossible, making this part of our original statement redundant. We thank the reviewer for making us aware of this redundancy. To address the issue and avoid confusion, we added the following footnote (footnote 4):

> "Note that these inference criteria are logically equivalent to 'significantly smaller than 0 and not statistically equivalent to a range from −6% to 0%': Since the first criterion (statistically smaller than 0) requires the 90% CI to end below 0, half of the equivalence range specified in the second criterion — from 0% to +6% — is redundant, (which we failed to notice before preregistering the analysis)."

Comment

The comparison with social sciences is methodologically questionable, given the heterogeneity of these groups (e.g. are we including Economics or not?).

Response

As we write in the manuscript, we based the smallest effect size of interest (SESOI) on the comparison between the positive result rates of Psychology and General Social Sciences reported in Fanelli (2010) because the latter discipline had the lowest positive result rate among the disciplines classified as 'soft' by Fanelli (the current reviewer). We thus reasoned that reducing the number of positive results by this amount would constitute a meaningful, non-trivial difference, because it would turn Psychology from the discipline with the highest positive result rate to the discipline with the lowest positive result rate in the soft sciences. Although coarse classifications by discipline cannot yield perfectly homogenous groups, this does not preclude informative comparisons between disciplines such as those conducted in Fanelli (2010) and Fanelli (2012), which inspired our analysis.

However, our SESOI and its justification are arbitrary to some degree (e.g., a SESOI of 5%, 7%, or 10% may have been as defensible). Specifying the alternative hypothesis at all is crucial to ensure the falsifiability of our prediction (see e.g. Lakens, Scheel, & Isager, 2018; doi: 10.1177/2515245918770963) and the SESOI we chose was intended as a first step in a new research line with very little prior knowledge to build on. The reviewer's comment alerted us to the risk that this provisional specification could be mindlessly imitated or even turn into convention. To protect against such misuse, we added the following elaboration (line 254-259):

> 'Specifying a smallest effect size of interest of 6% absolute risk reduction provides an initial yardstick to evaluate our results and make our prediction falsifiable. However, the value of ±6% does not possess an intrinsic theoretical meaning. As the emerging meta-psychological literature matures, we hope to see future research base the smallest effect size of interest on increasingly well-informed empirical and theoretical considerations.'

Comment

*In any case, accepting the need to define a point estimation to compare, I would use the proportional difference (100*91.5/85.5=107%, or even better the odds ratio: e.g. OR of reporting a positive results in psychology vs social science (0.915/(1-0.915)/(0.855/(1-0.855) = 1.825593*

Response

We specified the SESOI as an absolute risk reduction of 6% (absolute risk reduction being defined as $ARR = \frac{positive\ cases\ in\ group\ 1}{all\ cases\ in\ group\ 1} - \frac{positive\ cases\ in\ group\ 2}{all\ cases\ in\ group\ 2}$). The reviewer suggests to instead express the SESOI as a relative effect size such as relative risk ($RR = \frac{positive\ cases\ in\ group\ 1}{all\ cases\ in\ group\ 1} / \frac{positive\ cases\ in\ group\ 2}{all\ cases\ in\ group\ 2}$) or an odds ratio ($OR = \frac{positive\ cases\ in\ group\ 1}{negative\ cases\ in\ group\ 1} / \frac{positive\ cases\ in\ group\ 2}{negative\ cases\ in\ group\ 2}$). However, a measure of relative effect size is not appropriate for our research question because it ignores the base rate of the phenomenon in question: Our study is motivated by the excessively high proportion of positive results in the psychological literature reported by Fanelli (2010) and others, and we were thus interested to see if a non-trivial absolute reduction of this proportion would occur in Registered Reports. A measure of relative effect size would be appropriate if we were uninterested in the actual proportion of positive results in the literature, which is not the case. For an in-depth discussion of the use of absolute versus relative effect sizes, see e.g. Sprenger & Stegenga (2017; doi.org/10.1086/693930).

