# Stage 1 Registered Report: Restriction of researcher degrees of freedom through the Psychological Research Preregistration-Quantitative (PRP-QUANT) Template

Lisa Spitzer[1] & Stefanie Mueller[1]

[1] Leibniz Institute for Psychology (ZPID)

## Author note

Lisa Spitzer https://orcid.org/0000-0002-4925-7291

Stefanie Mueller https://orcid.org/0000-0002-3611-6190

We are submitting a Stage 1 Registered Report. To maximize transparency in the further process, we have already formulated the results section and a description of the results in the abstract in past tense, but the analyses of this study have yet to be carried out. The results section is based on dummy/blinded data and, thus, values are nonsensical. To facilitate review, we have highlighted text parts that will be edited in brackets and color. In Stage 2, we will change the tense to past and append discussion and conclusion sections.

**RRs involving existing data at PCI RR**: For our study, we want to compare a new dataset coded using PRP-QUANT preregistrations with existing data from Bakker et al. (2020). We assume a bias level of 3: We have already downloaded the data from Bakker et al. (2020), however, we did not look at them and blinded these datasets to write and test our analysis scripts (the script used for blinding is available in the supplemental material, https://doi.org/10.23668/psycharchives.14107). In addition, we have already downloaded the PRP-QUANT preregistrations that exist to date but will not begin coding until receiving IPA.

**Deleted:** https://doi.org/10.23668/psycharchives.14047

Correspondence concerning this article should be addressed to Lisa Spitzer, Universitaetsring 15, 54296 Trier. E-mail: ls@leibniz-psychology.org

## Abstract

Preregistration can help to restrict researcher degrees of freedom and thereby ensure the integrity of research findings. However, its ability to restrict such flexibility depends on whether researchers specify their study plan in sufficient detail and adhere to this plan. Previous research indicates higher restrictiveness when preregistrations are based on structured versus unstructured template formats, although there is room for further improvement. The planned study aims to build on these findings and investigate the restrictiveness of preregistrations based on the PRP-QUANT Template, an extensive template that aids the preregistration of quantitative studies in psychology. Preregistrations will be sampled from PsychArchives and coded for their level of restrictiveness using the coding scheme of Bakker et al. (2020) and Heirene et al. (2021). We predict that preregistrations based on the PRP-QUANT Template (*N* = [74]) are more restrictive than preregistrations based on the OSF Preregistration Template (*N* = 52, Bakker et al., 2020, hypothesis 1). We will also inspect whether peer review can contribute further to restricting flexibility and predict higher restrictiveness for peer-reviewed (*n* = [27]) than non-peer-reviewed preregistrations (*n* = [47], hypothesis 2), using nested Wilcoxon-Mann-Whitney tests. Additionally, we will examine adherence to the preregistered plans in the associated publications (*N* = [17]). [In line/in contrast] to hypothesis 1, PRP-QUANT preregistrations [had significantly/did not have] higher restrictiveness scores than OSF Preregistrations. Moreover, [consistent/inconsistent] with hypothesis 2, peer-reviewed preregistrations [had significantly/did not have] higher restrictiveness than non-peer-reviewed ones. […] percent of the associated articles included undeclared deviations. We discuss the implications of our findings for the PRP-QUANT Template and structured templates in general.

*Keywords:* preregistration, open science, meta-research, reproducibility, replicability

## Introduction

While conducting studies, researchers hold a substantial degree of flexibility in decision-making, often referred to as researcher degrees of freedom (RDF, Simmons et al., 2011; see Huntington-Klein et al., 2021 for an illustration). This flexibility can potentially compromise the validity of findings and drawn conclusions, especially in the event of data-driven decisions or other forms of exploitation (Simmons et al., 2011).

Preregistration, the practice of publishing a time-stamped research plan prior to data collection or analysis (see Parsons et al., 2022), helps limit RDF by predetermining and transparently disclosing decisions concerning the research process (as argued by Forstmeier et al., 2017; Hardwicke & Wagenmakers, 2023; Wicherts et al., 2016) and allows others to evaluate the severity of the hypothesis test (Lakens, 2019). 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).

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. In their study, Bakker et al. (2020) compared preregistrations created using a structured versus unstructured template format regarding their ability to restrict RDF. The inspected unstructured format was the “Standard Pre-Data Collection Registration” (https://osf.io/9j6d7), which only inquires about whether data have already been collected or examined, leaving other descriptions open. This was compared to the structured format of the

"OSF Preregistration" (formerly "Prereg Challenge Registration", version 4, https://osf.io/jea94) which consists of 26 items more closely assessing the hypotheses, sampling plan, variables, design, and planned analyses. To evaluate the inspected preregistrations' restrictiveness, they devised an extensive coding scheme based on the RDF defined by Wicherts et al. (2016). Based on this, they found better, but not yet exhaustive, restriction of RDF with the structured compared to the unstructured template format (Bakker et al., 2020). Other studies that compared the OSF Preregistration Template with less extensive templates found similar results (Toth et al., 2021; Van Den Akker et al., 2023). These findings suggest that structured templates are associated with higher RDF restriction, while also indicating room for further improvement.

**Restrictiveness of Preregistrations Created With the PRP-QUANT Template**

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.

By inspecting preregistrations created with this template, we aim to investigate the extent to which it restricts RDF and which RDF are more restricted than others (*research question 1*) and compare its restrictiveness to the OSF Preregistration Template inspected by Bakker et al.

(2020; *research question 2*). Because of its level of detail, we predict that preregistrations created with the PRP-QUANT Template restrict RDF more than preregistrations based on the OSF Preregistration Template (*hypothesis 1*).

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. 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. We will compare PRP-QUANT preregistrations that were peer-reviewed as part of this service with PRP-QUANT preregistrations published by authors without any additional review and predict that peer-reviewed preregistrations restrict RDF more than non-peer-reviewed preregistrations (*hypothesis 2*).

## Adherence to the Preregistered Plan and Reporting of Deviations

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. Given the emerging evidence of insufficient disclosure of deviations in research articles (e.g., Chan et al., 2004; Chan et al., 2008; Chen et al., 2019; Claesen et al., 2021; Goldacre et al., 2019; Ofosu & Posner, 2023; Van Den Akker et al., 2023; see TARG Meta-Research Group & Collaborators et al., 2023 for a review), we will inspect the published research articles associated with the sampled PRP-QUANT preregistrations, following the procedure of Heirene et al. (2021) who investigated the restriction of RDF in gambling studies' preregistrations. We aim to descriptively assess the extent to which researchers that used the PRP-QUANT Template adhered to their preregistered plan and how they reported deviations in their articles (*research question 4*).

# Methods

## Transparency Statement

We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. We meet Level 3 of the PCI RR bias control (https://rr.peercommunityin.org/help/guide_for_authors). Our study design is displayed in Table A1 in the appendix. All study materials, including the RMD file underlying this manuscript (https://doi.org/10.23668/psycharchives.14120), analysis scripts (https://doi.org/10.23668/psycharchives.14107), coding schemes (https://doi.org/10.23668/psycharchives.14046), an overview of the preliminary sample, and dummy/blinded data (https://doi.org/10.23668/psycharchives.14045), have been published alongside this manuscript (https://doi.org/10.23668/psycharchives.14119) on PsychArchives. The final data, that is, the list of all included PRP-QUANT preregistrations and coded RDF, will be made available on PsychArchives as a scientific use file after the coding process.

**Deleted:** https://doi.org/10.23668/psycharchives.14056

**Deleted:** https://doi.org/10.23668/psycharchives.14047

**Deleted:** https://doi.org/10.23668/psycharchives.14055

**Deleted:** sample

**Deleted:** , and a separate list of the

**Deleted:** also

**Deleted:** As it is not our intention to judge the quality of individual preregistrations, the list of RDF scores will not include identifying data and its rows will be shuffled (one preregistration corresponds to one row of scores).¶

## Sample

In this observational study, we will consider all existing preregistrations that were created with the PRP-QUANT Template and published in the digital research repository PsychArchives (https://psycharchives.org/). We will conduct a search for PRP-QUANT preregistrations in

PsychArchives using the corresponding metadata tag (“zpid.tags.visible:PRP-QUANT”), since the PRP-QUANT Template is made available through and closely linked to this repository (https://doi.org/10.23668/psycharchives.4584). Additionally, we will inspect all studies conducted via ZPID’s service PsychLab by referring to our internal documentation and conducting a search on PsychArchives (“zpid.tags.visible:PsychLab”).

From all identified preregistrations, we will include those in our coding that are based on the PRP-QUANT Template, are written in English or German, are publicly accessible (i.e., not under embargo), and are empirical studies that include at least one testable hypothesis (see Bakker et al., 2020; Heirene et al., 2021).

To inspect researchers’ adherence to the preregistered plan and reporting of deviations, we will also search for associated publications for all included preregistrations (e.g., by inspecting the PsychArchives record and conducting a Google search using the preregistration DOI).

We performed an initial search to assess the feasibility of our search strategy, yielding a total of $N = 89$ preregistrations, among which $n = 74$ met the eligibility criteria for coding (with $n = 27$ being peer-reviewed, and $n = 47$ non-peer-reviewed). For $n = 17$, we identified associated publications (see supplemental material for an overview of the preliminary sample, https://doi.org/10.23668/psycharchives.14045). We will perform a second search before the start of coding to include any eligible preregistrations and associated articles that may have been published by then.

