A preregistration of the project

'The effectiveness of preregistration: Assessing preregistration strictness and preregistration-study consistency'

Olmo van den Akker, Marcel van Assen, Marjan Bakker, Shilaan Alzahawi, Gustav Nilsonne, Charlotte Pennington, Alexandra Sarafoglou, Sarah Schoch, Leone Verweij, & Jelte Wicherts

Background

Preregistration has been lauded as one of the key solutions to the replication crisis in the social sciences, mainly because it has the potential to prevent *p*-hacking by restricting researcher degrees of freedom. The effectiveness of preregistration depends on at least two aspects: (1) the strictness of the preregistration (i.e., whether the information provided in the preregistration is comprehensive enough to prevent the opportunistic use of researcher degrees of freedom), and (2) the consistency between the preregistration and the resulting study (i.e., whether the study was carried out in line with the preregistered plan). When a preregistration only contains limited information, or when researchers do not adhere to the preregistered plan, preregistration is ineffective.

Empirical evidence on the effectiveness of preregistration in the social sciences is limited. Bakker et al. (2020) found that preregistrations generally do not restrict all relevant researcher degrees of freedom, while Claesen et al. (2020) found that almost all preregistered studies in the journal Psychological Science included undisclosed deviations from the preregistration. Both studies investigated only one of the two aspects of preregistration effectiveness, making it hard to draw conclusions about the effectiveness of preregistration overall. In this project, we will investigate both strictness and preregistration-study consistency for a large sample of published preregistrations in the field of psychology. Aside from this overall assessment, we will also assess how effectively the following study parts are preregistered: the operationalization of the variables, the data collection procedure, the statistical model, inference criteria, inclusion and exclusion criteria, how missing data is handled, how violations of statistical assumptions are handled, and how statistical outliers are handled.

Despite the lack of empirical evidence about the effectiveness of preregistration, the Center for Open Science (COS) has tried to facilitate the uptake of preregistration among researchers by defining so-called preregistration badges that may be issued by journals, alongside badges for open data and open materials. Papers are eligible to earn a Preregistration Badge if they meet a set of four criteria, briefly that: (1) the preregistration constitutes a public date-time stamped registration in an institutional registration system, (2) the preregistration pre-dates the data collection for the study, (3) the preregistered study design corresponds to the actual study design, and (4) papers include a full disclosure of the results in accordance with

the preregistration. In this project, we will also assess whether preregistrations are in fact registered in line with these criteria.

Finally, we will investigate whether preregistration effectiveness has improved over time, whether replication studies are preregistered more effectively than original studies, and whether more comprehensive preregistration templates (i.e., templates that prompt more details and guidance about what to include in the preregistration) yield more effective preregistrations than less comprehensive templates.

Research questions (RQs)

1) Is preregistration effective in restricting researcher degrees of freedom?
    a. Are studies strictly preregistered?
    b. Are studies consistent with their corresponding preregistrations?
2) For which parts of a study is preregistration most effective?
    a. Which study parts are most strictly preregistered?
    b. Which study parts have the highest preregistration-study consistency?
3) Do authors provide explanations for preregistration-study inconsistencies?
    a. If so, what explanations do authors provide for preregistration-study inconsistencies?
4) Are replication studies more effectively preregistered than original studies?
    a. Are preregistrations of replication studies stricter than preregistrations of original studies?
    b. Are replication studies more consistent with their preregistration than original studies?
5) Are studies based on more comprehensive preregistration templates more effectively preregistered?
    a. Are preregistrations based on more comprehensive templates stricter than preregistrations based on less comprehensive templates?
    b. Are studies based on more comprehensive preregistration templates more consistent with their preregistration than studies based on less comprehensive preregistration templates?
6) Has preregistration effectiveness improved over time?
    a. Has preregistration strictness improved over time?
    b. Has preregistration-study consistency improved over time?
7) Are preregistrations registered in line with the criteria for earning a preregistration badge?
    a. Do preregistrations constitute public date-time stamped registrations in an institutional registration system?
    b. Do preregistrations pre-date the data collection for the study?
    c. Do preregistered study designs correspond to the reported study designs? See RQ 1b and RQ 2b.

d. Do papers include a full disclosure of the results in accordance with the preregistration? See RQ1 in the preregistration of another study (https://osf.io/z4awv)

We do not have specific hypotheses regarding research questions 1, 2, 3, and 7. The hypotheses regarding research questions 4, 5, and 6 are specified below.

