**Peer Review and Communication History**

**MS Title**: The Effect of Preregistration and P-Value Patterns on Trust in Psychology and Biology Research

**Author Names**: Clare Conry-Murray, Ann Mcconnon, Morgan Bower

**Submitted:** March 2, 2021

**Editor First Decision**: Revise & Resubmit

Oct 12, 2021

Dear Clare Conry-Murray,

thank you for submitting your manuscript to Collabra. Please excuse the delay in this response.

In addition to inviting two reviewers to read your manuscript, I have carefully read the manuscript myself and agree with both reviewer’s assertion that the manuscript contributes to the growing knowledge of the impact of open science practices, but has some considerable weaknesses that should be resolved before publication in Collabra. Therefore I have decided to ask you to revise and resubmit this manuscript. Please note that in-principle acceptance (IPA) and progression to Stage 2 is not guaranteed and any revision is likely to be sent back to the referees for further review. You should, therefore, include a document with a point-by-point response to the reviewers’ comments, outlining each change made in your manuscript or providing a suitable rebuttal.

Both reviewers recommend restructuring the introduction to better depict the relationships between p-Hacking and HARKing and the possibilities provided by preregistration and p-curves on the other. Additionally I would like to emphasize the point made by Reviewer 2 that materials provided on OSF to accompany this paper should allow for an execution of the R script with the data provided. Currently this sadly not the case. As both reviews mention, I would also appreciate an earlier reference to the preregistration itself as well as a more detailed elaboration on the decisions made in cut-off selections and participant exclusions “to make groups approximately equal”.

In addition to the suggestions made by the reviewers I would also like to add the following questions from my own review of the manuscript.

1. Were the abstracts that you used during the study rated regarding their overall trustworthiness by independent biologists and psychologists? If not, the differences you found between the two areas may simply be an artifact of the differences in the baseline quality of the abstracts.
2. You provided links within the abstracts to preregistrations but state only that particpants “did not need to follow it”. If possible, please provide information on how often participants did follow these links and what impact it had on their rating.
3. You state that the linear-mixed effects model (LMM) “provided better fit” than the repeated-measures ANOVA (RM-ANOVA). Please elaborate on this. Given strictly positive variances for all random intercepts both analysis approach should provide identical F-Values if conducted in the same fashion. Additionally, marginal and conditional R² values for the LMM would be appreciated to provide a better understanding for the overall impact of the investigated conditions on the judgments of trustworthiness.
4. Regarding your explorative analysis you claim to investigate “whether one field is more or less sensitive to p-Hacking and whether our variables interact in the field”. However, your results section as well as the code provided on OSF clearly indicate that you investigated only main effects in this model, thereby not providing any insight into field-differences in the impact of preregistration and p-Curves. Instead the main effect only denotes base differences in estimated trustworthiness.
5. You refer to Field et al.’s (2020) results, but those were related to whole articles, not merely abstracts. It is plausible to assume that the impact of preregistration and p-Curves you found here are simply due to the fact that they are embedded in very little additional information with which to assess the quality of the study (in contrast to Field et al., 2020). Please discuss possible origins for this difference in more detail.

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to resubmission.

If you have any questions or difficulties during this process, please contact us at the editorial office editorialoffice@collabra.org.

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,
Martin Schultze

**Reviewer 1**

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

In their manuscript, the authors aim to evaluate whether a preregistration and a diverse series of p-values smaller than .05 mentioned in an abstract increase trust in the research presented in the abstract. The authors were especially interested, if this differs between undergraduates, researchers and practitioners. They sampled data from all three groups from the fields of psychology and biology.
I agree with the authors that the question, how do these factors influence the perceived trustworthiness, is certainly a relevant and important question. Especially now, that Open Science practices are used more frequently, it needs to be investigated how this is perceived by the readership. This will result in further recommendations concerning education of the readership but also for researchers who conduct and write up the research.
Moreover, I would like to praise the authors for their use of Open Science practices. Their study was preregistered, a power analysis was done, and code, data, and the questionnaire are shared online. I wish to see more studies in the future, making their practices transparent like this one.

However, I cannot recommend the manuscript for publication in its current form for several reasons, which I list below.

The authors start early on in their introduction to mention the concepts of both preregistration and p-curves, but unfortunately, neither is ever really explained. Readers interested in Open Science might be familiar with those concepts, but unfamiliar readers do not know what these concepts mean. These should be explained in sufficient detail early on. Moreover, the authors seem to mix up the words *p-curve* and *series of p-values*. P-curve always refer to a distribution, but the authors write, for example in their discussion explaining the results: “but p-curves were considered primarily by researchers and professionals.”, but participants were only given a series of p-values, they couldn’t consider the according p-curve. The usage should be clear throughout the manuscript, otherwise it is too confusing.

The authors mention the study by Field et al. (2020) three times, but only very roughly. However, this seems to be a highly crucial study and of very high importance for the authors’ study, since Field et al. (2020) already investigated the trustworthiness of preregistration on research. Given the importance of the Field et al. study, I would like to see the study discussed in more detail. What can be learned from this study already? And I would like the authors to make explicit within the manuscript how their study adds to that and where it is different. Also, there is the repliCATS project going on since 2019, that tries to evaluate how readers make judgements about the replicability of research. The authors mention the DARPA initiative only in the discussion, but it would be good to also see it reflected in the intro already. Recently, a preprint came out with the very first results (Wintle et al., 2021; references at the bottom of this review).
The authors also call preregistration “a marker of a higher quality study”, but then cite Claesen et al. (2019) who showed that a preregistration is not necessarily an indicator of higher quality. I would like to see this mentioned in a more nuanced way. It could very well be that a preregistration is UNDERSTOOD as an indicator of higher quality, but that doesn’t mean that it definitely IS.