All included PRP-QUANT preregistrations will be compared to the $N = 52$ OSF preregistrations sampled by Bakker et al. (2020) to test hypothesis 1 (accessible at Veldkamp et al., 2020). Our sample size of $N = 74$ PRP-QUANT preregistrations already surpasses that of

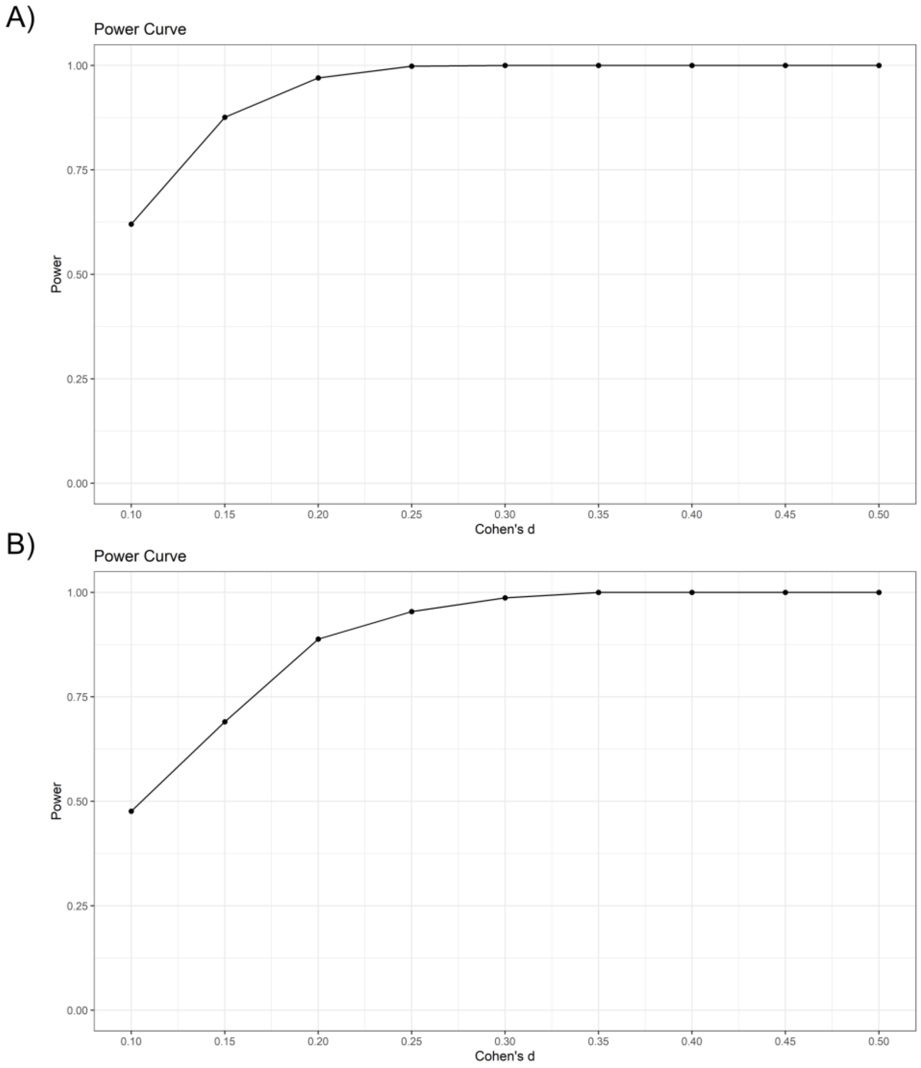
Bakker et al. (2020), which they determined through a power analysis for a Wilcoxon-Mann-Whitney test with α = .05 and a power of .8 to detect a medium effect size of Cohen's *d* = 0.5 (which corresponds to Cliff's *D* of approximately 0.33, Romano et al., 2006), a difference they defined as practically meaningful between two samples of preregistrations. Since our sample size is already determined by the number of available PRP-QUANT preregistrations, we conducted sensitivity analyses for our hypothesis tests (Lakens, 2022). Figure 1A shows a sensitivity curve depicting the relationship between effect size and power for testing hypothesis 1 given our current sample sizes, which was created in R (R Core Team, 2023) based on a power simulation with 1000 repetitions that incorporated the variability in the data from Bakker et al. (2020; see R script in the supplemental material, https://doi.org/10.23668/psycharchives.14107). This curve suggests that we would have a power of .97 to detect small effects of *d* = 0.2 for the overall difference in restrictiveness between templates, employing a nested Wilcoxon-Mann-Whitney test and α = .05. Meanwhile, an effect size of *d* = 0.5 would be detectable with a power above .99. Since the effect size found in Bakker et al. (2020) was even higher (*D* = 0.49, which resembles *d* of about 0.8, Romano et al., 2006), an effect of similar size could therefore also be detected with a high power. However, the difference between two structured templates is likely smaller than that between a structured and an unstructured template.

**Deleted:** ~~https://doi.org/10.23668/psycharchives.14047~~

To test hypothesis 2, we will compare all PRP-QUANT preregistrations that were peer-reviewed as part of PsychLab with the remaining PRP-QUANT preregistrations uploaded directly by researchers to PsychArchives without undergoing external review. For this comparison, the group sizes are limited by the number of available (non-)peer-reviewed preregistrations. However, the sensitivity curve in Figure 1B shows that with the current group sizes of 27 reviewed and 47 non-reviewed preregistrations, we would still have a power of .89 to detect small effects of $d = 0.2$ with α = .05, while an effect size of $d = 0.5$ could be detected with a power above .99.

**Figure 1**

*Sensitivity Curves*



*Note*. Sensitivity curves are provided for A) hypothesis 1 (PRP-QUANT vs. OSF preregistrations) and B) hypothesis 2 (peer-reviewed vs. non-peer-reviewed PRP-QUANT preregistrations). The calculations are based on the preliminary sample sizes. Power simulations were conducted in R (R Core Team, 2023).

[*NOTE: A paragraph describing the final sample, including the preregistrations identified during the second search, will be added here. We will also code the study type of preregistered studies for PRP-QUANT and OSF preregistrations and report the frequencies of different study types in both samples to assess their comparability.*]

## Measures and Coding Procedure

To ensure comparability, we will use the protocols provided by Heirene et al. (2021) which they adapted from Bakker et al. (2020), to code restrictiveness in the PRP-QUANT preregistrations, as well as adherence in their associated articles. These protocols are based on the 34 RDF defined by Wicherts et al. (2016) which encompass flexibility across five key stages: Theorizing, design, collection, analyses, and reporting (see Table 1).

**Table 1**

*Overview of RDF Inspected When Assessing Restrictiveness and Adherence*

| Code | RDF | Restrictiveness question | Adherence question |
|---|---|---|---|
| T1 | Conducting exploratory research without any hypothesis | Is at least one hypothesis specified such that it is clear what are the IV(s) and DV(s)? | Are the hypotheses reported the same as in the preregistration? |
| T2 | Studying a vague hypothesis that fails to specify the direction of the effect | Is the direction of the hypothesis specified? | Is the direction of each hypothesis the same? |
| D1 | Creating multiple manipulated independent variables and conditions | Does the text exclude the possibility that at least one of the manipulated variables will be omitted in the test of the hypothesis?<br>Does it specify exactly how the manipulated variable will be used in the analysis to test the hypothesis? | Are the manipulated independent variables operationalized in the same way as stated in the protocol? |
| D2 | Measuring additional variables that can later be selected as covariates, independent variables, mediators, or moderators | Does it exclude the possibility that at least one other variable (e.g., covariate) is included in the analysis? | Are all variables included in analyses testing hypotheses, consistent with the preregistered analysis plan? |
| D3 | Measuring the same dependent variable in several alternative ways | Does it specify which measurement instrument will be used as the main outcome variable? | Are the dependent variables measured in the same way as stated in the preregistration? |
| D4 | Measuring additional constructs that could potentially act as primary outcomes | Does it specify that the confirmatory analysis section of the paper will not include another DV than the ones specified in all hypotheses? | Are all dependent variables included in analyses reported in the preregistration? |
| D5 | Measuring additional variables that enable later exclusion of participants from the analysis (e.g., awareness or manipulation checks) | Does the preregistration indicate inclusion and exclusion criteria in selecting data points? | Are the criteria for including datapoints in analyses consistent? |
| D6 | Failing to conduct a well-founded power analysis | Is a power analysis reported? | Is the sample size involved in analyses consistent with the outcomes of the power analysis reported in the preregistration? |
| D7 | Failing to specify the sampling plan and allowing for running (multiple) small studies | Is the sampling protocol outlined, including the exact number of participants, recruitment strategy, eligibility criteria, and stopping rules? | Is the sampling protocol stated in the preregistration followed? |