Hypotheses and their rationales

1) Replication studies are more effectively preregistered than original studies (RQ4)
    a. Preregistrations of replication studies are stricter than preregistrations of original studies (RQ4)
    b. Replication studies are more consistent with their preregistration than original studies (RQ4)
2) Studies based on more comprehensive preregistration templates are more effectively preregistered than studies based on less comprehensive preregistration templates (RQ5)
    a. Preregistrations based on more comprehensive templates are stricter than preregistrations based on less comprehensive templates (RQ5)
        i. Preregistrations based on the OSF Prereg Template are stricter than preregistrations based on the AsPredicted template
        ii. Preregistrations based on the OSF Prereg Template are stricter than preregistrations based on the template for Pre-registration in Social Psychology
    b. Studies based on more comprehensive preregistration templates are more consistent with their preregistration than studies based on less comprehensive preregistration templates (RQ5)
        i. Studies based on the OSF Prereg Template are more consistent with their preregistration than studies based on the AsPredicted template
        ii. Studies based on the OSF Prereg Template are more consistent with their preregistration than studies based on the template for Pre-registration in Social Psychology
3) Preregistration effectiveness has improved over time (RQ6)
    a. Preregistration strictness has improved over time (RQ6)
    b. Preregistration-study consistency has improved over time (RQ6)

Hypothesis 1a is based on the idea that available information about the primary (to-be-replicated) study nudges researchers to specify more study details in the preregistration of the replication study, making such preregistrations stricter than preregistrations of original studies.

Hypothesis 1b is based on the idea that the principal goal of a replication study is to mimic the primary study. Given that the details of this primary study are specified in the

preregistration of the replication study, researchers doing replication studies are more likely to adhere to the preregistration than researchers doing original studies.

Hypotheses 2a.i and 2a.ii are based on the idea that more comprehensive preregistration templates nudge researchers to specify more study details in preregistrations, making them stricter than preregistrations of researchers using less comprehensive templates.

Hypotheses 2b.i and 2b.ii are based on the idea that researchers using a more comprehensive preregistration template value restricting researcher degrees of freedom more than researchers using a less comprehensive preregistration template and are therefore more likely to adhere to the preregistration more closely.

Hypotheses 3a and 3b are based on the idea that researchers are getting more familiar and more experienced with preregistration and are therefore becoming more effective at (a) making their preregistrations stricter and (b) ensuring preregistration-study consistency.

Sample of published preregistrations in psychology

We used two main sources to find published preregistrations. First, we looked at published papers that earned a Preregistration Challenge prize. The Preregistration Challenge was an educational campaign organized by the Center for Open Science (COS) in 2017 and 2018 where researchers could earn $1,000 if they published a study that was preregistered using a specific preregistration template (see https://cos.io/our-services/prereg-more-information for more information). A full list of Preregistration Challenge prizewinners (N = 180) can be found at https://www.zotero.org/groups/479248/osf/items/collectionKey/D77RMN4N.

Second, we looked at published papers that earned a Preregistration Badge in 2019 or before as part of the COS' Open Science Badges initiative (see https://cos.io/our-services/open-science-badges for more information). Papers are eligible to earn a Preregistration Badge if they meet a set of criteria (i.e., that a public time-stamped preregistration was made before data collection, and results are reported comprehensively and in accordance with the preregistered plan, see https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges). Journals decide themselves whether to check that papers claimed to be eligible for a preregistration badge meet the criteria, or whether to rely on researchers' self-report only. Preregistration + Analysis Plan Badges can be awarded if the preregistration also included a plan for the statistical analyses in the proposed study. We extracted 193 papers that earned a Preregistration Badge and 51 papers that earned a Preregistration + Analysis Badge in 2019 or before from a database with all papers that earned an Open Science Badge per 21 February 2020 (Kambouris et al., 2020).