Also, I think the introduction is a bit short when it comes to which populations (researchers, students, professionals) have which knowledge. I think it may be more mixed and heavily depends on the university and environment the participant is in. As the authors wrote, more and more teaching involves the education about questionable research practices and how to counteract them. The authors cited Chopik et al. (2018), but should also have a look at the paper by Sarafoglou et al. (2020). Moreover, there are more and more ReproducibiliTea Journal Clubs, which are bottom-up initiatives educating about Open Science, often attended by students and early career researchers (Orben, 2019). Therefore, individual differences among students could be explained whether their university has such offers or not. I would wish to see this reflected in the introduction.

It is commendable that the authors conducted a power analysis and also reported it, but I’m missing a justification for the effect size. Where does it come from? Was it informed by other studies (e.g., Field et al., 2020), or was it the smallest effect size of interest? Also, how do the authors explain that the Field et al. (2020) study with a little bit smaller sample size was underpowered based on their pilot study, while the authors here calculated a power of .90 for their effect of interest? Also, I was wondering why the preregistration states that there are 8 abstracts, but the manuscript reports 12?

For the results section, I would wish the results of the ANOVA would also be reported. I have to say that the authors followed their preregistration perfectly; i.e., they wrote that if they feel confident with the linear-mixed effects models, that those are going to be reported. In fact, authors wrote in the manuscript that “because the linear mixed-effects model provided better fit, we only report the mixed-effects analysis here”. I would like to ask that the ANOVA results get reported as well (e.g., in a footnote) for transparency reasons because then the reader can take that information into account as well.

Concerning the discussion section, I have to admit that I am not entirely convinced by the arguments why the here reported effect size of the preregistration effect is small. It is argued that it could be because participants might know that preregistration can be used inappropriately (as has been shown by Claesen et al., 2019). But personally, I believe that the things uncovered by Claesen et al. (2019) are not known to many, especially not outside of the field of Open Science. Unfortunately, this cannot be solved by common sense and not by the info the study of this manuscript offers. Therefore, I am unsure about the interpretation of the results. We can see the effects on trustworthiness, but the reasons for them are currently in the dark.

Thus, I do not think that the current manuscript offers enough “flesh” or substance to be published in the present form. I think a second follow up study that uses the same experimental manipulation, and then additionally asks the participants a lot of questions for potential reasons afterwards could really shed a lot of light on what’s going on here. What I’m thinking about, e.g., awareness of preregistration/p-hacking, experience with preregistration, belief in preregistration as a tool to combat p-hacking, knowledge about p-curves, and so on and so forth. Lastly, I think the title does not entirely convey the content of the study since the focus was not purely on p-hacking but also good research practices (i.e., preregistration).

Minor comments:

* Methods: in the participants characteristics table, the column Biology is not really self-explanatory
* Results: Table 3, to me it is entirely unclear what the planned contrasts between the subscript comparison looked like. The explanation does not help, please clarify in the table description. Also, it says Registration instead of Preregistration in there.
* General: The authors should check their citation manager. For how to cite links, please have a look at the latest APA guidelines. Also, I would recommend to cite the Open Science Collaboration (2015) as such and not just Collaboration (2015).
* In the discussion, sometimes single letters appear (such as r and f)
* Spelling errors (such as intellence and practitioner sand student)

References
Sarafoglou, A., Hoogeveen, S., Matzke, D., & Wagenmakers, E. J. (2020). Teaching good research practices: Protocol of a research master course. *Psychology Learning & Teaching, 19*(1), 46-59.
<https://journals.sagepub.com/doi/pdf/10.1177/1475725719858807>

Orben, A. (2019). A journal club to fix science. *Nature, 573*(7775), 465-466.
<https://www.nature.com/articles/d41586-019-02842-8>

Wintle, B., Mody, F., Smith, E., Hanea, A., Wilkinson, D. P., Hemming, V., … & Fidler, F. (2021). Predicting and reasoning about replicability using structured groups. *MetaArXiv*.
<https://osf.io/preprints/metaarxiv/vtpmb/>

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  | ✔ |  |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

**Reviewer 2**

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

This article is timely, informative, and will make an important contribution to the literature on researcher practices & perceptions of the trustworthiness of published papers. I recommend that it be given a ‘revise and resubmit’ decision addressing the points below. The most critical point to address would be the questions around sample size and your stopping rule - I think this could be most straightforwardly addressed with an additional replication study, but a version without any additional data collection would also be possible in my view (details below).

My comments are bulleted below; the final section describes my review of the materials shared alongside the manuscript.