| Code | RDF | Restrictiveness question | Adherence question |
|---|---|---|---|
| C1 | Failing to randomly assign participants to conditions | Is it specified how randomization is implemented? | Is the randomization procedure used consistent with that reported in the preregistration? |
| C2 | Insufficient blinding of the participants and/or experimenters | Does it describe procedures to blind participants to and/or experimenters to conditions? | Is the blinding procedure used consistent with that reported in the preregistration? |
| C3 | Correcting, coding, or discarding data during data collection in non-blinded manner | Does it include protocols concerning coding of data, discarding of cases, or correction of scores during data collection? | Are the procedures used to code and manage data during the data collection process consistent? |
| C4 | Determining the data collection stopping rule on the basis of desired results or intermediate significance testing | Is the sampling protocol outlined, including the exact number of participants, recruitment strategy, eligibility criteria, and stopping rules? (same as D7) | Is the sampling protocol stated in the preregistration followed? (same as D7) |
| A1 | Choosing between different options of dealing with incomplete or missing data on ad hoc grounds | Does it indicate how the study deals with incomplete or missing data? | Are the procedures used to deal with missing data consistent with those reported in the preregistration? |
| A2 | Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, and motion correction) in an ad hoc manner | Does it offer a protocol for pre-processing the data when required (e.g., corrected for motion and other artifacts)? | Are the procedures used to preprocess data consistent? |
| A3 | Deciding how to deal with violations of statistical assumptions in an ad hoc manner | Does it indicate how to test for and deal with violations of statistical assumptions? | Are the procedures used to test for statistical assumptions consistent? |
| A4 | Deciding on how to deal with outliers in an ad hoc manner | Does it indicate how to detect outliers and how they should be dealt with? | Are the procedures used to identify and deal with outliers consistent? |
| A5 | Selecting the dependent variable out of several alternative measures of the same construct | Does it specify which measurement instrument will be used as the main outcome variable? (same as D3) | Are the dependent variables measured in the same way as stated in the preregistration? (same as D3) |
| A6 | Trying out different ways to score the chosen primary dependent variable | Is the method used to measure the primary outcome variable(s) fully described? | Are the dependent variables scored in a way that is consistent? |
| A7 | Selecting another construct as the primary outcome | Does it specify that the confirmatory analysis section of the paper will not include another DV than the ones specified in all hypotheses? (similar to D4) | Are the dependent variables used in primary analyses all the same as reported in the preregistration? |
| A8 | Selecting independent variables out of the set of manipulated independent variables | Does the text exclude the possibility that at least one of the manipulated variables will be omitted in the test of the hypothesis? (similar to D1) | Are the independent variables used in primary analyses all the same? |

| Code | RDF | Restrictiveness question | Adherence question |
|---|---|---|---|
| A9 | Operationalizing manipulated independent variables in different ways (e.g., by discarding or combining levels of factors) | Does it specify exactly how the manipulated variable will be used in the analysis to test the hypothesis? (similar to D1) | Are the manipulated independent variables operationalized in the same way as stated in the protocol? (same as D1) |
| A10 | Choosing to include different measured variables as covariates, independent variables, mediators, or moderators | Does it exclude the possibility that at least one other variable (e.g., covariate) is included in the analysis? (same as D2) | Are all variables included in analyses testing hypotheses, consistent with the preregistered analysis plan? (same as D2) |
| A11 | Operationalizing non-manipulated independent variables in different ways | Are the methods to measure non-manipulated IV(s) fully described? | Are non-manipulated IVs operationalized in a way consistent with the preregistration? |
| A12 | Using alternative inclusion and exclusion criteria for selecting participants in analyses | Does the preregistration indicate inclusion and exclusion criteria in selecting data points? (same as D5) | Are the criteria for including datapoints in analyses consistent? (same as D5) |
| A13 | Choosing between different statistical models | Does it specify the statistical model(s) that will be used to test the hypothesis (e.g., logistic regression)? | Are the statistical tests used to test hypotheses consistent? |
| A14 | Choosing the estimation method, software package, and computation of SEs | Does it indicate details of the estimation technique used to estimate the statistical model and compute standard errors? | Are the estimation techniques used to estimate the statistical model(s) consistent? |
| | | Does it specify which statistical software package and version is used for running the analyses? | Is the statistical software used to conduct analyses consistent with the preregistered plan? |
| A15 | Choosing inference criteria (e.g., Bayes factors, alpha level) | Does it indicate the inference criteria (e.g., Bayes factors, Alpha level)? | Are the inference criteria used consistent? |
| R6 | Presenting exploratory analyses as confirmatory (HARKing) | Does it specify that the confirmatory analysis section of the paper will not include another DV than the ones specified in all hypotheses? (same as A7) | |

*Note.* Questions are abbreviated. The full coding scheme is available in the supplemental material. RDF = Researcher degree of freedom. T = Theorizing. D = Design. C = Collection. A = Analyses. R = Reporting.

For assessing restrictiveness and adherence, we will focus on the RDF that are applicable to preregistrations (cf. Table 1, restrictiveness: T1-A15, R6; adherence: T1-A15). For example, for the RDF “T1: Conducting exploratory research without any hypothesis”, restrictiveness will be coded with the question “Is at least one hypothesis specified such that it is clear what are the IV(s) and DV(s)?”, while adherence will be coded with “Are the hypotheses reported the same as in the preregistration?”.

Overall, 23 questions will be used to code restrictiveness (i.e., there are dependencies in that some questions inform multiple RDF). The coding will be based on the dimensions outlined in Table 2. 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).

Twenty-four questions will be used to code adherence. If an article comprises multiple studies, adherence will be assessed based on the level of preregistrations (i.e., if an article includes two preregistered studies, adherence will be evaluated for each preregistration-article pair). We will distinguish between three types of deviations from preregistration to article: Modifying, additive, and omitting (see Table 2). If the methods presented in the article differ from those outlined in the preregistration, deviations are coded as ‘modifying’. They are labeled as ‘additive’ if the article introduces information not included in the preregistration and as ‘omitting’ if information provided in the preregistration is absent in the associated article. For modifying deviations, we will furthermore examine in more detail whether they were disclosed and justified. The full coding scheme is available in the supplemental material (https://doi.org/10.23668/psycharchives.14046).

**Table 2**

*Scoring of Restrictiveness, Adherence, and Deviation Type*

| Coding | Score | Description |
|---|---|---|
| Restrictiveness | 0 | Not specified: opportunistic use of RDF not restricted at all |
| | 1 | Some specification but lacking details: opportunistic use of RDF is restricted to some extent |
| | 2 | Detailed specification: opportunistic use of RDF is completely restricted, but no explicit statement confirming that authors will not deviate from this plan by adding additional methods/processes |
| | 3* | Detailed specification and statement that authors will not deviate from their plan by adding additional methods/processes: opportunistic use of RDF is completely restricted |
| | NA | RDF item not relevant to preregistration |
| Adherence | 0 | Not consistent with preregistration—deviation |
| | 1 | Consistent with preregistration—no deviation |
| | $U_P$ | Unable to conclusively assess deviations because information is not provided in the preregistration |
| | $U_A$ | Unable to conclusively assess deviations because information is not provided in the article |
| | $U_B$ | Unable to conclusively assess deviations because information is not provided in both the preregistration and article |
| | NA | Not applicable |
| Deviation Type | 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 = $U_P$), for example, randomization procedure is not described in the preregistration but in the article |
| | Omitting | Information about an RDF was included in the preregistration (restrictiveness > 0) but was subsequently omitted in the article (adherence = $U_A$), 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 = $U_B$) |
| | NA | Not applicable |

*Note.* Scores adapted from Heirene et al. (2021). For some RDF, only a subset of restrictiveness scores are possible (see coding scheme in the supplemental material). * Scores of 3 will be coded for comparability with Bakker et al. (2020), but will be recoded to 2, because explicit statements that authors will adhere to their planned methods and avoid additional processes are not common in preregistrations.

Each preregistration will be coded independently by two persons. Inconsistencies will be discussed and solved in pairs. As a measure of inter-coder reliability, a pilot coding phase will be conducted using a randomly selected 10% of the sample. Krippendorff's ɑ will be calculated to assess inter-coder reliability. If ɑ exceeds the threshold of 0.7, the coding process will proceed as planned. If the inter-coder reliability falls below this threshold, the coding protocols and strategies will be revised by discussing ambiguities. [*NOTE: This paragraph will be revised to include the results of the pilot.*]

## Data Analysis

### *R Packages and Scripts*

This manuscript is written with the R package *papaja* (Version 0.1.1.9001, Aust & Barth, 2022). We will use R (Version 4.3.1; R Core Team, 2023) and the R-packages *effsize* (Version 0.8.1; Torchiano, 2020), *irr* (Version 0.84.1; Gamer et al., 2019), *lme4* (Version 1.1.34; Bates et al., 2015), *mice* (Version 3.16.0; van Buuren & Groothuis-Oudshoorn, 2011), *nestedRanksTest* (Version 0.2.9000; Scofield, 2016), *pastecs* (Version 1.3.21; Grosjean & Ibanez, 2018), *psych* (Version 2.3.6; William Revelle, 2023), *RColorBrewer* (Version 1.1.3; Neuwirth, 2022), *tidyverse* (Version 2.0.0; Wickham et al., 2019), and *xfun* (Version 0.39; Xie, 2023) for all our analyses.

Our analysis scripts are based on the scripts provided by Heirene et al. (2021). To adapt and test these, we used a blinded version of the OSF Preregistration data provided by Bakker et al. (2020), where all numbers were replaced with random values within the coding range, and a dummy data set for the coded PRP-QUANT preregistrations. Our analysis scripts (https://doi.org/10.23668/psycharchives.14107), the blinded/dummy data employed for testing them (https://doi.org/10.23668/psycharchives.14045), and the R Markdown file that underlies this

**Deleted:** https://doi.org/10.23668/psycharchives.14047 manuscript – incorporating the code used to generate all outputs displaying the results (https://doi.org/10.23668/psycharchives.14120) – are accessible in the supplemental material.

