Recommender comments

1) If the sample size is fixed, there is no use in doing an a-priori power analysis. After all, you are not determining the sample size. Sensitivity analyses seem the better choice (see Lakens, 2022, Sample Size Justification, Collabra, for suggestions (no need to cite it)).

Thank you very much for this recommendation and reference to the paper. We have created sensitivity curves based on power simulations for both hypotheses and included them in the manuscript. Please note that we also revised our analysis strategy for our hypothesis tests which now have a higher sensitivity (see Figure 1).

2) The coding now seems to indicate that no deviation from a preregistered plan is a good thing. However, that is often not true in practice, as preregistrations are often made before the measure, manipulation, etc are fully pilot tested. So, sometimes the data will inform us deviations are an improvement (e.g., transforming data, additional outlier removals, or even just because the preregistration was not very thought through). A deviation can thus lead to better science, and the scoring should reflect this.

We agree that it is important to highlight that deviations can be useful for improving studies. We have therefore added the following sentence in the introduction: "Deviations from the preregistered plan can be useful and necessary for improving studies, however, it is important that such deviations are transparently reported to ensure interpretability." (p. 5, l. 93 f). Additionally, instead of talking only about adherence, we have made it more consistent in the manuscript to write about "adherence and reporting of deviations" to highlight that both adhering as well as deviating and reporting these deviations are acceptable options. Based on a comment by reviewer #1, we also revised our adherence coding and now inspect different types of deviations more closely (see below).

I recommend you carefully examine the additional reviewer comments, and incorporate the recommendations in an improved Stage 1 protocol.

We would like to thank you and the reviewers for your valuable and helpful suggestions. We have addressed and responded to all comments from the reviewers below.

Review #1

Major comments

*1. An assumption running through the manuscript is that preregistrations should maximally constrain researcher degrees of freedom. I agree that this is an important goal of preregistration — constraining RDFs reduces the risk of bias from data-dependent decision-making; however, in my view it is not *necessary* for a preregistration to be useful. A preregistration with little detail can still be useful because it provides transparency. As such, a preregistration that contained barely any detail would still be appropriate and useful, because it would indicate that most research decisions have not been made in advance, and consequently there is a high risk of bias. In practice, I expect most studies fall somewhere along a continuum in the extent to which research decisions can be made in advance, and it might not always be possible to maximally constrain all RDFs (for example, if the researchers have already seen the data, if the purpose of the study is to generate, rather than test hypotheses etc).*

We agree that shorter preregistrations also increase transparency and aim to clearly express in the manuscript that these preregistrations are also useful. We have therefore revised the following section: "In practice, it is not always possible to make all research decisions in advance and thus completely limit RDF, for example, if the focus is on hypothesis generation rather than testing. In these cases, brief preregistrations can already substantially increase transparency by signaling which decisions were made in advance and which were not. Nonetheless, whenever feasible, more extensive and detailed preregistrations may be particularly effective in restricting RDF (as proposed by Wicherts et al., 2016)." (p. 3, l. 40 ff).

*2. I think the concept of "specificity" needs to be clarified and perhaps renamed as "restrictiveness". I recognize that this conceptual scheme is adopted from prior research (e.g., Wicherts and Bakker), but some improvement would be helpful. In paragraph two, the authors state that preregistrations should ideally be "specific (i.e., providing a detailed description), precise (i.e., allowing only one interpretation), and exhaustive (i.e., excluding the possibility of using other methods)." I find this is a bit confusing — "specific" and "precise" are highly synonymous, and the definition in brackets also seem to overlap substantially. Additionally, the manuscript later states that "specificity" will in fact be used not just to mean specific, but also to mean precise and exhaustive (collectively). Not only is this a bit confusing, it also doesn't seem to cover an important aspect of preregistration that the authors are trying to capture, which is the *extent* to which RDFs are constrained ("...investigate the extent to which it restricts RDF...") — that relies not just on being specific/precise, but also on being comprehensive. Perhaps just calling this concept "restrictiveness" would make things clearer?*

Based on your suggestion, we have renamed "specificity" to "restrictiveness". Furthermore, we have removed the ambiguous definition of restrictive preregistrations and instead introduce them with: "Nonetheless, whenever feasible, more extensive and detailed preregistrations may be particularly effective in restricting RDF (as proposed by Wicherts et al., 2016)." (p. 3, l. 44 f).