* I have a comment on the title you’ve given to this manuscript, which is unusual itself, and my advice is also maybe a little unusual: I think you should include that you have this finding “In biology and psychology papers.” This is technichally a less general claim, but it would avoid the (wrong) assumption that many people may make that studies like this are only in psychology.
* I would advice taking another pass with more attention to structure of the introduction; an unfamiliar reader may not understand how preregistration, HARKing, and p-hacking are specifically related to each other and to p-curves. In particular, the paragraph on p-curve is placed after the section on preregistration, instead of with the paragraphs on p-hacking and HARKing. Relatedly, I don’t think it would necessarily be clear to a naive reader that preregistration is both a method for avoiding p-hacking, and then also a separate signal of credibility that you’ll be measuring in your study.
* You might be able to solve some of this by separating this set of topics (roughly pp. 2-5) instead into two parts: Actions that researchers can take (HARKing and p-hacking on one hand, or preregistration on the other, even though it’s an imperfect tool.) And then signals you can see in the paper, namely an unexpected p-curve shape or a statement of preregistration.
* You mention that you didn’t preregister the exclusions (thank you for including this), but I wasn’t clear for several pages whether you preregistered your own study or not, and how you justified your sample size. I finally found a reference to it, and to the power analysis you ran much later in the methods section! I would recommend that you move these links to earlier in the methods section or end of introdcution.
* In your preregistration, you stated an intent to collect data from “at least 206 participants, but as many as 300.” There is no information about a stopping rule included in your preregistration. In the end you collected data from 318, and used exclusion criteria you developed post data collection. I also see a reference to excluding or re-classifyiing some data points in order to make group sizes approximately equal. This is a potential threat to the validity of your study. First, the preregistered sample size gives you the flexibility to stop at anywhere between 2/3 and all the way through a data collection process - how did you decide when to end, and had you seen any initial results at that point? The most serious concern here is whether your knowledge of results could have impacted (even unconsciously) the analysis decisions you made. Because of this, I would like to see at least one of the following in a revision:

(1) Give a more complete account of the process that lead to the final sample. This might include e.g. how many batches of participants you ran, how you decided to end data collection overall, and what information you looked at to determine the need for data dropping and re-classification. Then, calculate and report all of your results in two versions: one with the data as you have it, and one with all data as originally coded. Finally, you would need to address any differences that result from these two different versions of your dataset. The stating of all statistical findings of one of these datasets could be ‘demoted’ to a supplemental section, but the main text should discuss their existance and any difference in their interpretation.

(2) Preregister and conduct an exact replication using the group and inclusion criteria you have now developed, with a single intended target size and a more exact stopping rule that can be implemented cleanly given the logistics of your data collection. (There are many possiblities here! This could be anything from intentionally over-collecting based on your empirical observation of how much data is likely to be lost, to running in small batches until the inclusion criteria result in a dataset that is larger than your intended sample size.)

* A second potential threat to validity is the fact that you preregistered to conduct both linear mixed model and ANOVA analysis, but then to report the LMM only “If the results are the same for both analyses, and we feel confident about the analysis, we will report the linear mixed effects model.” However, in the main paper you reported only the LMM, but no ANOVA results, “because the linear mixed-effects model provided better fit.” I wasn’t sure if ‘fit’ here referred to a later decision that the LMM is a more appropriate analysis for this data, or the fact that the inferential statistics were weaker in the ANOVA. In either case, I would ask for a revision to include a supplement reporting all statistics, and for the main text to discuss this and any resulting differences.
* In the discussion, a point on inclusivity: you reference participants’ age as related to their progression through school (and thus level of familiarity with statistics) - note that not all 22 year olds are close to the end of an undergraduate degree!
* On a small citation note, if you want you can replace the Rogers (2019) Wired article with this preprint(<https://osf.io/preprints/socarxiv/46mnb/>) for a reference to the SCORE project.

**Replicability & materials**

Thank you for including the full text of the survey as well as cleaned data & analysis scripts! This is greatly appreciated. A few notes:

* Thank you for including a codebook! One column was opaque to me and not defined in the codebook -DS. Could you please provide a definition?
* Please note that you have a potentially identifiable column (PID - I don’t know whether these are identifiable to participants or not, but in general I would recommend not using the software-generated ID strings.) One line includes a participant’s full name!
* If possible, can you share a version of this data that has not yet been cleaned/reduced to analyzable participants? This will improve the independent reproducibility of your work.
* Additionally, I was unable to re-run your analysis scripts, because the provided XLSX cannot be read into R the way it looks like your script intends. In particular, the early part of the script refers to columns including Prereg, phacked, and Trust\_score, which are not present in this version. Can you share a version of the dataset that this analysis works over

[For clarity, I’m asking for two versions of the data: A version as close to raw as possible (redacting identifiable columns), and since your pipeline seems to involve manual work to prep for analysis, a final version of the data that runs with your R script.]

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  |  | ✔ |

**Author Response**
Nov 12, 2021

Dear Dr. Schultze

Thank your or the opportunity to revise our manuscript. We found the reviews extremely helpful, and we hope you will find our manuscript improved. We have included the comments from you and the reviewers here and we respond to each point below:

From the editor:

Both reviewers recommend restructuring the introduction to better depict the relationships between p-Hacking and HARKing and the possibilities provided by preregistration and p-curves on the other.

RESPONSE: We have restructured the introduction in response to these comments.

Additionally I would like to emphasize the point made by Reviewer 2 that materials provided on OSF to accompany this paper should allow for an execution of the R script with the data provided. Currently this sadly not the case.

RESPONSE: OSF now has the raw data, as well as the clean data that is ready for the R script to be run. We apologize for posting the incorrect data previously.

As both reviews mention, I would also appreciate an earlier reference to the preregistration itself as well as a more detailed elaboration on the decisions made in cut-off selections and participant exclusions “to make groups approximately equal”.

RESPONSE: We have now noted at the end of the introduction on p. 8 that the study was preregistered. We also now explain how our groups were defined differently from the preregistration on p. 9-10. We note that this may have made our hypothesis more difficult to confirm. We made the group size/criteria decisions before data analysis. From page 9-11: “These groups differ from the preregistration in that Master’s level students were included with researchers but we preregistered that they would be included with students. We made the decision to include Master’s level students with researchers to make the expertise groups sizes more equal. The analysis with master’s students in the “student” group is available in the supplementary materials.”