We identified 26 papers in our sample that earned both a Preregistration Challenge prize and a Preregistration (+ Analysis Plan) Badge. After deleting these duplicate papers, the total number of papers in our sample was 180 + 193 + 51 - 26 = 398. This initial sample of papers can be found in the fourth sheet of the Excel-file uploaded on https://osf.io/e2bjp.

To assess whether these papers were from the field of psychology we looked up their Research Areas as listed in the Web of Science Core Collection. If the paper was not listed in that database, we categorized the Research Area ourselves based on the journal the paper was published in or the departmental affiliation of the authors. In total, 329 papers were categorized as psychology papers, meaning that only 69 of the published preregistrations in our initial sample were from other fields. This sample of preregistered psychology papers can be found in the third sheet of the Excel-file uploaded on https://osf.io/e2bjp.

The papers in our sample often contained multiple preregistered studies. We consider a study separate from other studies in a paper when that study was based on a different sample of participants. Each of the preregistered studies is coded separately. In total, the 329 papers in our sample included 613 preregistered studies, an average of 1.86 preregistered studies per paper. This sample of preregistered psychology studies can be found in the second sheet of the Excel-file uploaded on https://osf.io/e2bjp.

Of these 613 preregistered studies we omitted 43 studies because they were conducted in a registered report framework (where the studies are peer reviewed before data collection), 52 studies because they were part of a multi-lab paper that did not focus on the individual studies but only on the bigger picture (e.g., Many Labs 2, Klein et al., 2018), 13 studies because we were not able to locate a preregistration, and 8 studies because it was unclear which study was described in which (part of a) preregistration.

Finally, we excluded 13 preregistered studies that were based on secondary data (i.e., data that already existed and was gathered to answer another research question from the one in the study). We excluded these studies because the preregistration of studies using secondary data is qualitatively different from studies using primary data (Weston et al., 2019; Van den Akker et al., 2021) and would therefore require different coding procedures. All our exclusions left us with a final sample of 484 studies from 280 papers. This final sample of studies can be found in the first sheet of the Excel-file uploaded on https://osf.io/e2bjp.

A PRISMA flow diagram outlining the full sample selection procedure can be found at https://osf.io/3qupe.

Measuring preregistration effectiveness (RQ1 and RQ2)

We will use several coding protocols to collect data with regard to preregistration effectiveness. First, we use a protocol to identify the hypotheses in each preregistration-study pair (see https://osf.io/fdmx4). This protocol is also used in another study (see https://osf.io/z4awv) in which we look at selective hypothesis reporting. Using the protocol, coders will copy-paste the text from the preregistration that includes the hypothesis and based on that text extract the variables from each hypothesis (independent variable, dependent variable, mediating variable, moderating variable, control variable). From the hypotheses that are consistent between the preregistration and the published study, we randomly select one from each study to assess preregistration effectiveness for that study. During this process we

excluded hypotheses for which coders could not clearly determine what the relationship between the hypothesis variables was---i.e., an association, effect, moderation, or mediation---as well as univariate hypotheses. We did so because our method for computing preregistration effectiveness requires a hypothesis with at least two variables (see below).

Preregistration effectiveness will be coded using a newly developed protocol that assesses both the strictness of the preregistration (adapted from Bakker et al., 2020) and the consistency between the preregistration and the published study. The static version of this Qualtrics protocol can be found at https://osf.io/dpg3v.