*I'll also note that "exhaustive" (the idea introduced in previous papers that a preregistration should state everything that the researchers are *not* going to do, as well as what they will do) seems far too strict and impractical to me. It would require researchers to imagine every possible thing they could do and then explicitly say that they won't do it. Not only does that sound time-consuming and unwieldy, its probably unlikely that researchers can imagine every possibility anyway. Wouldn't it be easier to assume that if you do something in your study that you did not pre-specify, then that is a deviation? Perhaps there's a useful distinction between 'additive' deviations (you're doing something that wasn't specified) and 'modifying' deviations (you're changing something that was specified to something else)? The Bakker and Heirene studies seem to back-up my point, most prereg's they examined were not 'exhaustive', and the authors seems to recognize that its unreasonable to expect them to be — the analysis plan states that scores of 3 (awarded for being exhaustive) will be converted to scores of 2. I'd argue to jettison the idea of exhaustive altogether.*

Thank you for this crucial feedback. We acknowledge that assessing "exhaustiveness" is not feasible. Still, we would like to maintain the score of 3 during the coding procedure, as this ensures the highest level of comparability with the original studies, allowing us to conduct sensitivity analyses and potentially revisit these scores in an exploratory manner. For our main analyses, we will recode scores of 3 to 2 as described in the manuscript.

In light of your input, we have furthermore decided to revise our adherence/deviation coding: We will first determine if there was a deviation for each RDF (adherence score: 1 = no deviation present, 0 = deviation present). Subsequently, considering both restrictiveness and adherence scores, we will assign different deviation types using the following scoring procedure (see also Table 2 in the manuscript):
- *No deviation*: Information is provided and the identical in preregistration and article (restrictiveness > 0, adherence = 1)
- *Modifying*: Information about the RDF was given in the preregistration (restrictiveness > 0) and differs between preregistration and article (adherence = 0), for example, different randomization procedures are described in the preregistration and article
- *Additive*: No information about an RDF was provided in the preregistration (restrictiveness = 0), but this information appears in the article (adherence = UP), for example, randomization procedure is not mentioned in the preregistration but described in the article
- *Omitting*: Information about an RDF was included in the preregistration (restrictiveness > 0) but was subsequently omitted in the article (adherence = UA), for example, randomization procedure is described in the preregistration, but not mentioned in the article
- *U*: No information provided in both the preregistration and article (restrictiveness = 0, adherence = UB)
- *NA*: Not applicable

3. Restrictiveness will be assessed using a complex coding scheme. Firstly, there is a list of items such as "Failing to randomly assign participants to conditions" and "Insufficient blinding the participants and/or experiments". The restrictiveness of the preregistration for each of these items is then assessed by assigning a score from 0 to 3. Zero is awarded when noting is specified about the item. One is assigned if RDFs are restricted "to some extent". Two is awarded if RDFs are "completely restricted". And three is awarded when the restrictions are "exhaustive". I've commented on exhaustiveness above. More generally, this coding and scoring scheme seems problematic to me:
3a) It seems super challenging to me to imagine the universe of possible RDFs (see e.g., multiverse and ManyAnalyst studies) and I'm a bit worried about the the arbitrariness/subjectivity of many of the items, e.g., "Failing to conduct a well-founded power analysis" — how is "well founded" being judged? Its also not clear how that particular item relates to RDFs - a clearly specified power analysis could constrain RDFs regardless of whether it is well-founded. The items in Table 1 seem to make a lot of assumptions about the study being evaluated — for example several items are about hypotheses, but what if the study is not testing hypotheses? Additionally, what if the study uses alternative statistical methods that do not require power analysis? (e.g., Bayesian, estimation, or a purely descriptive study).