In addition to the suggestions made by the reviewers I would also like to add the following questions from my own review of the manuscript.

1. Were the abstracts that you used during the study rated regarding their overall trustworthiness by independent biologists and psychologists? If not, the differences you found between the two areas may simply be an artifact of the differences in the baseline quality of the abstracts.

RESPONSE: We used “typicality” as a measure of validity of the abstracts, rather than trustworthiness, since trustworthiness is our DV. To test whether experts in each field viewed the abstracts differently, we have now examined whether typicality of the abstracts was judged differently by Psychology researchers compared to Biology researchers. In fact, the 35 Psychology researchers rated the abstracts as more typical (M= 5.37 (SD = 2.13, where 0= very unusual to 10 very typical) than the 39 Biology researchers (M = 4.82, SD = 2.43); however this difference was not significant, *t*(72)= 1.3, *p* = .305. (The difference in typically was also not significantly different between fields in the whole sample.)

We also note that the abstracts were almost exactly the same-- with novel acronyms describing a biological test or a personality tests for the sleep abstracts, and drugs or therapy for the depression abstracts, but otherwise the abstracts described the same study design, and used the same wording etc. We chose to focus on sleep and depression because they both have biological and psychological elements.

1. You provided links within the abstracts to preregistrations but state only that participants “did not need to follow it”. If possible, please provide information on how often participants did follow these links and what impact it had on their rating.

RESPONSE: The links in the abstracts did not work. They were fake links. Therefore, if anyone had tried to follow them, they would have received an error message. Unfortunately, we do not have data on how many tried to follow the links.

You state that the linear-mixed effects model (LMM) “provided better fit” than the repeated-measures ANOVA (RM-ANOVA). Please elaborate on this. Given strictly positive variances for all random intercepts both analysis approach should provide identical F-Values if conducted in the same fashion. Additionally, marginal and conditional R² values for the LMM would be appreciated to provide a better understanding for the overall impact of the investigated conditions on the judgments of trustworthiness.

RESPONSE: We now explain better that our model was improved by including random effects over including only the intercept (p. 14). We have now also included both the marginal and conditional R² values for the LMM (see p. 14).

1. Regarding your explorative analysis you claim to investigate “whether one field is more or less sensitive to p-Hacking and whether our variables interact in the field”. However, your results section as well as the code provided on OSF clearly indicate that you investigated only main effects in this model, thereby not providing any insight into field-differences in the impact of preregistration and p-Curves. Instead the main effect only denotes base differences in estimated trustworthiness.

RESPONSE: We have now included interactions in our exploratory analysis, and a table (Table 4) to show all differences between the two fields in terms of judgements with and without preregistration and by p-value patterns. These are located on p. 15-16. We have also updated our posted code to include this new analysis.

1. You refer to Field et al.’s (2020) results, but those were related to whole articles, not merely abstracts. It is plausible to assume that the impact of preregistration and p-Curves you found here are simply due to the fact that they are embedded in very little additional information with which to assess the quality of the study (in contrast to Field et al., 2020). Please discuss possible origins for this difference in more detail.

RESPONSE: We have now added more discussion to the reasons for differences with Field et al. (2020) on p. 17. Field et al. used descriptions of studies and situations that were more detailed than abstracts and also included context about the familiarity of the researchers to the participants (former collaborators or not). We now add more details about how our study differs from Field et al., and the ways in which the two studies found similar results (i.e. no increase in trustworthiness for preregistration from researchers).

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

RESPONSE: Thank you!

From Reviewer 1:

In their manuscript, the authors aim to evaluate whether a preregistration and a diverse series of p-values smaller than .05 mentioned in an abstract increase trust in the research presented in the abstract. The authors were especially interested, if this differs between undergraduates, researchers and practitioners. They sampled data from all three groups from the fields of psychology and biology.

I agree with the authors that the question, how do these factors influence the perceived trustworthiness, is certainly a relevant and important question. Especially now, that Open Science practices are used more frequently, it needs to be investigated how this is perceived by the readership. This will result in further recommendations concerning education of the readership but also for researchers who conduct and write up the research.

Moreover, I would like to praise the authors for their use of Open Science practices. Their study was preregistered, a power analysis was done, and code, data, and the questionnaire are shared online. I wish to see more studies in the future, making their practices transparent like this one.

RESPONSE: Thank you!

However, I cannot recommend the manuscript for publication in its current form for several reasons, which I list below.

The authors start early on in their introduction to mention the concepts of both preregistration and p-curves, but unfortunately, neither is ever really explained. Readers interested in Open Science might be familiar with those concepts, but unfamiliar readers do not know what these concepts mean. These should be explained in sufficient detail early on.

RESPONSE: WE have now added a short explanation of each to the first paragraph on p. 2.

Moreover, the authors seem to mix up the words *p-curve* and *series of p-values*. P-curve always refer to a distribution, but the authors write, for example in their discussion explaining the results: “but p-curves were considered primarily by researchers and professionals.”, but participants were only given a series of p-values, they couldn’t consider the according p-curve. The usage should be clear throughout the manuscript, otherwise it is too confusing.

RESPONSE: Thank you for this suggestion! We agree that your suggested wording is much more precise. We now use “series of p-values” or “pattern of p- values,” rather than “p-curve analysis,” where appropriate.

The authors mention the study by Field et al. (2020) three times, but only very roughly. However, this seems to be a highly crucial study and of very high importance for the authors’ study, since Field et al. (2020) already investigated the trustworthiness of preregistration on research. Given the importance of the Field et al. study, I would like to see the study discussed in more detail. What can be learned from this study already? And I would like the authors to make explicit within the manuscript how their study adds to that and where it is different.