**Deleted:** https://doi.org/10.23668/psycharchives.14056

### *Preprocessing*

For each preregistration, the responses to the questions in our coding scheme will be translated into restrictiveness scores for each RDF.

Subsequently, we will adjust all restrictiveness scores of 3 to 2 for both the PRP-QUANT and OSF preregistrations. A score of 3 requires an explicit statement from authors that they will adhere to their planned methods and avoid additional processes. Heirene et al. (2021) reported that scores of 3 were rarely achieved due to the scarcity of these explicit statements from the authors and thus suggested this adjustment for future studies. To evaluate the impact of this decision on the results, we will conduct sensitivity analyses by re-running the hypothesis tests with the non-recoded data and reporting differences.

### *Restrictiveness*

To assess the extent to which the PRP-QUANT Template restricts RDF (*research question 1*), we will inspect the distribution of restrictiveness scores of PRP-QUANT preregistrations across all RDF. In addition, stacked bar plots of restrictiveness scores for each RDF are displayed for PRP-QUANT and OSF preregistrations in Figure 2, and for peer-reviewed and non-peer-reviewed PRP-QUANT preregistrations in Figure 3. We will also examine the number of preregistrations where the minimum and maximum number of hypotheses varies when viewed as single versus interconnected but independent predictions, providing means, standard deviations, medians, minimum, and maximum values for both interpretations.

To test our two hypotheses (*research question 2/hypothesis 1*: higher restrictiveness in PRP-QUANT than OSF preregistrations; *research question 3/hypothesis 2*: higher restrictiveness in peer-reviewed than non-peer-reviewed preregistrations), we will largely adopt the methods employed by Bakker et al. (2020) and Heirene et al. (2021). Duplicate information (i.e., RDF based on the same questions as others: C4, A5, A10, A12, R6) will be excluded from these analyses.

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).

To compare the restrictiveness scores between 1) PRP-QUANT and OSF preregistrations, and 2) peer-reviewed and non-peer-reviewed PRP-QUANT preregistrations, we will perform one-tailed nested Wilcoxon-Mann-Whitney tests, using the R package *nestedRanksTest* (Scofield, 2016). The nested ranks test treats the template (PRP-QUANT vs. OSF) as a fixed effect, and the 24 RDF as a random effect. First, group-specific Z-scores are calculated by comparing the ranks between templates. Additionally, distributions of Z-scores are generated by bootstrapping, for which ranks are assigned without considering the template. The Z-scores are then aggregated across groups. Lastly, the *p* value is determined by assessing the percentage of cases where the bootstrapped aggregated Z-score is higher than the observed one (for more information, see Scofield, 2015). To determine significance, a criterion of α = .05 will be applied. Besides these nested tests, we will assess restrictiveness in individual RDF by conducting 24 additional one-tailed Wilcoxon-Mann-Whitney tests for each of the two hypotheses. For these analyses, *p* values will be corrected for multiple tests using the Benjamini-Hochberg correction technique (Benjamini & Hochberg, 1995). As effect size, we will use Cliff's delta (*D*, Cliff, 1993).

**Deleted:** To determine significance, a criterion of α = .05 will be applied. As effect size, we will use Cliff's delta (

### *Adherence*

Adherence to the preregistered plans and reporting of deviations (*research question 4*) will be analyzed descriptively. We will focus on two aspects: The number of preregistration-article pairs with deviations and the total deviations across all pairs. At the level of preregistration-article pairs, we will analyze the number of studies that included modifying, additive, or omitting deviations. We will provide the average number of deviations, along with their corresponding standard deviations, minimum, and maximum values. At the level of total deviations across pairs, we will report percentages and frequencies of different deviation types (see Table 5). For modifying deviations, we will also assess the proportion of justified, unjustified, and nondisclosed deviations.

## Results

[*NOTE: The results section was written based on a generated dummy data set of PRP-QUANT preregistrations and a blinded version of the Bakker et al. (2020) data (i.e., random numbers were generated for each score, the R script used for this generation is available in the supplemental material). Reported scores will be adjusted accordingly after data collection.*]

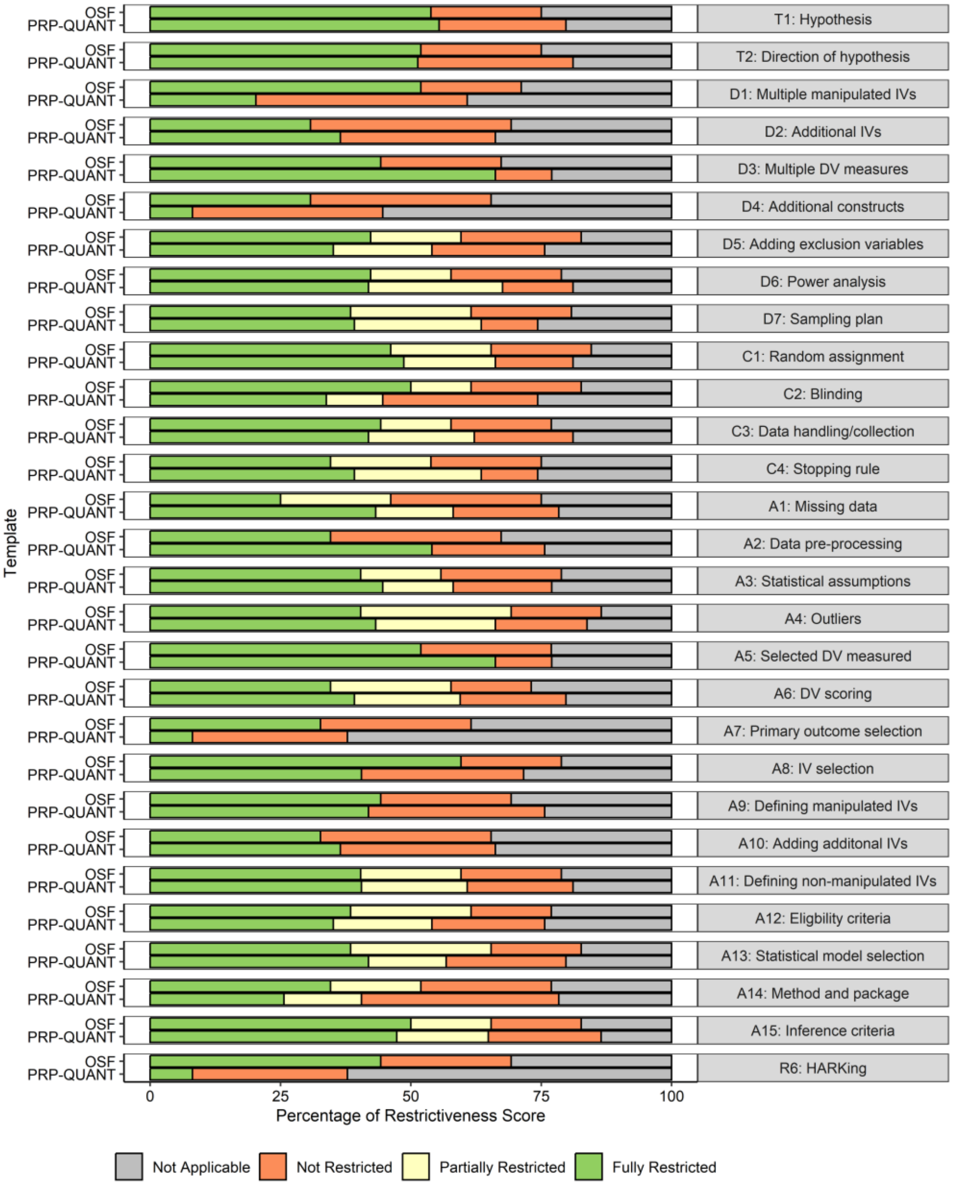
### Restrictiveness

### *Overall Restriction of RDF Through the PRP-QUANT Template*

Across all PRP-QUANT preregistrations, 503 of the 2146 coded RDF were not restricted (23.44%), while 222 were partially restricted (10.34%). For 839 RDF, full restriction according to the used coding scheme was achieved (39.10%). In 582 cases (27.12%), RDF were not applicable for the coded preregistrations. Full restrictiveness was particularly prevalent for […], while […] were often not restricted. The distribution of restrictiveness scores for PRP-QUANT, in comparison with the OSF preregistrations, is displayed in Figure 2.

**Figure 2**

*Distribution of Restrictiveness Scores for PRP-QUANT and OSF Preregistrations*



For 30 preregistrations (40.54%), the hypotheses were not specified clearly. Specifically, the number of hypotheses differed depending on whether they were interpreted as single predictions (*Mean* = 5.62, *SD* = 3.01, *Median* = 5.5, *min* = 1, *max* = 10) or multiple linked but autonomous predictions that could be tested separately (*Mean* = 5.2, *SD* = 2.86, *Median* = 5, *min* = 1, *max* = 10).