We assess preregistration strictness and preregistration-study consistency by answering questions about five essential parts of the preregistration/study:

1. the operationalization of the independent variable (in case the hypothesis implies a causal link between two or more variables) or the first variable (in case the hypothesis doesn't imply a causal link between two or more variables).
2. the operationalization of the dependent variable (in case the hypothesis implies a causal link between two or more variables) or the second variable (in case the hypothesis doesn't imply a causal link between two or more variables).
3. the data collection procedure
4. the statistical model used
5. the inference criteria used

We selected these study parts because they represent the whole process of testing a hypothesis - study design (operationalization of the variables), data collection, and statistical analysis (model and inference) - and it is thus crucial to restrict researcher degrees of freedom for these study parts particularly.

To assess a preregistration's strictness (RQ1a), we will score the five study parts above based on whether they are described in a specific (all steps that will be taken are described) and precise (each of the described steps allows only one interpretation or implementation) manner (Wicherts, et al., 2016) in the preregistration. When any one part of a preregistration is described in a specific and precise manner that part of the preregistration is said to be producible and is scored with 2 points. When some but not all elements relevant to that part of the preregistration are producible, we award 1 point. And, finally, when a part of the preregistration is not deemed producible at all it is scored with 0 points. One exception is the question about the data collection procedure, for which the protocol asks about two elements: sample size and sampling frame. If either one of these two elements is producible (so not necessarily both) the data collection procedure as a whole is scored with 2 points. We implemented this exception because researchers can choose to preregister either an exact sample size *or* a specific and precise sampling method, both of which would not leave researcher degrees of freedom open and would be producible. After summing all scores on the

five essential parts of the study, the strictest preregistration would score 10 points while the least strict preregistration would score 0 points.

To assess the consistency between a preregistration and a study (RQ1b), we will score whether the description of a study part in the preregistration and the description of the corresponding part in the paper are consistent. However, we will only score those parts of the study that scored 1 point or 2 points on preregistration strictness. A preregistration and a study are considered 'consistent' on any one part only when that part is described such that the researcher's action as promised in the preregistration and the researcher's action as stated in the published papers are equivalent. In the preregistration-study consistency part of the protocol any one part can earn 0 points (inconsistent) or 1 point (consistent), so the maximum preregistration-study consistency score is 5, whereas the minimum score is 0.

To compute preregistration effectiveness (RQ1), we will first multiply the score for preregistration strictness with the score for preregistration-study consistency for each part separately. These multiplied scores will signify how effectively each individual study part was preregistered. We will then sum all of these partial effectiveness scores to get a score that indicates how effectively the study was preregistered as a whole. For example, let us suppose a preregistration-study pair scores as follows for preregistration strictness: 1 point for the operationalization of the independent variable, 2 points for the operationalization of the dependent variable, 1 point for the data collection protocol, and 0 points for the statistical model and inference criteria; and as follows for preregistration-study consistency: 1 point for the operationalizations of the independent and dependent variable, and 0 points for the data collection protocol. The preregistration effectiveness score will then be 1*1 + 2*1 + 1*0 + 0 + 0 = 3. This method results in a maximum score for preregistration effectiveness of 10 and a minimum score of 0.

Aside from the five 'essential' study parts outlined above, we will also score several 'non-essential' parts of a study: the operationalization of the mediating variable / moderating variable / control variable, the inclusion and exclusion criteria, how missing data is handled, how violations of statistical assumptions are handled, and how statistical outliers are handled. We will score these non-essential parts in the same way as the essential parts, but the scores for these parts will not be used to calculate a score for the preregistration/study overall. As such, they will only provide information about preregistration strictness (RQ2a), preregistration-study consistency (RQ2b) and preregistration effectiveness (RQ2) of the individual study parts.

Measuring authors' explanations for preregistration-study inconsistencies (RQ3)

We will check what the most common inconsistencies are for both the essential and non-essential study parts, and using open questions to the coders we will elicit what explanations the authors offer for these inconsistencies. These open questions are listed below. A static version of the full protocol can be found at https://osf.io/qmgau.