RESPONSE: Thank you for pointing this out. We have now included much more discussion of Field et al, (2020) including differences in our methods and results, as well as similarities in our results. This is primarily on p. 17.

 Also, there is the repliCATS project going on since 2019, that tries to evaluate how readers make judgements about the replicability of research. The authors mention the DARPA initiative only in the discussion, but it would be good to also see it reflected in the intro already. Recently, a preprint came out with the very first results (Wintle et al., 2021; references at the bottom of this review).

RESPONSE: We have added reference to the Wintle et al (2020) paper on page 6-7. We appreciate the suggestion!

The authors also call preregistration “a marker of a higher quality study”, but then cite Claesen et al. (2019) who showed that a preregistration is not necessarily an indicator of higher quality. I would like to see this mentioned in a more nuanced way. It could very well be that a preregistration is UNDERSTOOD as an indicator of higher quality, but that doesn’t mean that it definitely IS.

RESPONSE: We appreciate this comment and we agree that preregistration is UNDERSTOOD as a marker of a higher quality study rather than a guarantee. We have therefore changed the wording in this way on p. 8.

Also, I think the introduction is a bit short when it comes to which populations (researchers, students, professionals) have which knowledge. I think it may be more mixed and heavily depends on the university and environment the participant is in. As the authors wrote, more and more teaching involves the education about questionable research practices and how to counteract them. The authors cited Chopik et al. (2018), but should also have a look at the paper by Sarafoglou et al. (2020). Moreover, there are more and more ReproducibiliTea Journal Clubs, which are bottom-up initiatives educating about Open Science, often attended by students and early career researchers (Orben, 2019). Therefore, individual differences among students could be explained whether their university has such offers or not. I would wish to see this reflected in the introduction.

RESPONSE: We have expanded the section on the three groups (researchers, students and practitioners) on pages 6-7 of the introduction. We thank you for the suggestion to cite Orben, 2019 and Sarafoglou et al. (2020), which we have done.

It is commendable that the authors conducted a power analysis and also reported it, but I’m missing a justification for the effect size. Where does it come from? Was it informed by other studies (e.g., Field et al., 2020), or was it the smallest effect size of interest? Also, how do the authors explain that the Field et al. (2020) study with a little bit smaller sample size was underpowered based on their pilot study, while the authors here calculated a power of .90 for their effect of interest?

RESPONSE: We now provide more information about the effect size in our power analysis. We chose a small to medium effect size because it was the smallest effect size of interest, given our financial situation. This is now explained on p. 11.

Also, I was wondering why the preregistration states that there are 8 abstracts, but the manuscript reports 12?

RESPONSE: Thank you for noticing this discrepancy! The manuscript was incorrect. There were 8 abstracts, as the preregistration states. This has been corrected on page 12.

For the results section, I would wish the results of the ANOVA would also be reported. I have to say that the authors followed their preregistration perfectly; i.e., they wrote that if they feel confident with the linear-mixed effects models, that those are going to be reported. In fact, authors wrote in the manuscript that “because the linear mixed-effects model provided better fit, we only report the mixed-effects analysis here”. I would like to ask that the ANOVA results get reported as well (e.g., in a footnote) for transparency reasons because then the reader can take that information into account as well.

RESPONSE: Thank you! The ANOVA results are now in the footnote on p. 13-14.

Concerning the discussion section, I have to admit that I am not entirely convinced by the arguments why the here reported effect size of the preregistration effect is small. It is argued that it could be because participants might know that preregistration can be used inappropriately (as has been shown by Claesen et al., 2019). But personally, I believe that the things uncovered by Claesen et al. (2019) are not known to many, especially not outside of the field of Open Science. Unfortunately, this cannot be solved by common sense and not by the info the study of this manuscript offers. Therefore, I am unsure about the interpretation of the results. We can see the effects on trustworthiness, but the reasons for them are currently in the dark.

RESPONSE: We agree that the current study cannot shed light on the reasons for people’s judgments and who have now stated as such on p. 19. We also note that for people less familiar with research, a link may have seemed like an indication or more information, regardless of its real purpose as a preregistration. We now discuss these issues on p. 19-20.

Thus, I do not think that the current manuscript offers enough “flesh” or substance to be published in the present form. I think a second follow up study that uses the same experimental manipulation, and then additionally asks the participants a lot of questions for potential reasons afterwards could really shed a lot of light on what’s going on here. What I’m thinking about, e.g., awareness of preregistration/p-hacking, experience with preregistration, belief in preregistration as a tool to combat p-hacking, knowledge about p-curves, and so on and so forth.

RESPONSE: Thank you for the suggestion. Unfortunately, we are unable to do a follow-up study right now because some authors are no longer available and funding is also not currently available. However, we agree that future studies should examine reasoning and beliefs, and we have suggested as such in the discussion on p. 19-20.

 Lastly, I think the title does not entirely convey the content of the study since the focus was not purely on p-hacking but also good research practices (i.e., preregistration).