### *[Higher/No Higher] RDF Restriction in PRP-QUANT Than OSF Preregistrations*

Our first hypothesis was that preregistrations based on the PRP-QUANT Template constrain RDF more than preregistrations based on the OSF Preregistration Template. [In line with/In contrast to] our hypothesis, the PRP-QUANT preregistrations [had/did not have] a [significantly] higher restrictiveness than the OSF preregistrations, *Z* = -0.04, *p* = .971, $Median_D$ = -0.02. For nine of the 24 tested RDF, restrictiveness was descriptively higher in the PRP-QUANT preregistrations. The difference was statistically significant for two RDF based on the sensitivity of our test, and remained significant in zero cases after correcting for multiple tests (see Table 3). [*NOTE: A short description of which RDF are more restricted in the PRP-QUANT preregistrations will be added.*]

**Deleted:** two

**Deleted:** flexibility was more restricted in PRP-QUANT than in OSF preregistrations (see Table 3). [

A sensitivity analysis showed that recoding the restrictiveness scores from 3 to 2 [did not affect/affected] the results [in that …]. [*NOTE: If the sensitivity analysis shows an influence on the results, it is described in more detail here.*]

**Table 3**

*Comparisons Between PRP-QUANT and OSF Preregistration Restrictiveness Scores for Individual RDF*

| RDF | *W* | *p* | Corrected *p* | *D* | 95% CIs |
|---|---|---|---|---|---|
| T1: Hypothesis | 1,867.00 | .628 | > .999 | -0.03 | -0.21, 0.16 |
| T2: Direction of hypothesis | 1,736.00 | .856 | > .999 | -0.10 | -0.28, 0.09 |
| D1: Multiple manipulated IVs | 956.50 | > .999 | > .999 | -0.50 | -0.66, -0.3 |
| D2: Additional IVs / A10: Adding additional IVs | 1,939.50 | .468 | > .999 | 0.01 | -0.2, 0.21 |
| D3: Multiple DV measures / A5: Selected DV measured | 2,280.00 | .019 | .23 | 0.18 | 0, 0.36 |
| D4: Additional constructs | 1,386.50 | .997 | > .999 | -0.28 | -0.47, -0.06 |
| D5: Adding exclusion variables / A12: Eligibility criteria | 1,807.00 | .729 | > .999 | -0.06 | -0.26, 0.14 |
| D6: Power analysis | 2,176.00 | .094 | .386 | 0.13 | -0.08, 0.33 |
| D7: Sampling plan / C4: Stopping rule | 2,333.50 | .017 | .23 | 0.21 | 0, 0.4 |
| C1: Random assignment | 1,992.00 | .359 | > .999 | 0.04 | -0.16, 0.23 |
| C2: Blinding | 1,568.00 | .968 | > .999 | -0.18 | -0.37, 0.01 |
| C3: Data handling/collection | 2,177.00 | .094 | .386 | 0.13 | -0.07, 0.32 |
| A1: Missing data | 1,697.50 | .887 | > .999 | -0.12 | -0.3, 0.08 |
| A2: Data pre-processing | 1,822.00 | .718 | > .999 | -0.05 | -0.24, 0.14 |
| A3: Statistical assumptions | 2,183.50 | .088 | .386 | 0.14 | -0.07, 0.33 |
| A4: Outliers | 1,954.00 | .438 | > .999 | 0.02 | -0.18, 0.21 |
| A6: DV scoring | 1,869.00 | .614 | > .999 | -0.03 | -0.22, 0.17 |
| A7: Primary outcome selection / R6: HARKing | 1,923.00 | .503 | > .999 | 0.00 | -0.22, 0.22 |
| A8: IV selection | 1,540.00 | .982 | > .999 | -0.20 | -0.38, 0 |
| A9: Defining manipulated IVs | 1,450.00 | .996 | > .999 | -0.25 | -0.42, -0.06 |
| A11: Defining non-manipulated IVs | 1,914.50 | .521 | > .999 | 0.00 | -0.2, 0.2 |
| A13: Statistical model selection | 1,931.00 | .486 | > .999 | 0.00 | -0.19, 0.2 |
| A14: Method and package | 1,805.00 | .733 | > .999 | -0.06 | -0.26, 0.14 |
| A15: Inference criteria | 2,172.00 | .097 | .386 | 0.13 | -0.07, 0.32 |

*Note.* *W* = test statistic of the Wilcoxon-Mann-Whitney test. *D* = Cliff's delta, for which values can range between -1 (all PRP-QUANT preregistrations score lower than all OSF preregistrations) to 1 (all PRP-QUANT preregistrations score higher than all OSF preregistrations). CIs = 95% confidence intervals of effect sizes. Hypothesis tests were conducted with imputed data. *p* values were corrected using the Benjamini-Hochberg method.

### *[Higher/No Higher] Restriction of RDF in Peer-Reviewed Than Non-Peer-Reviewed Preregistrations*

Secondly, we predicted that peer-reviewed PRP-QUANT preregistrations restrict RDF more than non-peer-reviewed preregistrations created with the same format. [Consistent/Inconsistent] with our hypothesis, restrictiveness was [significantly/not] higher for peer-reviewed preregistrations than non-peer-reviewed preregistrations, $Z$ = -0.05, $p$ = .959, $Median_D$ = -0.06. Six of the 24 tested RDF showed a descriptively higher restrictiveness for peer-reviewed preregistrations. For zero RDF, this difference reached statistical significance, which remained significant in zero cases after correcting for multiple tests (see Table 4). [*NOTE: A short description of which RDF are more restricted in the peer-reviewed preregistrations will be added.*] Figure 3 shows the distribution of restrictiveness scores for peer-reviewed and non-peer-reviewed PRP-QUANT preregistrations.

As shown in a sensitivity analysis, recoding the restrictiveness scores from 3 to 2 had [no/an] effect on this analysis [in that …]. [*NOTE: If the sensitivity analysis shows an influence on the results, it is described in more detail here.*]

**Deleted:** 7

**Deleted:** Zero

**Deleted:** benefited from peer review, that is, they showed higher restrictiveness in the peer-reviewed preregistrations

**Table 4**

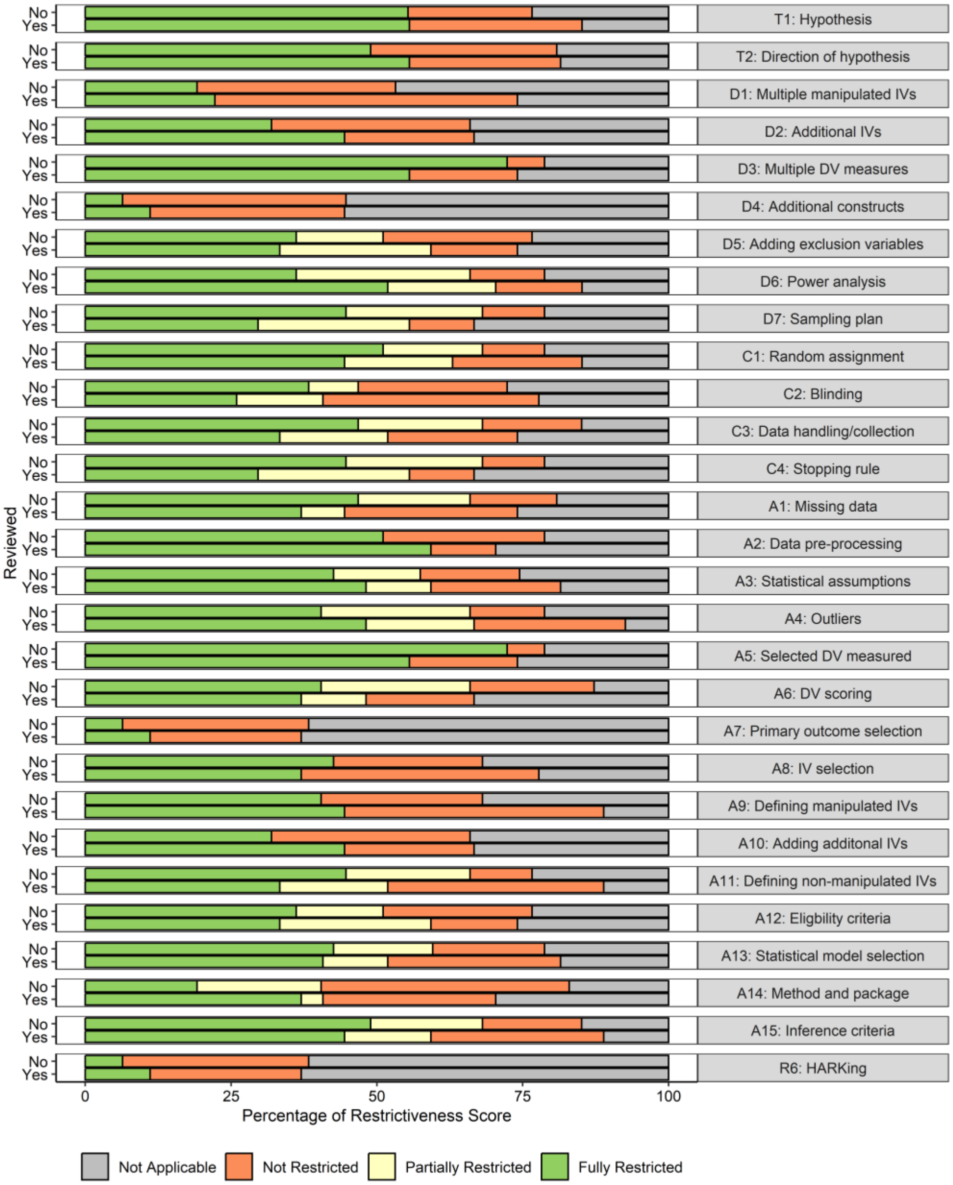
*Comparisons Between Peer-Reviewed and Non-Peer-Reviewed PRP-QUANT Preregistration Restrictiveness Scores for Individual RDF*