Thank you for this important remark. Coding purely on the basis of the RDF, as shown in Table 1, would indeed be very subjective. We aim to achieve greater objectivity by using a detailed coding scheme (see online materials, https://doi.org/10.23668/psycharchives.14046) which we have adopted from the authors of the original studies (Bakker et al., 2020; Heirene et al., 2021). For example, the coding for "D6 Failing to conduct a well-founded power analysis" would be as follows:

Is a power analysis reported?
- No --> D6 = 0
- YES but power level used for the power analysis < .8 --> D6 = 1
- YES the effect size estimate used for the power analysis is based on ((a representative

preliminary study or meta-analytical results) OR (set at medium or smaller)) AND (at the same time the power analysis is used to make a sample size decision) --> D6 = 2
- YES like previous AND the text indicates no other power analysis will be included in the paper than this one --> D6 = 3
Additional decision rule: If the authors state that a power analysis was conducted, but don't explicitly state the parameters sufficiently to be able to reproduce it, then we will score this as 0. Minimum required details: alpha, beta, estimated effect size, test used
---
We also agree that many of the RDF would not be applicable if the coded study did not test hypotheses. Therefore, we will only include "empirical studies that include at least one testable hypothesis" (p. 7, l. 131). Additionally, if items are not applicable for the specific study design (e.g., power analyses for Bayesian statistics), we will assign the coding "NA" (i.e., RDF item not relevant to preregistration, see Table 2 in the manuscript).

3b) On top of this complex coding scheme, we have a scoring scheme. Zero is assigned when a particular item is not mentioned — but what if that item is simply not relevant? For example, what if blinding would not make a difference in this particular study design so the researchers don't mention it?

The coding scheme takes a very structured approach here: In the case that blinding is not mentioned, "NA" is assigned. If blinding is mentioned but no exact procedure for it is described, the score 0 is given, if the procedure is not described in detail/a reproducible manner, a score of 1 is given. A score of 2 is assigned if the described procedure is detailed and reproducible, and a 3 if it is additionally stated that no other blinding procedures are used (see coding scheme in the online materials, https://doi.org/10.23668/psycharchives.14046). Thus, by assigning the score NA, the coding protocol also takes into account that such aspects as blinding may not be relevant for all studies (and thus may not be mentioned at all).

One is assigned when RDFs are restricted to "some extent" and two when RDFs are completely restricted. The rules for determining what "to some extent" is seem pretty arbitrary to me and I don't see how one can know when RDFs have been completely restricted — one would need to know the entire possible universe of forking paths to make that determination, which seems impossible.

The decision rules are precisely defined in the coding scheme (see https://doi.org/10.23668/psycharchives.14046). By using these, we aim to make the decisions less arbitrary. We recognize, however, that despite the elaborate coding scheme, no definitive assessment of restrictiveness is possible, as one would indeed need to know the complete garden of forking paths to do so. The coding is therefore only an approximation of the actual restrictiveness, but can hopefully still provide an interesting comparison since both preregistration formats will be coded based on the same rules. Nevertheless, we will include a discussion of this limitation in the Discussion section.

Note also that the difference in amount of detail between a score of 0 and a score of 1 is quite small compared to the difference between a score of 1 and a score of 2 — as the intervals are not equal, is it meaningful to calculate summary scores based on an unweighted mean? Does an unweighted mean also make sense when items will be more or less easier to restrict? Perhaps it would be more meaningful and interpretable to simply report the proportion in each category (e.g., "item one was completely unrestricted in X% of preregistrations,

partially restricted in X% of preregistrations, and strongly restricted in X% of preregistrations).

We agree that calculating a mean based on our data does not make that much sense, and will therefore report the distribution of scores as percentages and in stacked bar plots instead of providing means, SD, medians, min, and max, for the descriptive reports of restrictiveness (see section "Overall Restriction of RDF Through the PRP-QUANT Template", p. 20 ff). Additionally, following the same rationale, we have decided against using mean restrictiveness scores in our hypothesis tests, and will instead use nested Wilcoxon-Mann-Whitney tests (see description of restrictiveness analyses in the section "Data Analysis", p. 18 ff).

4. The power analyses seem to involve somewhat arbitrarily selected numbers — for example in one analysis the effect size is a Cohen's d of 0.5 which appears to have been chosen because Bakker et al. considered that to be a meaningful difference. Its not clear to me why an effect less than 0.5 wouldn't be meaningful. Perhaps it would be more helpful to plot power curves, then we can explore the sensitivity of the design to detect a range of different effect sizes.

We have created sensitivity curves for both hypotheses based on power simulations and included them in the manuscript (see Figure 1). Additionally, we revised our analyses (i.e., we are now using nested Wilcoxon-Mann-Whitney tests instead of calculating mean restrictiveness scores and comparing those with regular Wilcoxon-Mann-Whitney tests). As a result, our analyses now have a higher sensitivity.