1. In what way is the [study part] inconsistent?
2. Please copy-paste the authors' explanation for the inconsistency. If the authors do not provide an explanation, please fill out the letter 'n'. To find the authors' explanation you may find it helpful to use the search terms "deviat", "discrep", and "inconsist".

Measuring whether a hypothesis is part of a replication study (RQ4)

We measure whether a hypothesis is part of a replication study or an original study by first searching the preregistration and paper for the string "replic" and assessing whether the authors refer to the study involving the hypothesis as a replication. If they do, in either the preregistration or the paper, we code the hypothesis as a replication hypothesis. If they do not, we code the hypothesis as an original hypothesis.

For replication hypotheses, we additionally check the contents of the paper to see whether they are part of a direct replication or conceptual replication. We code the hypothesis as part of a direct replication when the authors use the same methods (materials *and* procedure) to test the hypothesis as in a prior study. The methods have to be truly identical except that the replication study uses a different sample and except for any translations of study materials. If the methods are not identical in this way, we code the hypothesis as part of a conceptual replication. If our assessment categorizes at least 20% of all replication hypotheses as direct, the variable *replic* in our statistical models will contain three levels (0 = original hypothesis, 1 = conceptual replication hypothesis, 2 = direct replication hypothesis). If our assessment categorizes less than 20% of all replication hypotheses as direct, we will lump conceptual and direct replication hypotheses together and the variable *replic* will contain only two levels (0 = original hypothesis, 1 = replication hypothesis).

A static version of the protocol used to assess whether a hypothesis is part of a direct replication can be found at https://osf.io/cen93. This protocol is also used for a project investigating selective hypothesis reporting (see https://osf.io/5x8ca).

Measuring the comprehensiveness of preregistration templates (RQ5)

To identify the preregistration templates used to draft preregistrations we searched papers for the keyword "regist" to find the link to the preregistration. We then looked at the preregistration link and the surrounding paragraph to identify any references to a preregistration template. If there were no such references, we looked at the preregistration itself to identify which template had been used.

The three preregistration templates with the highest frequency were scored on their comprehensiveness (i.e., their potential to restrict researcher degrees of freedom) using a newly developed protocol where we assessed whether the template includes a prompt, additional instructions, and an example for nine important study elements (see https://osf.io/rtuvb for the filled-out protocol). The maximum possible comprehensiveness score using this protocol is 27 (all nine study elements are included, including additional

instructions and an example). Scoring was done by two independent coders (OA and CP) who resolved three initial coding discrepancies among each other. For one discrepancy an independent third coder (MB) made the final call. Table 1 provides an overview of the preregistration templates we identified, their frequency and comprehensiveness score.

Table 1

Template	Freq.	Comprehensiveness
OSF Prereg template (Bowman et al., 2016)	189	24
AsPredicted (https://aspredicted.org)	167	10
Pre-Registration in Social Psychology (Van 't Veer & Giner-Sorolla, 2016)	30	14
OSF's Open Templates (https://osf.io/9j6d7; https://osf.io/haadc)	11	0
Happy Lab Pre-Registration Template (https://osf.io/yvsj8)	7	-
ClinicalTrials.gov Template (https://prsinfo.clinicaltrials.gov/ProtocolDetailedReviewItems.pdf)	4	-
Replication Recipe (Brandt et al., 2013) (https://osf.io/4jd46)	3	-
ELTE Decision Lab Preregistration Form	1	-
TESS Proposal	1	-
Unknown	69	
Total	484	-

Note: The OSF-Standard Pre-Data Collection Registration is combined with OSF's Open-Ended Registration into OSF's Open Templates because they share a minimalistic setup. This minimalistic setup also means they automatically score 0 on comprehensiveness.