RESPONSE: We have changed the title of the paper to The Effect of Preregistration and P-Value Patterns on Trust in Psychology and Biology Research

Minor comments:

* Methods: in the participants characteristics table, the column Biology is not really self-explanatory
* RESPONSE: We have changed Table 1 to include “Field: Biology” to make it clear that the column reports on people in the field of Biology.
* Results: Table 3, to me it is entirely unclear what the planned contrasts between the subscript comparison looked like. The explanation does not help, please clarify in the table description. Also, it says Registration instead of Preregistration in there.
* RESPONSE: The typo has been fixed. We have added notes to make it more clear how our labeling in the tables works.
* General: The authors should check their citation manager. For how to cite links, please have a look at the latest APA guidelines. Also, I would recommend to cite the Open Science Collaboration (2015) as such and not just Collaboration (2015).
* RESPONSE: APA style has been checked more thoroughly.
* In the discussion, sometimes single letters appear (such as r and f)
* Spelling errors (such as intellence and practitioner sand student)
* RESPONSE: These have been fixed and we have checked the spelling of the entire document.

References

Sarafoglou, A., Hoogeveen, S., Matzke, D., & Wagenmakers, E. J. (2020). Teaching good research practices: Protocol of a research master course. *Psychology Learning & Teaching, 19*(1), 46-59.

<https://journals.sagepub.com/doi/pdf/10.1177/1475725719858807>

Orben, A. (2019). A journal club to fix science. *Nature, 573*(7775), 465-466.

<https://www.nature.com/articles/d41586-019-02842-8>

Wintle, B., Mody, F., Smith, E., Hanea, A., Wilkinson, D. P., Hemming, V., … & Fidler, F. (2021). Predicting and reasoning about replicability using structured groups. *MetaArXiv*.

<https://osf.io/preprints/metaarxiv/vtpmb/>

From Reviewer 2:

This article is timely, informative, and will make an important contribution to the literature on researcher practices & perceptions of the trustworthiness of published papers. I recommend that it be given a ‘revise and resubmit’ decision addressing the points below. The most critical point to address would be the questions around sample size and your stopping rule - I think this could be most straightforwardly addressed with an additional replication study, but a version without any additional data collection would also be possible in my view (details below).

My comments are bulleted below; the final section describes my review of the materials shared alongside the manuscript.

· I have a comment on the title you’ve given to this manuscript, which is unusual itself, and my advice is also maybe a little unusual: I think you should include that you have this finding “In biology and psychology papers.” This is technically a less general claim, but it would avoid the (wrong) assumption that many people may make that studies like this are only in psychology.

RESPONSE: We agree! Our new title is: The Effect of Preregistration and P-Value Patterns on Trust in Psychology and Biology Research

· I would advice taking another pass with more attention to structure of the introduction; an unfamiliar reader may not understand how preregistration, HARKing, and p-hacking are specifically related to each other and to p-curves. In particular, the paragraph on p-curve is placed after the section on preregistration, instead of with the paragraphs on p-hacking and HARKing. Relatedly, I don’t think it would necessarily be clear to a naive reader that preregistration is both a method for avoiding p-hacking, and then also a separate signal of credibility that you’ll be measuring in your study.

RESPONSE: We have moved the section on p-curves to make it closer to the discussion of the problem. We hope the introduction now flows better., and provides a better justification for our study·

 You might be able to solve some of this by separating this set of topics (roughly pp. 2-5) instead into two parts: Actions that researchers can take (HARKing and p-hacking on one hand, or preregistration on the other, even though it’s an imperfect tool.) And then signals you can see in the paper, namely an unexpected p-curve shape or a statement of preregistration.

RESPONSE: Thank you for this suggestion. We have used this phrasing in the introduction to help readers think about the difference between preregistration and analysis of p-value patterns.

· You mention that you didn’t preregister the exclusions (thank you for including this), but I wasn’t clear for several pages whether you preregistered your own study or not, and how you justified your sample size. I finally found a reference to it, and to the power analysis you ran much later in the methods section! I would recommend that you move these links to earlier in the methods section or end of introduction.

RESPONSE: We now mention our preregistration at the end of the Introduction. Thank you for the suggestion!

· In your preregistration, you stated an intent to collect data from “at least 206 participants, but as many as 300.” There is no information about a stopping rule included in your preregistration. In the end you collected data from 318, and used exclusion criteria you developed post data collection. I also see a reference to excluding or re-classifyiing some data points in order to make group sizes approximately equal. This is a potential threat to the validity of your study. First, the preregistered sample size gives you the flexibility to stop at anywhere between 2/3 and all the way through a data collection process - how did you decide when to end, and had you seen any initial results at that point? The most serious concern here is whether your knowledge of results could have impacted (even unconsciously) the analysis decisions you made. Because of this, I would like to see at least one of the following in a revision:

(1) Give a more complete account of the process that lead to the final sample. This might include e.g. how many batches of participants you ran, how you decided to end data collection overall, and what information you looked at to determine the need for data dropping and re-classification. Then, calculate and report all of your results in two versions: one with the data as you have it, and one with all data as originally coded. Finally, you would need to address any differences that result from these two different versions of your dataset. The stating of all statistical findings of one of these datasets could be ‘demoted’ to a supplemental section, but the main text should discuss their existence and any difference in their interpretation.

(2) Preregister and conduct an exact replication using the group and inclusion criteria you have now developed, with a single intended target size and a more exact stopping rule that can be implemented cleanly given the logistics of your data collection. (There are many possibilities here! This could be anything from intentionally over-collecting based on your empirical observation of how much data is likely to be lost, to running in small batches until the inclusion criteria result in a dataset that is larger than your intended sample size.)

RESPONSE: We chose to use your first suggestion. We tried to be very clear about all exclusions, so that it was evident that leaving these participants in would be a threat to validity. We also changed our grouping to make our groups more equal sizes. We did all of this before looking at the data (except in terms of group sizes). We note above and in the manuscript (on p. 9) that the change was to include master’s level students in the “researcher” group. This might be seen to decreases the level of expertise in the research group making it harder to find evidence of expertise, which is what we preregistered as our expertise hypothesis.