5. In the introduction, the authors argue that structured template = better constraint of RDFs. That seems plausible to me. However, it also seems plausible to me that unstructured = better constraint of RDFs because researchers can specify as much as they like without having to comply with a preregistration template which does not fit well with their research design. This is why for my own preregistrations I usually write a detailed protocol and register it, rather than using a template. This also raises the possibility the structured templates work better in some cases (e.g., when the researchers are just starting out with preregistration or when the template is a good fit for their research design), than others. Focusing only on average effects might obscure this somewhat.

This is a valid argument - we agree that in individual cases, and especially at higher levels of experience, preregistrations without templates could restrict RDF to the same or an even greater extent. However, as shown in previous studies, more structured templates seem to be generally better equipped to restrict RDF (Bakker et al., 2020; Toth et al., 2021; Van Den Akker et al., 2023), which we would like to follow up on here.

6. "To assess the risk of bias in reporting within the associated articles, we will evaluate the remaining six RDF proposed by Wicherts et al. (2016)" — its unclear to me how several of these items relate to preregistration or risk of reporting bias, for example: "Failing to assure reproducibility (verifying the data collection and data analysis)" is about reproducibility, "Failing to enable replication (re-running of the study)", is about replication, and "Misreporting results and p values" is about statistical reporting inconsistencies.

We agree that these aspects do not fit well under the term "risk of bias in reporting" and that overall this has less to do with preregistration. Since we really want to focus on

preregistration, we have removed the aspect "risk of bias in reporting" from our study and will instead only focus on the restrictiveness of preregistrations and the adherence to preregistered plans/reporting of deviations in the associated articles.

7. It doesn't look like the coders will be blind to whether the preregistrations they are assessing are PRP-QUANT or OSF, so there is a risk of bias here (especially as one author is an author of PRP-QUANT). I'm not sure if it is possible to blind the preregistrations, if not then the risk of bias should be noted as a limitation. I wonder if it would be possible to deconstruct the preregistrations into units, and do the assessment on the individual units, thus blind to the type of preregistration? The units can then be recombined for analyses.

We agree that there is a risk of bias when coding the PRP-QUANT preregistrations. Even deconstructing the items would not solve this as we will only code the PRP-QUANT preregistrations (the OSF preregistrations have already been coded by Bakker et al.). We will proceed to the best of our knowledge and belief and mention this limitation in the Discussion section.

8. For the PRP-QUANT vs OSF comparison and the reviewed vs non-reviewed comparison, there are a number of serious threats to internal validity arising from the fact that these groups could reasonable differ on a number of relevant variables other than the variables of interest. For example, PRP-QUANT was introduced relatively recently, and many researchers will have more experience with preregistration now; many of those using the OSF Prereg Challenge template were doing so for the first time. It could also be that researchers preferentially submit confirmatory studies relative to exploratory studies for preregistration review, and its easier to restrict RDFs in confirmatory studies. These are just examples, there are likely many more, and none of them needs to fully account for an observed effect to undermine internal validity. The threat is so severe, that I think the combination of known and unknown threats could easily swamp any effect of the feature of interest. Perhaps that's too pessimistic, but minimally I think these issues should be acknowledged in the manuscript (if they cannot be addressed).

Thank you for raising this important point. The distinction between exploratory and confirmatory studies cannot impact our analyses, as both Bakker et al. (2020) and we only include preregistrations that have at least one hypothesis. However, you are right about the other possible confounding variables, which we will discuss in the Limitations section. Additionally, based on your comment and a comment by Reviewer #2, we have decided to include "study type" in our coding. We will report study type for our sample as well as for the Bakker sample to assess their comparability.

9. Although the research aims are stated clearly, its a bit unclear to me what the broader purpose of the study is. I assume that if the authors get the results they expect (PRP-QUANT > OSF) then this will support a recommendation that folks use the former rather than the latter? Perhaps that does make sense, but I'd just note that PRP-QUANT is designed for specifically psychologists. The OSF template certainly had a bunch of psychologists involved in making it, but there were folks from other disciplines too, and it was not only intended for psychologists. So this might be worth bearing in mind.

Based on your comment, we have added a sentence explaining the scope of the PRP-QUANT Template: "In contrast to the OSF Template, whose scope covers various disciplines, the PRP-QUANT Template is specifically tailored to the field of psychology." (p. 4, l. 66 ff).