Measuring registration dates (RQ6 an RQ7b)

To see whether preregistration effectiveness has increased over time (RQ6) and whether the preregistration was created before data collection (RQ7b) we coded the date the preregistration was formally registered. For frozen registrations on OSF, that information is clearly listed on the right-side of the preregistration document behind the word "registered"). For frozen registrations on AsPredicted, that information is clearly listed on the top of the preregistration document behind the word "public"). For frozen registrations on clinicaltrials.org, we retrieved the information by looking at the Study Record Versions and selecting the last version before data collection began (i.e., before the study became "active"). For frozen registrations on tessexperiments.org, we took the first day of the Field period (the period that the study was going to be conducted). Finally, for non-frozen registrations we used the date at which the preregistration was last modified. The registration dates were recoded to the number of months since the first preregistration in our sample was registered, which was in April 2014 (Van Zant & Moore, 2015).

To answer RQ7b, we extracted from the paper the month and year in which data for the study was collected. To make comparison with the registration date straightforward, we also

recoded this information to the number of months since the first preregistration in our sample was registered.

Measuring whether preregistrations are registered in line with the criteria for earning a preregistration badge (RQ7)

To answer RQ7a, we coded how the preregistrations in our sample (N = 484) were registered. They could either be '*frozen and directly accessible*' (i.e., the paper directly linked to a timestamped and non-editable document at osf.io, aspredicted.org, clinicaltrials.gov, or tessexperiments.org), '*frozen and indirectly accessible*' (i.e., the paper linked to osf.io or another repository, but not to a frozen and non-editable document - finding that document required some searching through the OSF-website), '*non-frozen*' (i.e., the paper linked to osf.io or another repository, but we could only find a document that was editable and/or non-timestamped), or *'inaccessible'* (we were not able to locate the preregistration, or it was unclear which study was described in which (part of a) preregistration).

To answer RQ7b, we will code the registration date of the preregistration and the data collection period reported in the paper. For more details, see the section 'Measuring registration dates'.

To answer RQ7c, we will code the extent to which preregistrations and papers are consistent. For more details, see the section 'Measuring preregistration effectiveness'.

To answer RQ7d, we will make use of data of another study in which we look at selective hypothesis reporting (i.e., the extent to which the results of preregistered hypotheses are included in the final paper) using the same sample of published preregistrations outlined above. For more details, see RQ1 in the preregistration of that study (https://osf.io/z4awv).

Data analysis

Research questions 1 and 2 (RQ1 and RQ2)

To answer RQ1 we will calculate the average scores on preregistration strictness, preregistration-study consistency, and preregistration effectiveness for all the studies in our sample, and to answer RQ2 we will calculate these scores per study part. We will present these results like in Table 2.

Table 2

	Strictness	Consistency	Effectiveness
Independent variable (N=)			
Third variable (N=)			
Dependent variable (N=)			
First control variable (N=)			

Data collection procedure (N=)
Inclusion / exclusion criteria (N=)
Incomplete / missing data (N=)
First statistical model (N=)
Violations of statistical assumptions (N=)
Inference criteria (N=)
Total of essential parts (N=)

Research question 3 (RQ3)

To answer RQ3 we will count the number of times a study part is assessed by the coders as consistent between the preregistration and the paper. For all identified inconsistencies, we will analyze the explanations the authors provided for them in the paper (RQ3a). We will do so for each study part separately. How we will present this data depends on the results, so we will not speak to that in this preregistration.

Research question 4 (RQ4)

To test Hypothesis 2 (RQ4) we will run three multilevel regressions (study is the first level, paper is the second level), one with preregistration strictness, one with preregistration-study consistency, and one with preregistration effectiveness as the dependent variable.

As we mentioned in the section 'Measuring whether a hypothesis is part of a replication study' the main independent variable *replic* will have either two levels or three levels, meaning that the multilevel regressions will include either one dummy (replication vs. original study, RvO) or two dummies (direct replication vs. original study, DvO; and conceptual replication vs. original study, CvO) respectively.