Comment

I am also doubtful about the validity of testing this type of point hypothesis for replication purposes. At the very least, a prediction interval ought to be used, but more generally there are well-known issues with ignoring the measurement error that goes into these types of comparisons. Work amply discussed in the context of replication studies…

Response

The issues the reviewer refers to with respect to replication studies do not apply to the current project. First, we should point out that the suggestion by the reviewer to use prediction intervals (as suggested in Patil, Peng, and Leek, 2016) has been severely criticized by Morey and Lakens (2016; doi: 10.5281/zenodo.838685) because, in addition to the arbitrariness of a confidence level and the difficulty of interpreting the 'success rate', the test is based on a non-significant test of no difference -- which is a statistical fallacy (we solve this problem by using an equivalence test to interpret the non-significant test of no difference). We have seen no uptake of the recommendation by Patil et al. in the last 5 years. Second, the comparison we report does not ignore the variability in the original estimate: we do not treat the value observed by Fanelli (2010) as the population mean, but compare his and our estimate in a regular significance test (i.e., taking sampling error into account).

Comment

p14 l16 I find this part of the analysis rather confusing. Apart from the technical questions raised above, how can observing a difference of less than 5% not be compatible with a null hypothesis of a range of ±6%?

Response

The reason that we fail to accept the hypothesis that the difference between the positive result rate we observed in standard reports and the one reported for Psychology in Fanelli (2010) is less than ±6% is that the CI of the observed difference overlaps with the upper bound of +6%. Accepting the hypothesis whenever the point estimate lies within the equivalence bounds would lead to very high and uncontrolled error rates — this would be akin to rejecting a null hypothesis of zero effect whenever a point estimate is not exactly 0. An accessible introduction to the logic and practical applications of equivalence testing can be found in Lakens, Scheel, & Isager (2018, doi: 10.1177/2515245918770963).

Comment

p20 l 11 The paragraph discussing limitations does not go at the core of the study's main problem.
The core problem is that the two samples being compared are sampled in fundamentally different ways, and measured in different ways, too.
As already mentioned above, Fanelli 2010 used "test the hypothes" as a proxy of how authors draw conclusions from studies as a hole. The proxy could have implications for statistical tests, too, but these are very indirect, and multiple other possible explanations could apply to differences observed (all discussed in Fanelli 2019 and 2012). The whole point of using such a proxy at the time was that it allowed to compare disciplines and samples, using the same measure.
But if one sample obtains papers and measures "positive" results in one way, and the other uses a completely different measure, there is little to be learned. This is true of any study, as the old "apples vs oranges" adage reminds us.
Therefore, contrary to what the authors conclude at the end, this is not a "systematic comparison".

Overall, the study makes an interesting contribution for the part that seems to replicate results of multiple studies, Fanelli on one hand, and other surveys of RR on the other. But the comparison between the two groups is seems quite clearly flawed to me. The study in my opinion should remove or re-think the latter.

Perhaps the authors could, for example, compare RR with a random sample of studies that are actually testing multiple statistical hypotheses? Or maybe compare RR with pre-registered studies?

Response

As discussed above, this fundamental concern with our study seems to be based on a misunderstanding of our coding procedure. The single major difference in how we treated the two groups of papers in our study is the sampling procedure: Standard reports were sampled by replicating the sampling strategy reported in Fanelli (2010), whereas Registered Reports were identified by means of a curated database. In both cases, however, we made sure to only include papers which reported testing at least one hypothesis.

A central question for interpreting our results is whether Fanelli's (2010) sampling method yields a sufficiently representative sample of hypothesis-testing papers from the Psychological literature. We discuss this concern in depth throughout the manuscript, for example in the Introduction section on page 7, in the Method section on page 11, in the Results section on page 14-16, and in the Discussion section on page 21. It inspired our exploratory analysis of the phrases used to introduce hypotheses in Registered Reports (reported on page 14-16), and we openly discuss that conceptual replications using alternative search phrases are needed to establish the generalisability of our estimate of the positive result rate of standard reports. However, as we discuss on page 22, a growing number of studies provide converging evidence that the ballpark estimate of a positive result rate around/above 90% is reasonably robust to a range of different sampling procedures.

We believe that the nuance and caveats we provide throughout the manuscript are adequate to justify our conclusions, and that the presentation of our findings leaves room for unconvinced readers to interpret the results for standard reports and Registered Reports independently of each other.

We thank the reviewer for his comprehensive and challenging remarks, which we believe have made our manuscript stronger.