Additionally, our study aim is not only to assess whether the PRP-QUANT Template is better equipped to restrict RDF, but on a more general level we want to see whether it is worthwhile to continue developing/using more structured and extensive templates. We have added a description of this in the study design table: "If the preregistrations created with the PRP-QUANT format restrict RDF more (i.e., have an overall higher restrictiveness score) compared to the OSF preregistrations sampled by Bakker et al. (2020, support for hypothesis 1), it will be concluded that the PRP-QUANT format is indeed more effective in reducing RDF than the previous format, in the field of psychology. It therefore appears worthwhile to develop/use highly structured templates in the future. [...]" (p. 40, “Interpretation given different outcomes”).

Minor comments

** is it correct to say in the abstract that PRP-QUANT is "comprehensive"? This implies that it covers every possible researcher degree of freedom, which is probably impossible (you'd have to anticipate every possible research decision). Also "they devised a comprehensive coding scheme based on the RDF defined by Wicherts et al. (2016)." Wicherts et al. say in their own paper: "We created a list of 34 researcher DFs, but our list is in no way exhaustive for the many choices that need be made during the different phases of a psychological experiment"*

We replaced "comprehensive" with "extensive" in the relevant places.

** "These findings highlight the positive impact of structured templates on preregistration specificity while also indicating room for further improvement." This is a causal claim, but the findings referred to are all from non-randomised, observational studies; consequently they are prone to bias from confounding and self-selection. Given the uncertainty about these study's findings, the claim should probably be more tentative.*

We have reworded the sentence accordingly: "These findings suggest that structured templates are associated with higher RDF restriction, while also indicating room for further improvement." (p. 4, l. 60 f).

** "identifying missing restrictions" - I assume this means RDFs that could have been restricted but weren't, but its not clear how this is defined operationally.*

We have revised this sentence to offer a more precise example: "Furthermore, we aim to assess whether peer review of preregistrations further restricts RDF (as suggested by Bakker et al., 2020; research question 3), for example, by reviewers identifying gaps in the preregistration and recommending that the authors provide additional information." (p. 5, l. 79 ff). However, this is only one possible process that could take place. Because we will only compare the overall RDF restriction between peer-reviewed and non-peer-reviewed preregistrations, we will be unable to draw conclusions about the specific processes through which restrictiveness is increased in greater detail.

** "...which entails the failure to ensure reproducibility and replicability and misreporting of the preregistration..." - meaning unclear, what is the reproducibility and replicability of a preregistration? What exactly is meant by misreporting?*

The terms "failure to ensure reproducibility and replicability" and "misreporting of the preregistration" were meant as separate items here. We agree that the sentence structure was confusing. Since we removed the aspect "risk of bias in reporting" from our study, this description was also removed.

** "we will conduct a search for PRP-QUANT preregistrations in this repository using the "PRP-QUANT" metadata tag" — could you explain a bit more about what this tag is. Will it reliably identify all instances of the PRP-QUANT template? Or does it e.g., require authors to actively tag their registrations? (which would create further opportunity for selection biases).*

The metadata tag is assigned by trained administrators of the PsychArchives team who check and approve every submission to the repository. For additional certainty, we conducted a keyword search using "prp" and verified the presence of the tag in the found PRP-QUANT preregistrations. Consequently, all PRP-QUANT preregistrations should be reliably identifiable by this tag.

** "empirical studies that include at least one testable hypothesis" - how will this be operationalized?*

This is operationalized by inspecting the item "I3 Hypothesis" in the PRP-QUANT preregistrations. Only if at least one hypothesis is mentioned here, the study is included.

** "We conducted an initial search to validate our search strategy" — this is not really a validation (e.g., we dont know how many PRP-QUANT instances are being missed) and perhaps better described as a feasibility check.*

We have changed this sentence to: "We performed an initial search to assess the feasibility of our search strategy..." (p. 7, l. 136).

** Table 1 — its unclear what the codes refer to — what are T, C, D etc? And it would be super helpful to have a column with the specific coding questions in it so we can see how they correspond to the various RDFs.*

We have included an explanation for the codes in the table note, and added the coding questions to the table (see Table 1 in the manuscript).