In case the regression coefficient of the RvO dummy is found to be statistically significant ($p < .025$), we will conclude that replication status is associated with the dependent variable in the respective regression (strictness, consistency, or effectiveness). Moreover, we will add columns to Table 2 to present the data for replications and original studies separately. The alpha level of .025 is based on a Bonferroni correction ($\frac{.05}{2} = .025$) where we assume two independent analyses (the regressions involving the variables strictness and consistency). The analysis involving effectiveness is not independent from the other analyses because effectiveness is computed based on the strictness and consistency scores of the different study parts (see the section 'Measuring preregistration effectiveness').

In case the regression coefficient for DvO is found to be statistically significant ($p < .025$), we will conclude that a difference in strictness, consistency, or effectiveness exists between direct replications and original studies. In case the regression coefficient for CvO is found to be statistically significant ($p < .025$), we will conclude that a difference in strictness, consistency, or effectiveness exists between conceptual replications and original studies. Moreover, if either one of these regression coefficients is found to be statistically significant, we will add columns to Table 2 to present the data for direct replications, conceptual replications, and original studies separately.

VIOLATION OF STATISTICAL ASSUMPTIONS / ESTIMATION TECHNIQUE

In R-code (version 3.6.1) the three regressions outlined above look as follows:

```
strictnessReplic <- lmer(strictness ~ replic + (1 | paper), data = PPP)
summary(strictnessReplic)
consistencyReplic <- lmer(consistency ~ replic + (1 | paper), data = PPP)
summary(consistencyReplic)
effectivenessReplic <- lmer(effectiveness ~ replic + (1 | paper), data = PPP)
summary(effectivenessReplic)
```

Research question 5 (RQ5)

For each separate template comparison that is part of Hypothesis 2 (RQ5) we will run three multilevel regressions (study is first level, paper is second level), one with preregistration strictness, one with preregistration-study consistency, and one with preregistration effectiveness as the dependent variable. Replication status will be included as a control variable. For each regression we only include data for the templates that are directly compared. In the below R-code (version 3.6.1) *OSF1* represents the OSF Prereg Template, *AP* represents the AsPredicted template, and *SP* represents the Pre-Registration in Social Psychology template. The template mentioned before the 'vs' in the variable name is coded with a 1, and the template mentioned after the 'vs' in the variable name is coded with a 0.

```
strictness.OSF1vsAP <- lmer(strictness ~ OSF1vsAP + replic + (1 | paper),
data = PPP)
summary(strictness.OSF1vsAP)
consistency.OSF1vsAP <- lmer(consistency ~ OSF1vsAP + replic + (1 | paper),
data = PPP)
summary(consistency.OSF1vsAP)
effectiveness.OSF1vsAP <- lmer(effectiveness ~ OSF1vsAP + replic + (1 |
paper), data = PPP)
summary(effectiveness.OSF1vsAP)

strictness.OSF1vsSP <- lmer(strictness ~ OSF1vsSP + replic + (1 | paper),
data = PPP)
summary(strictness.OSF1vsSP)
consistency.OSF1vsSP <- lmer(consistency ~ OSF1vsSP + replic + (1 | paper),
data = PPP)
summary(consistency.OSF1vsSP)
effectiveness.OSF1vsSP <- lmer(effectiveness ~ OSF1vsSP + replic + (1 |
```

```
paper), data = PPP)
summary(effectiveness.OSF1vsSP)
```

In case a regression coefficient is found to be statistically significant ($p < .01$) we will conclude that a difference in strictness, consistency, or effectiveness exists between the templates that were compared in that regression. Our particular hypothesis is supported if that's the case and the effect is in the expected direction (higher strictness / consistency / effectiveness for the more comprehensive template). The alpha level of .01 is based on a Bonferroni correction ($\frac{.05}{4} \approx .01$) where we assume four independent analyses (the regressions involving the variables strictness and consistency). The analysis involving effectiveness is not independent from the other analyses because effectiveness is computed based on the strictness and consistency scores of the different study parts (see the section 'Measuring preregistration effectiveness').