However on your suggestion, we also have now analyzed our data as preregistered and found almost the same results. The only difference was that students (when M.S. students are included with students) distinguished between p-hacked patterns of p-values, which they did not do when students only included undergraduates. These results are now described in the supplementary materials (which is now explained on p. 9, along with a reference to the SM). The data and code for this analysis is also available on OSF.

· A second potential threat to validity is the fact that you preregistered to conduct both linear mixed model and ANOVA analysis, but then to report the LMM only “If the results are the same for both analyses, and we feel confident about the analysis, we will report the linear mixed effects model.” However, in the main paper you reported only the LMM, but no ANOVA results, “because the linear mixed-effects model provided better fit.” I wasn’t sure if ‘fit’ here referred to a later decision that the LMM is a more appropriate analysis for this data, or the fact that the inferential statistics were weaker in the ANOVA. In either case, I would ask for a revision to include a supplement reporting all statistics, and for the main text to discuss this and any resulting differences.

RESPONSE: We now include the RM ANOVA in a footnote on p. 13-14. We note that the RM ANOVA confirmed more of our hypotheses than the LMM, but we reported the LMM in the main manuscript because it seems to better account for random effects.

· In the discussion, a point on inclusivity: you reference participants’ age as related to their progression through school (and thus level of familiarity with statistics) - note that not all 22 year olds are close to the end of an undergraduate degree!

RESPONSE: Thank you for pointing this out! We have removed this sentence--we agree that it is problematic.

· On a small citation note, if you want you can replace the Rogers (2019) Wired article with this preprint(<https://osf.io/preprints/socarxiv/46mnb/>) for a reference to the SCORE project.

RESPONSE: Thank you for directing us towards this paper! We have replaced the Rogers (2019) citation with Alipourfard et al. (2021).

Replicability & materials

Thank you for including the full text of the survey as well as cleaned data & analysis scripts! This is greatly appreciated. A few notes:

· Thank you for including a codebook! One column was opaque to me and not defined in the codebook -DS. Could you please provide a definition?

RESPONSE: The column called DS was to indicate the data source. Because we used filters on Prolific and collects data from different filters into different files, DS was informative about participants’ jobs and student status. You will notice that the (newly posted) raw data comes from files that indicate their source. We have now added this to our codebook, which we also posted separately so it is easier to find.

· Please note that you have a potentially identifiable column (PID - I don’t know whether these are identifiable to participants or not, but in general I would recommend not using the software-generated ID strings.) One line includes a participant’s full name!

RESPONSE: Thank you! We have now removed these columns from all posted data.

· If possible, can you share a version of this data that has not yet been cleaned/reduced to analyzable participants? This will improve the independent reproducibility of your work.

RESPONSE: Yes! We have now added a folder on OSF called “Raw data.” It includes each of the datasets by source.

· Additionally, I was unable to re-run your analysis scripts, because the provided XLSX cannot be read into R the way it looks like your script intends. In particular, the early part of the script refers to columns including Prereg, phacked, and Trust\_score, which are not present in this version. Can you share a version of the dataset that this analysis works over

RESPONSE: Yes! We’re very sorry that we uploaded the wrong data! The correct file has now been uploaded to our OSF page.

[For clarity, I’m asking for two versions of the data: A version as close to raw as possible (redacting identifiable columns), and since your pipeline seems to involve manual work to prep for analysis, a final version of the data that runs with your R script.]

**Editor First Decision**: Accept

April 1, 2022

Dear Clare Conry-Murray,

The reviewers and I have now read your revised manuscript. While all three of us appreciate your the changes you made in response to the valid points raised by the reviewers, the two reviewers came to different conclusions regarding the final decision. After another careful investigation of the manuscript I have chosen to accept the manuscript for publication, conditional on several revisions, which are listed below. Please be aware that, as also pointed out by reviewer 2, there are still several typos and grammatical errors in the manuscript. I would strongly advise giving the manuscript to a person not involved in its writing for a thorough proof-read in addition to the small things we point out below.

Minor revisions:

* Please indicate as a footnote the main differences in the results when including master’s students as students vs. when including them as researchers in addition to referencing the supplementary material to alleviate some of reviewer 1’s concerns.
* Please go into more detail on your exclusion criteria, to respond to reviewer 2:
“You justified the sample size and reclassification of Master’s students more clearly, but I still found the explanations of data removal unclear - it was described as ‘necessary for the validity of the results’ and ‘unusual data patterns’. I also didn’t see columns in the raw data indicating which participants are being included or excluded. Because the number of participants excluded is fairly large, a skeptical reader might wonder if this results in changes to the outcomes (You write that you did not examine the outcomes before making these decisions, so this wouldn’t be a concern about QRP, just the robustness of the finding.)”
* Please rephrase the sentence “However, because the linear mixed-effects model provided better
fit, we report the mixed-effects analysis here.” (p. 13), which was already indicated in the first round of reviews to be irritating. In line with the new results you provide, you could say, for example that you report mixed-effects models, because model comparisons indicated the need for random effects.
* Regarding the report of the LMMs please provide model equations as asked for by reviewer 2:
“You describe a number of LMM, and describe their structure in paragraph form, without quotations and using the ‘x’ character for interactions. I would recommend that you either present these using numbered equations presented on their own line, or else use a math/code font to distinguish references to factors and model structure from prose paragraphs.”
* Some minor things in the text that should be addressed:
	+ p. 14: please cite R and the lme4 package correctly
	+ p. 14: “a chi-square test” please also use the χ symbol here, which you use later in the same sentence.
	+ p. 14: “The then tested a model that […]” please fix this sentence.
	+ p. 14: “by-trail variability” should be “by-trial variability”
	+ p. 16: “was not madke” should be “was not made”
	+ Table 4: “Coghen’s d” should be “Cohen’s d”

Please let me know, whether you are willing and able to include these revisions into your manuscript. I look forward to receiving your final revision and accepting it for publication in Collabra: Psychology.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Martin Schultze

**Reviewer 1**

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

First, I would like to thank the authors for implementing most of my suggestions and I think it improved the manuscript. However, my main point from the first review – that the origins of the observed effects in this study are entirely unknown and remain in the dark – still holds true. Meaning, I still believe the findings in the manuscript are not sufficient to grant publication.