| RDF | *W* | *p* | Corrected *p* | *D* | 95% CIs |
|---|---|---|---|---|---|
| T1: Hypothesis | 617.00 | .589 | .966 | -0.03 | -0.28, 0.22 |
| T2: Direction of hypothesis | 679.00 | .295 | .966 | 0.07 | -0.18, 0.31 |
| D1: Multiple manipulated IVs | 548.00 | .845 | .966 | -0.14 | -0.39, 0.14 |
| D2: Additional IVs / A10: Adding additional IVs | 725.00 | .147 | .966 | 0.14 | -0.13, 0.39 |
| D3: Multiple DV measures / A5: Selected DV measured | 453.50 | .992 | .992 | -0.28 | -0.49, -0.05 |
| D4: Additional constructs | 625.50 | .544 | .966 | -0.01 | -0.28, 0.26 |
| D5: Adding exclusion variables / A12: Eligibility criteria | 620.00 | .569 | .966 | -0.02 | -0.28, 0.24 |
| D6: Power analysis | 735.00 | .119 | .966 | 0.16 | -0.11, 0.41 |
| D7: Sampling plan / C4: Stopping rule | 554.00 | .828 | .966 | -0.13 | -0.38, 0.14 |
| C1: Random assignment | 561.00 | .813 | .966 | -0.12 | -0.37, 0.15 |
| C2: Blinding | 521.00 | .907 | .99 | -0.18 | -0.42, 0.09 |
| C3: Data handling/collection | 562.00 | .805 | .966 | -0.11 | -0.36, 0.15 |
| A1: Missing data | 556.00 | .824 | .966 | -0.12 | -0.38, 0.15 |
| A2: Data pre-processing | 732.50 | .115 | .966 | 0.15 | -0.09, 0.38 |
| A3: Statistical assumptions | 631.50 | .517 | .966 | 0.00 | -0.27, 0.26 |
| A4: Outliers | 620.50 | .568 | .966 | -0.02 | -0.29, 0.25 |
| A6: DV scoring | 636.00 | .495 | .966 | 0.00 | -0.26, 0.26 |
| A7: Primary outcome selection / R6: HARKing | 674.00 | .329 | .966 | 0.06 | -0.21, 0.33 |
| A8: IV selection | 556.00 | .825 | .966 | -0.12 | -0.38, 0.15 |
| A9: Defining manipulated IVs | 571.00 | .777 | .966 | -0.10 | -0.36, 0.18 |
| A11: Defining non-manipulated IVs | 469.50 | .974 | .992 | -0.26 | -0.5, 0.02 |
| A13: Statistical model selection | 581.00 | .737 | .966 | -0.08 | -0.34, 0.19 |
| A14: Method and package | 716.00 | .172 | .966 | 0.13 | -0.15, 0.38 |
| A15: Inference criteria | 569.00 | .785 | .966 | -0.10 | -0.36, 0.16 |

*Note.* *W* = test statistic of the Wilcoxon-Mann-Whitney test. *D* = Cliff's delta, for which values can range between -1 (all peer-reviewed preregistrations score lower than all non-peer-reviewed preregistrations) to 1 (all peer-reviewed preregistrations score higher than all non-peer-reviewed preregistrations). CIs = 95% confidence intervals of effect sizes. Hypothesis tests were conducted with imputed data. *p* values were corrected using the Benjamini-Hochberg method.

**Figure 3**

*Distribution of Restrictiveness Scores for (Non-)Peer-Reviewed PRP-QUANT Preregistrations*



## Adherence [*NOTE: Heading might be updated to better present key results*]

In 17 of the preregistration-article pairs (100%), the preregistration, the article, or both were not specified in sufficient detail for completely assessing the adherence between them. For 11.76% of RDF, no information was provided in the preregistration ($U_P$ scores per preregistration-article pair: *Mean* = 3.35, *SD* = 1.8), and for 16.91%, information was lacking in the article ($U_A$ scores: *Mean* = 5.06, *SD* = 1.95). In 11.27% of cases, the information was not provided in both ($U_B$ scores: *Mean* = 3.06, *SD* = 2.25).

Zero of the 17 inspected research articles adhered to their preregistration (0%), that is, followed exactly the procedure described in the preregistration. Meanwhile, 17 displayed modifying deviations (100%). Within this group, 16 articles contained declared deviations. On average, the articles included 1.53 declared and justified deviations (*SD* = 1.59, *min* = 0, *max* = 7), and 1.53 declared but unjustified deviations (*SD* = 1.23, *min* = 0, *max* = 4). In the case of 14 articles, undeclared deviations were present (82.35%), with an average of 1.35 undeclared deviations per article (*SD* = 0.93, *min* = 0, *max* = 3). In addition, 17 articles included additive deviations (100%), that is, information not pre-specified in the preregistration appeared in the article, and 17 articles comprised omitting deviations (100%), meaning that information provided in the preregistration was absent in the article. On average, articles included 3.35 additive (*SD* = 1.8, *min* = 1, *max* = 8) and 5.06 omitting deviations (*SD* = 1.95, *min* = 3, *max* = 9).

Examining the adherence scores across preregistration-article pairs at the level of RDF, it was observed that for 73 RDF, no deviations were present (17.89% of the 408 coded RDF). Meanwhile, a total of 60 modifying deviations were found (14.71%). Out of these, 20 were justified (33.33%) and 21 were not justified (35%). We identified a total of 19 undeclared deviations, which accounted for 31.67% of all modifying deviations (see Table 5).

[Declared/Undeclared] deviations were most common for […]. In addition, we identified 48 additive (11.76%) and 69 omitting deviations (16.91%).

**Table 5**

*Deviation Types Present in the PRP-QUANT Preregistrations by RDF*

| Code | Abbreviated question | No deviation | Modifying | Additive | Omitting | U | NA |
|---|---|---|---|---|---|---|---|
| T1 | Are the hypotheses reported the same as in the preregistration? | 23.53 (4) | 5.88 (1) | 29.41 (5) | 23.53 (4) | 11.76 (2) | 5.88 (1) |
| T2 | Is the direction of each hypothesis the same? | 17.65 (3) | 11.76 (2) | 5.88 (1) | 11.76 (2) | 23.53 (4) | 29.41 (5) |
| D1 | Are the manipulated independent variables operationalized in the same way as stated in the protocol? | 23.53 (4) | 5.88 (1) | 23.53 (4) | 5.88 (1) | 0 (0) | 41.18 (7) |
| D2 | Are all variables included in analyses testing hypotheses, consistent with the preregistered analysis plan? | 17.65 (3) | 5.88 (1) | 17.65 (3) | 5.88 (1) | 11.76 (2) | 41.18 (7) |
| D3 | Are the dependent variables measured in the same way as stated in the preregistration? | 17.65 (3) | 17.65 (3) | 5.88 (1) | 47.06 (8) | 0 (0) | 11.76 (2) |
| D4 | Are all dependent variables included in analyses reported in the preregistration? | 0 (0) | 0 (0) | 17.65 (3) | 0 (0) | 11.76 (2) | 70.59 (12) |
| D5 | Are the criteria for including datapoints in analyses consistent? | 17.65 (3) | 17.65 (3) | 17.65 (3) | 5.88 (1) | 5.88 (1) | 35.29 (6) |
| D6 | Is the sample size involved in analyses consistent with the outcomes of the power analysis reported in the preregistration? | 11.76 (2) | 35.29 (6) | 5.88 (1) | 5.88 (1) | 11.76 (2) | 29.41 (5) |
| D7 | Is the sampling protocol stated in the preregistration followed? | 29.41 (5) | 17.65 (3) | 0 (0) | 0 (0) | 11.76 (2) | 41.18 (7) |
| C1 | Is the randomization procedure used consistent with that reported in the preregistration? | 23.53 (4) | 11.76 (2) | 5.88 (1) | 41.18 (7) | 5.88 (1) | 11.76 (2) |
| C2 | Is the blinding procedure used consistent with that reported in the preregistration? | 23.53 (4) | 5.88 (1) | 11.76 (2) | 11.76 (2) | 17.65 (3) | 29.41 (5) |
| C3 | Are the procedures used to code and manage data during the data collection process consistent? | 23.53 (4) | 35.29 (6) | 17.65 (3) | 5.88 (1) | 0 (0) | 17.65 (3) |
| A1 | Are the procedures used to deal with missing data consistent with those reported in the preregistration? | 17.65 (3) | 5.88 (1) | 11.76 (2) | 17.65 (3) | 17.65 (3) | 29.41 (5) |