** "As an additional measure of specificity, we will count the number of hypotheses specified in the preregistrations and assess their clarity and distinctiveness" - this seems highly subjective, how will it be operationalized?*

We have added a description for our operationalization: "As an additional measure of restrictiveness, we will assess the clarity and distinctiveness of preregistered hypotheses, similar to Heirene et al. (2021). Specifically, we will examine the number of preregistrations where the number of hypotheses differs depending on whether they are interpreted as single or as several linked but autonomous predictions (e.g., in cases where several predicted effects are mentioned within a single statement)." (p. 15, l. 190 ff).

** "First, we will impute missing values using a two-way imputation procedure based on row and column means (see Heirene et al., 2021)." - more detail on that needed in this manuscript I feel*

We have added a more detailed description of our imputation procedure: "First, we will impute missing values using a two-way imputation procedure based on row and column means. Specifically, the overall mean, the mean for each RDF, and the mean for each preregistration will be computed based on available values, and missing values will be imputed using the formula *RDF mean + preregistration mean - overall mean* (Bernaards & Sijtsma, 2000)." (p. 19, l. 256 ff).

** The conflict of interest statement says that "The authors declare that there are no conflicts of interest with respect to the authorship or the publication of this article", but then also states a conflict of interest "Stefanie Mueller was a member of the task force that created the PRP-QUANT Template but has no financial interest in the results of the presented studies."*

We have removed the first sentence from the Conflict of Interest statement. After consideration, we also feel that we should state our affiliation here, which we have added in the statement. It now reads: "Lisa Spitzer and Stefanie Mueller work for the Leibniz Institute for Psychology (ZPID) that distributes the PRP-QUANT Template, and Stefanie Mueller was a member of the task force that created the PRP-QUANT Template. The template is available free of charge, and none of the authors has a financial interest in the results of this study." (p. 32, l. 358 ff).

** "but they vary in the level of detail that is requested." - perhaps clarify here whether the templates *require* certain information be entered, or simply have a text box for each item that can be ignored.*

To express that providing answers to each item is mostly voluntary, we have used the term "prompting for" instead of "outlining": "Preregistration templates, prompting for information to include in the preregistration, can assist researchers in creating such restrictive preregistrations, but they vary in the level of detail that is requested." (p. 3, l. 46 ff).

** "...which were evaluated by external reviewers..." — could you provide a bit more detail here? Were the reviewers specifically asked to evaluate the restrictiveness of the preregistration?*

Reviewers were not explicitly instructed to evaluate restrictiveness, but they conducted an assessment of the methodology and the quality of operationalization, among other criteria. We have edited the description in the manuscript slightly: "To answer this question, we will inspect PRP-QUANT preregistrations that were submitted to ZPID's service PsychLab in order to apply for a free-of-charge data collection. As PsychLab aimed to promote preregistration by offering this incentive for high-quality preregistrations, the submitted preregistrations underwent evaluation by external reviewers prior to acceptance, assessing their 1) originality and incremental value, 2) relationship to the literature, 3) methodology, 4) quality of the questionnaire and definition of research constructs, and 5) implications of the proposed study." (p. 5, 82 ff).

** Looking through Table 1, its not clear to me what happens when an item is not applicable to a particular design*

See above. If a RDF is not applicable, we will code "NA (RDF item not relevant to preregistration)" (see also Table 2 in the manuscript).

Review #2 (Marjan Bakker)

The authors include a power analysis based on the currently available studies. For the second hypothesis, they do this with a sensitivity power analysis. It would be good to do this for the first hypothesis as well (what should the effect size be at least, when alpha is .05, power .8, and the current sample size?) and compare that with the effect sizes found in Heirene and Bakker. It is, of course, a given that we cannot increase the sample sizes as all preregistrations that have used the PRP-QUANT template are already included. Thus, this study might not be able to detect small effects. Nevertheless, it is an important study into this topic, which might be extended in a few years when more studies use these templates.

Thank you very much for this suggestion. We have created sensitivity curves based on power simulations for both hypotheses and included them in the manuscript (see Figure 1). Additionally, we have compared the sensitivity of hypothesis 1 with the effect size found in Bakker et al. (2020). Please note that because we revised our hypothesis tests based on a comment by Reviewer #1, our tests now have a higher sensitivity.