Second, I have to revise a statement in my previous review, that is, that the authors followed their preregistration perfectly. Despite reading the manuscript and the preregistration, I didn’t notice that changes to the sample had been made that weren’t preregistered. Even though it is now made transparent in the revision, I oppose the author’s action to include Master students in the researcher group and I further wonder why this happened, since the authors already suggested two options in their preregistration in case they fail to gather enough data from one group. That is – and I quote from the preregistration – “If we are unable to collect data from 69 participants in each of the three groups, we will exclude that group or combine the post- BA/BS groups and test differences in the remaining two groups. However, if that is the case, we will need at least 85 in each group. We will make these decisions before looking at the data (except we will look at the sample size in each group).“ So, I can only wonder, for what reason did the authors deviate from their preregistration and did not follow one of those two options or tried to collect more data specifically from the researcher group? In addition, I oppose their approach because it not only made the hypothesis “more difficult to confirm” as noted by the authors in the reply letter, it also renders the results less interpretable and meaningful because groups are not that distinct anymore, so what do differences between groups then actually mean?

Taking both points together, I now believe even more that another replication study with strictly predefined groups and a subsequent questionnaire is necessary to establish the findings and shed some light on the reasons driving the effects. Since the authors wrote in their reply letter that another study is not possible, I, unfortunately, have to suggest a rejection.

Last but not least, I wanted to mention that there were many flaws and mistakes in this manuscript, ranging from simple typo and grammar mistakes up to wrong datafiles, supposedly anonymous datasets containing full names, and not disclosed deviations from the preregistration in the first round. Still, after mentioning all of these, the revision still contains typos and grammar mistakes, copy-pasted typos in the supplemental material, incorrect citation, missing number and bracket in a table, referring to the wrong table, etc. and some of these even in newly added parts. What I want to say is that to me as a reviewer this creates a very careless and sloppy impression and feels a little bit disrespectful. I am sure the authors do not want to bring that impression across. We all make mistakes and any of them as a single mistake by accident would not be an issue at all and would be considered normal, but here there are so many that I am wondering whether all authors have proofread the manuscript prior to submission? Or if the responsible supervisor has given constructive feedback to the first author? For future submissions I would suggest to have the manuscript proofread by all contributing authors, maybe even by someone else who was not involved in the study, and especially look for such careless mistakes.

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) | ✔ |  |  |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) | ✔ |  |  |  |  |

**Reviewer 2**

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

Thanks very much to the authors for their response to reviews, and I enjoyed reading it a second time. I was also able to reproduce the results of the primary LMM as reported in the manuscript, thanks to the provided R code. I am recommending and ‘accept’: this version is exceptionally clear about the scope & impact of the data presented. I have two strong recommendations to accompany this decision: one to increase transparency, and one to increase readability.

(1) You justified the sample size and reclassification of Master’s students more clearly, but I still found the explanations of data removal unclear - it was described as ‘necessary for the validity of the results’ and ‘unusual data patterns’. I also didn’t see columns in the raw data indicating which participants are being included or excluded. Because the number of participants excluded is fairly large, a skeptical reader might wonder if this results in changes to the outcomes (You write that you did not examine the outcomes before making these decisions, so this wouldn’t be a concern about QRP, just the robustness of the finding.)

I realize that the reasons for exclusions are likely very obvious, such as nonsense or blank answers, but these decisions should still be stated and justified. Possible solutions would be to describe the criteria ('we dropped responses where participants gave identical answers to all items",“we dropped responses from participants who listed a job title judged to be joking or unreasonable”), and/or to include a separate data file showing the removed data.

(2) You describe a number of LMM, and describe their structure in paragraph form, without quotations and using the ‘x’ character for interactions. I would recommend that you either present these using numbered equations presented on their own line, or else use a math/code font to distinguish references to factors and model structure from prose paragraphs.

Finally, one small typo, you refer to “Coghen’s d” instead of “Cohen’s d” in table 4.

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  |  | ✔ |

**Author Response**
April 14, 2022

Dear Dr. Schultze,

Thank you for the opportunity to revise our manuscript.  We have addressed each of the issues you mentioned:

·      There is now a footnote on the difference between master’s students as student vs researchers on p. 9.

·      Page 8-9 also goes into detail about all data exclusions.

·      We now include your suggested wording to justify using linear mixed effects models (p. 14).

·      We now use Cambia Math font for equations and interactions throughout the Results section.

·      All minor concerns have been fixed, and the manuscript has been proofread very carefully by several readers in additional to all authors.

We appreciate the feedback and we hope you agree that manuscript is improved.

Best,

Clare Conry-Murray

**Editor Final Decision:** Accept

May 20, 2022

Dear Clare,

Please excuse my delayed response. Thank you for the revisions you made on your manuscript, it is now accepted in this version.

Best,
Martin