| Code | Abbreviated question | No deviation | Modifying | Additive | Omitting | U | NA |
|---|---|---|---|---|---|---|---|
| A2 | Are the procedures used to preprocess data consistent? | 17.65 (3) | 17.65 (3) | 11.76 (2) | 11.76 (2) | 5.88 (1) | 35.29 (6) |
| A3 | Are the procedures used to test for statistical assumptions consistent? | 17.65 (3) | 5.88 (1) | 11.76 (2) | 35.29 (6) | 17.65 (3) | 11.76 (2) |
| A4 | Are the procedures used to identify and deal with outliers consistent? | 23.53 (4) | 23.53 (4) | 5.88 (1) | 29.41 (5) | 5.88 (1) | 11.76 (2) |
| A6 | Are the dependent variables scored in a way that is consistent? | 17.65 (3) | 11.76 (2) | 5.88 (1) | 35.29 (6) | 0 (0) | 29.41 (5) |
| A7 | Are the dependent variables used in primary analyses all the same as reported in the preregistration? | 0 (0) | 0 (0) | 5.88 (1) | 0 (0) | 23.53 (4) | 70.59 (12) |
| A8 | Are the independent variables used in primary analyses all the same? | 23.53 (4) | 23.53 (4) | 5.88 (1) | 23.53 (4) | 5.88 (1) | 17.65 (3) |
| A11 | Are non-manipulated IVs operationalized in a way consistent with the preregistration? | 17.65 (3) | 23.53 (4) | 5.88 (1) | 17.65 (3) | 17.65 (3) | 17.65 (3) |
| A13 | Are the statistical tests used to test hypotheses consistent? | 23.53 (4) | 17.65 (3) | 29.41 (5) | 5.88 (1) | 5.88 (1) | 17.65 (3) |
| A14.1 | Are the estimation techniques used to estimate the statistical model(s) consistent? | 0 (0) | 17.65 (3) | 17.65 (3) | 29.41 (5) | 17.65 (3) | 17.65 (3) |
| A14.2 | Is the statistical software used to conduct analyses consistent with the preregistered plan? | 17.65 (3) | 11.76 (2) | 11.76 (2) | 17.65 (3) | 23.53 (4) | 17.65 (3) |
| A15 | Are the inference criteria used consistent? | 23.53 (4) | 23.53 (4) | 0 (0) | 17.65 (3) | 17.65 (3) | 17.65 (3) |
| | % of total scores (summation) | 17.89 ( 73) | 14.71 ( 60) | 11.76 ( 48) | 16.91 ( 69) | 11.27 ( 46) | 27.45 (112) |

*Note.* Percentage (frequency) of different deviation types made with respect to each RDF. Modifying = RDF was restricted in the preregistration (restrictiveness > 0) and deviation occurred between preregistration and article (adherence = 0). Additive = RDF was not restricted in the preregistration (restrictiveness = 0), but related information was described in the article (adherence = $U_P$). Omitting = RDF was restricted in the preregistration (restrictiveness > 0), but not mentioned in the article (adherence = $U_A$). U = Unable to determine, no information in neither the preregistration nor the article (restrictiveness = 0, adherence = $U_B$). NA = Not applicable. Twenty-four questions were used to code adherence for 29 RDF (i.e., there were some dependencies in that the same questions informed multiple RDF). Duplicate answers were excluded from analyses.

## Authors' Contributions

Conceptualization: L. Spitzer, S. Mueller; Methodology: L. Spitzer, S. Mueller; Software: L. Spitzer; Validation: L. Spitzer; Formal Analysis: L. Spitzer; Investigation: L. Spitzer; Resources: S. Mueller; Data Curation: L. Spitzer, Writing – Original Draft: L. Spitzer; Writing – Review & Editing: S. Mueller; Visualization: L. Spitzer; Supervision: S. Mueller, Project Administration: L. Spitzer

## Conflicts of Interest

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.

## References

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Aust, F., & Barth, M. (2022). *Papaja: Prepare reproducible APA journal articles with R Markdown*. https://github.com/crsh/papaja

Bakker, M., Veldkamp, C. L. S., Assen, M. A. L. M. van, Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), e3000937. https://doi.org/10.1371/journal.pbio.3000937

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bernaards, C. A., & Sijtsma, K. (2000). Influence of Imputation and EM Methods on Factor Analysis when Item Nonresponse in Questionnaire Data is Nonignorable. *Multivariate Behavioral Research*, *35*(3), 321–364. https://doi.org/10.1207/S15327906MBR3503_03

Bosnjak, M., Fiebach, C. J., Mellor, D., Mueller, S., O'Connor, D. B., Oswald, F. L., & Sokol, R. I. (2022). A template for preregistration of quantitative research in psychology: Report of the joint psychological societies preregistration task force. *American Psychologist*, *77*(4), 602–615. https://doi.org/10.1037/amp0000879

Chan, A.-W., Hrobjartsson, A., Jorgensen, K. J., Gotzsche, P. C., & Altman, D. G. (2008). Discrepancies in sample size calculations and data analyses reported in randomised trials: Comparison of publications with protocols. *BMJ*, *337*, a2299–a2299. https://doi.org/10.1136/bmj.a2299

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles. *JAMA*, *291*(20), 2457–2465. https://doi.org/10.1001/jama.291.20.2457

Chen, T., Li, C., Qin, R., Wang, Y., Yu, D., Dodd, J., Wang, D., & Cornelius, V. (2019). Comparison of Clinical Trial Changes in Primary Outcome and Reported Intervention Effect Size Between Trial Registration and Publication. *JAMA Network Open*, *2*(7), e197242. https://doi.org/10.1001/jamanetworkopen.2019.7242

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), 211037. https://doi.org/10.1098/rsos.211037

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. https://doi.org/10.1037/0033-2909.114.3.494

Forstmeier, W., Wagenmakers, E., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews*, *92*(4), 1941–1968. https://doi.org/10.1111/brv.12315

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *Irr: Various coefficients of interrater reliability and agreement*. https://CRAN.R-project.org/package=irr

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., Marston, C., Powell-Smith, A., Heneghan, C., & Mahtani, K. R. (2019). COMPare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, *20*(1), 118. https://doi.org/10.1186/s13063-019-3173-2

Grosjean, P., & Ibanez, F. (2018). *Pastecs: Package for analysis of space-time ecological series*. https://CRAN.R-project.org/package=pastecs

Hardwicke, T. E., & Wagenmakers, E.-J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour*, *7*(1), 15–26. https://doi.org/10.1038/s41562-022-01497-2

Heirene, R., LaPlante, D., Louderback, E. R., Keen, B., Bakker, M., Serafimovska, A., & Gainsbury, S. M. (2021). *Preregistration specificity & adherence: A review of preregistered gambling studies & cross-disciplinary comparison* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/nj4es

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, *59*(3), 944–960. https://doi.org/10.1111/ecin.12992

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *心理学評論*, *62*(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer palettes*. https://CRAN.R-project.org/package=RColorBrewer

Ofosu, G. K., & Posner, D. N. (2023). Pre-Analysis Plans: An Early Stocktaking. *Perspectives on Politics*, *21*(1), 174–190. https://doi.org/10.1017/S1537592721000931

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., … Aczel, B. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behaviour*, *6*, 312–318. https://doi.org/10.1038/s41562-021-01269-4

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Romano, J., Kromrey, J. D., Coraggio, J., Skowronek, J., & Devine, L. (2006). Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate choices? *Annual Meeting of the Southern Association for Institutional Research*.

Scofield, D. G. (2015). Using nestedRanksTest. In *http://cran.nexr.com/*. http://cran.nexr.com/web/packages/nestedRanksTest/vignettes/nestedRanksTest.html

Scofield, D. G. (2016). *Mann-whitney-wilcoxon test for nested ranks*. https://github.com/douglasgscofield/nestedRanksTest

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

TARG Meta-Research Group & Collaborators, Robert T Thibault, Robbie Clark, Hugo Pedder, Olmo van den Akker, Samuel Westwood, Jacqueline Thompson, & Marcus Munafo. (2023). Estimating the prevalence of discrepancies between study registrations and publications: A systematic review and meta-analyses. *medRxiv*, 2021.07.07.21259868. https://doi.org/10.1101/2021.07.07.21259868

Torchiano, M. (2020). *Effsize: Efficient effect size computation*. https://doi.org/10.5281/zenodo.1480624

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., DeHaven, A., Bochantin, J., & Borns, J. (2021). Study Preregistration: An Evaluation of a Method for Transparent Reporting. *Journal of Business and Psychology*, *36*(4), 553–571. https://doi.org/10.1007/s10869-020-09695-3

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Van Den Akker, O., Bakker, M., Van Assen, M. A. L. M., Pennington, C. R., Verweij, L., Elsherif, M. M., Claesen, A., Gaillard, S. D. M., Yeung, S. K., Frankenberger, J.-L., Krautter, K., Cockcroft, J. P., Kreuer, K. S., Evans, T. R., Heppel, F., Schoch, S. F., Korbmacher, M., Yamada, Y., Albayrak-Aydemir, N., … Wicherts, J. M. (2023). *The effectiveness of preregistration in psychology: Assessing preregistration strictness and preregistration-study consistency* [Preprint]. MetaArXiv. https://doi.org/10.31222/osf.io/h8xjw

Veldkamp, C. L. S., Mellor, D. T., Bakker, M., Assen, M. A. L. M. van, Wicherts, J., Nosek, B. A., Ong, H. H., Crompvoets, E. A. V., & Soderberg, C. K. (2020). *Ensuring the quality and specificity of preregistrations* [Data]. OSF. https://osf.io/hbze5

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Van Aert, R. C. M., & Van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

William Revelle. (2023). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Xie, Y. (2023). *Xfun: Supporting functions for packages maintained by ’yihui xie’.* https://CRAN.R-project.org/package=xfun