On page 4, the PRP-Quant Template's development is only mentioned shortly. It might be good to extend this a bit more to clarify how the PRP-Quant Template and the OSF preregistration template differ. Does it cover more research parts, and what are these added parts? Or are the questions more guided/specific? Was it developed taking the results of Bakker and Heirene into account? Should we expect an increase of specificity scores on specific items? And how is it currently implemented and used?

We have expanded the description in the manuscript based on your comment: "In 2022, the "Psychological Research Preregistration-Quantitative (PRP-QUANT) Template" was published by a Joint Psychological Societies Preregistration Task Force (Bosnjak et al., 2022). It was developed based on the APA's Journal Article Reporting Standards (JARS, Appelbaum et al., 2018) and previous preregistration templates. In contrast to the OSF Template, whose scope covers various disciplines, the PRP-QUANT Template is specifically tailored to the field of psychology. Compared to previous templates, various items underwent description revisions, some items were divided into smaller sub-questions, and new items were introduced. As the PRP-QUANT Template is very extensive (including overall 45 items) and was specifically designed to prompt for many details and enable precise planning (see Bosnjak et al., 2022), our objective is to investigate whether it can indeed contribute to achieving higher restrictiveness." (p. 4, l. 63 ff).
In addition, here is some more information based on your other questions that we did not include in the manuscript for the sake of brevity:
- *Was it developed taking the results of Bakker and Heirene into account?* The findings of Bakker et al. (2019) were not considered in the development of the PRP-QUANT Template. However, the knowledge gained from our study together with the previous findings can now be used to revise the template in the future.
- *Should we expect an increase of specificity scores on specific items?* We expect higher restrictiveness for various items (e.g., C3: "Correcting, coding, or discarding detail during data collection in nonblinded manner" and A3: "Deciding how to deal with violations of statistical assumptions in an ad hoc manner", since this is not directly inquired in the OSF but in the PRP-QUANT Template). However, since we want to have an overall picture, we will

compare all RDF between the templates.
- *And how is it currently implemented and used?* The PRP-QUANT Template is available through PsychArchives (https://doi.org/10.23668/psycharchives.4584) and the ZPID preregistration platform PreReg (https://prereg-psych.org/). It is available in different formats (text, table, online questionnaire, R Markdown, Jupyter Notebook).

This is also important to discuss possible confounders that might explain differences in specificity (different types of study/different study fields/different time periods). Thus, it would be good to give more information about how well these two sets of preregistration can be compared. Or collect some additional information so that the authors can evaluate the possible influence of these confounders.

We agree that there are likely confounding variables in our study. Based on your suggestion, we have added the variable "study_type" to our coding scheme, and will also code this variable for the OSF preregistrations. We will report the frequencies of different study types for both samples in the "Sample" section to assess their comparability. We also considered including the study field as an extra variable in the coding process, however, we ultimately decided against it due to the possibly very high degree of variance, which would make coding this variable challenging. We will discuss the study type as well as other possible confounding variables in detail in the Limitations Section (see also our comments to Reviewer #1).

Relatedly, their second hypothesis is about the influence of peer review. It is currently unclear to me how and why these preregistrations were peer-reviewed. Thus the part on page 5 should be extended. Again, this is important to evaluate possible confounders in this observational study.

We have extended the description of the peer review process accordingly: "To answer this question, we will inspect PRP-QUANT preregistrations that were submitted to ZPID's service PsychLab in order to apply for a free-of-charge data collection. As PsychLab aimed to promote preregistration by offering this incentive for high-quality preregistrations, the submitted preregistrations underwent evaluation by external reviewers prior to acceptance, assessing their 1) originality and incremental value, 2) relationship to the literature, 3) methodology, 4) quality of the questionnaire and definition of research constructs, and 5) implications of the proposed study. " (p. 5, l. 82 ff).

On page 14, row 241, the authors mention that they will only do the 29 separate tests if they found an overall difference. Given the limited power, wouldn't it be good always to present these results on the individual RDFs?

Thank you for your advice. Based on your comment, we have decided to carry out the follow-up analyses in any case and have adjusted the description in the manuscript accordingly. Please also note that we no longer correct for multiple testing, as we assume after detailed discussion that the restrictiveness in different RDF describes individual constructs (e.g. restrictiveness of the hypotheses, restrictiveness of the sampling plan, etc.).