513 **Appendix**

**Table A1 [NOTE: Table will be updated with the final sample sizes etc. in Stage 2]**

*Study Design, Based on the Template Provided by PCI RR*

| Question | Hypothesis | Sampling Plan | Analysis Plan | Rationale for deciding the sensitivity of the hypothesis test | Interpretation given different outcomes | Theory that could be shown wrong by the outcomes |
|---|---|---|---|---|---|---|
| *Research question 1*: To what extent does the PRP-QUANT Template restrict RDF and which RDF are more restricted than others? | None | We aim to sample all PRP-QUANT preregistrations published on PsychArchives. We will include all preregistrations that meet our inclusion criteria (i.e., preregistrations that are based on the PRP-QUANT Template, are written in English or German, are publicly accessible, are empirical studies, and include at least one testable hypothesis). An initial search identified $N = 74$, to which all other preregistrations published up to the start of coding will be added. | The distribution of restrictiveness scores of PRP-QUANT preregistrations across all RDF will be inspected. In addition, stacked bar plots of restrictiveness scores for each RDF will be displayed for PRP-QUANT and OSF preregistrations, as well as for peer-reviewed and non-peer-reviewed PRP-QUANT preregistrations. We will also examine the number of preregistrations where the minimum and maximum number of hypotheses varies when viewed as single versus interconnected but independent predictions, providing means, standard deviations, medians, minimum, and maximum values for both interpretations. | Descriptive analyses of the PRP-QUANT preregistrations' restrictiveness scores will be used to answer this research question. No hypothesis tests will be conducted. | The results will be reported descriptively. | N/A |

| Question | Hypothesis | Sampling Plan | Analysis Plan | Rationale for deciding the sensitivity of the hypothesis test | Interpretation given different outcomes | Theory that could be shown wrong by the outcomes |
|---|---|---|---|---|---|---|
| *Research question 2*: Are RDF more restricted in preregistrations created with the PRP-QUANT Template, compared to the OSF Preregistration Template studied by Bakker et al. (2020)? | *Hypothesis 1 (primary)*: Preregistrations created with the PRP-QUANT Template restrict RDF more (i.e., have higher restrictiveness scores) than preregistrations based on the format inspected by Bakker et al. (i.e., the OSF Preregistration Template). | All included PRP-QUANT preregistrations (currently $N = 74$) will be compared to the $N = 52$ OSF preregistrations sampled by Bakker et al. (2020). A sensitivity analysis indicates that with the current sample sizes, we would have a power of .97 to detect a small effect size of Cohen's $d = 0.2$, and a power above .99 to detect $d = 0.5$ (which corresponds to Cliff's $D$ of approximately 0.33, Romano et al., 2006). | We will conduct a nested one-tailed Wilcoxon-Mann-Whitney test to compare restrictiveness scores between PRP-QUANT and OSF preregistrations, using the R package *nestedRanksTest* (Scofield, 2016). In this model, template will be treated as a fixed effect and RDF as a random effect. First, group-specific Z-scores are calculated by comparing the ranks between templates. Additionally, distributions of Z-scores are generated by bootstrapping, for which ranks are assigned without considering the template. The Z-scores are then aggregated across groups. Lastly, the $p$ value is determined by assessing the percentage of cases where the bootstrapped aggregated Z-score is higher than the observed one. To determine significance, a criterion of α = .05 will be applied. Additionally, we will conduct 24 more Wilcoxon-Mann-Whitney tests to compare the restrictiveness scores for the individual RDF. For these follow-up tests, $p$ values will be corrected for multiple tests using the Benjamini-Hochberg correction technique. As effect size, we will use Cliff's delta ($D$, Cliff, 1993). | Bakker et al. (2020) determined their sample size of 53 by conducting a power analysis for a Wilcoxon-Mann-Whitney test with α = .05 and a power of .8 to detect a medium effect size of Cohen's $d = 0.5$, which they defined to be a practically meaningful difference between two samples of preregistrations (however, since one preregistration was withdrawn, their final group size was $n = 52$). We will use all PRP-QUANT preregistrations fulfilling our criteria, that is, at least 74. Thus, our sample size already surpasses that of Bakker et al. (2020). Additionally, we will implement a nested Wilcoxon-Mann-Whitney test, resulting in a higher | 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. However, if contrary to our predictions, the PRP-QUANT preregistrations do not have significantly higher | This test is not grounded in a clear-cut theory but is based on the assumption that employing more structured templates is linked to higher restrictiveness, as initially described by Bakker et al (2020). Our objective is to examine whether a template even more structured and detailed than the one previously studied by Bakker et al. (2020) can even better restrict RDF. |

**Deleted:** To determine significance, a criterion of α = .05 will be applied. As effect size, we will use Cliff's delta (

**Deleted:** ¶
¶

| Question | Hypothesis | Sampling Plan | Analysis Plan | Rationale for deciding the sensitivity of the hypothesis test | Interpretation given different outcomes | Theory that could be shown wrong by the outcomes |
|---|---|---|---|---|---|---|
| | | | | power than in the original study. | restrictiveness scores than the OSF ones, we will conclude that there is no evidence that the PRP-QUANT Template achieves a higher level of restrictiveness. We will also further examine for how many of the individual RDF, restrictiveness is higher in PRP-QUANT than OSF preregistrations, and will conclude that the benefit of the PRP-QUANT Template might be most pronounced for all RDF showing significant differences. | |
| *Research question 3*: Can peer review of preregistrations help to restrict RDF? | *Hypothesis 2 (secondary)*: Peer-reviewed preregistrations created with the PRP-QUANT Template restrict RDF more (i.e., have higher restrictiveness scores) than non- | All PRP-QUANT preregistrations that were reviewed will be compared with the remaining non-peer-reviewed PRP-QUANT preregistrations. A sensitivity analysis shows that with the current group sizes | Similar to the analysis of hypothesis 1, we will conduct a one-tailed nested Wilcoxon-Mann-Whitney test to compare the restrictiveness scores between peer-reviewed versus non-peer-reviewed PRP-QUANT preregistrations (procedure is detailed above). Review status will be treated as a fixed effect and RDF as a random effect. To determine significance, a criterion of α = .05 will be applied. Additionally, we will conduct | For this comparison, the group sizes are limited by the number of available (non-)peer-reviewed preregistrations. However, our sensitivity analysis indicates that we will still have high power to detect even | If our analysis reveals that peer-reviewed preregistrations exhibit a higher level of restrictiveness (i.e., have an overall higher restrictiveness score) compared to | This test is also not based on a formulated theory, but rather on the observation made by Bakker et al. (2020) that peer review could potentially have a positive effect on the restrictiveness of |

| Question | Hypothesis | Sampling Plan | Analysis Plan | Rationale for deciding the sensitivity of the hypothesis test | Interpretation given different outcomes | Theory that could be shown wrong by the outcomes |
|---|---|---|---|---|---|---|
| | peer-reviewed preregistrations created with the same format. | of 27 reviewed and 47 non-reviewed preregistrations, we would have a power of .89 to detect small effects of *d* = 0.2 with α = .05, while an effect size of *d* = 0.5 could be detected with a power above .99. | 24 more Wilcoxon-Mann-Whitney tests to compare the restrictiveness scores for the individual RDF. For these follow-up tests, *p* values will be corrected for multiple tests using the Benjamini-Hochberg correction technique. Cliff's delta (*D*, Cliff, 1993) will be used as effect size. | small effects (e.g., a power of .89 to detect effects of *d* = 0.2 with α = .05). | non-peer-reviewed preregistrations (supporting hypothesis 2), we will conclude that peer review is indeed a valuable tool for enhancing the quality of preregistrations, a potential that is currently underused. If we find no significant difference in the overall restrictiveness between peer-reviewed and non-peer-reviewed preregistrations, we will conclude that there is insufficient evidence to support the necessity of peer review for achieving high restrictiveness. As for hypothesis 1, we will also inspect for how many of the individual RDF, restrictiveness is higher in peer-reviewed than non-peer-reviewed preregistrations. | preregistrations. |

**Deleted:** To determine significance, a criterion of α = .05 will be applied. Cliff's delta (

| Question | Hypothesis | Sampling Plan | Analysis Plan | Rationale for deciding the sensitivity of the hypothesis test | Interpretation given different outcomes | Theory that could be shown wrong by the outcomes |
|---|---|---|---|---|---|---|
| | | | | | Based on these analyses, we will conclude that the benefit of peer review for increasing restrictiveness might be most evident for RDF exhibiting significant differences. | |
| *Research question 4*: To what degree do researchers that used the PRP-QUANT Template adhere to their preregistered plan, what deviations occur, and how are these reported? | None | We will search for associated publications for all included preregistrations by examining the PsychArchives record of each preregistration and searching for the preregistration DOI on the Internet (currently identified: *N* = 17, other publications will be searched for until the coding begins). | Researchers' adherence to their preregistered plans and reporting of deviations will be analyzed descriptively. We will focus on two aspects: The number of preregistration-article pairs with deviations and the total deviations across all pairs. At the level of preregistration-article pairs, we will analyze the number of studies that include modifying, additive, or omitting deviations. We will provide the average number of deviations, along with their corresponding standard deviations, minimum, and maximum values. At the deviations level, we will calculate percentages and frequencies of different types of deviations for each RDF and overall, across all preregistration-article pairs, presenting the results in a table. For modifying deviations, we will also assess the proportion of justified, unjustified, and nondisclosed deviations. | Descriptive analyses of the PRP-QUANT preregistrations' adherence and deviation type scores will be used to answer this research question. No hypothesis tests will be conducted. | The results will be reported descriptively. | N/A |