Table 3 presents the estimated power for each template comparison. The power calculations were done using G*Power 3.1.9.4 with settings as shown in Figure 1. We only varied the sample sizes of the groups, which amount to the effective frequencies of the templates (i.e., the number of studies using the different templates taking into account that studies are nested in papers in our data). These effective frequencies are 144, 87, and 23 for OSF1, AP, and SP, respectively (based on ICC = .5).

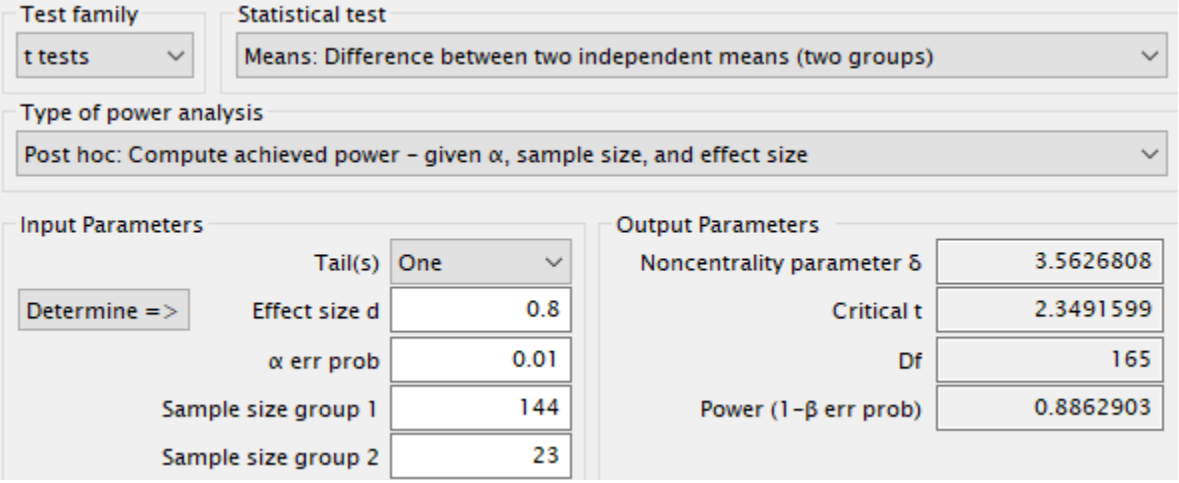
Table 3

Estimated power for each template comparison in Hypothesis 2

	Small effect (*d*=.2)	Medium effect (*d*=.5)	Large effect (*d*=.8)
OSF1 vs. AP	0.19	0.91	1.00
OSF1 vs. SP	0.07	0.45	0.87

Figure 1

G*Power settings for the power analyses denoted in Table 3



Research question 6 (RQ6)

To test Hypothesis 2 (related to RQ6) we will run three multilevel regressions (study is first level, paper is second level), one with preregistration strictness, one with preregistration-study consistency, and one with preregistration effectiveness as the dependent variable. The number of months between the study's preregistration date and the preregistration date of the first published study in our sample (see the section 'Measuring registration dates') will be included as the main independent variable, and replication status will be included as a control variable. In R-code (version 3.6.1) this look as follows:

```
strictnessMonths <- lmer(strictness ~ months + replic + (1 | paper), data =
PPP)
summary(strictnessMonths)
consistencyMonths <- lmer(consistency ~ months + replic + (1 | paper), data =
PPP)
summary(consistencyMonths)
effectivenessMonths <- lmer(effectiveness ~ months + replic + (1 | paper),
data = PPP)
summary(effectivenessMonths)
```

If the regression coefficient of months is found to be statistically significant ($p < .025$) we will conclude that strictness, consistency, or effectiveness has changed over time. Our particular hypothesis is supported if that's the case and the effect is in the expected direction (higher strictness / consistency / effectiveness over time). The alpha level of .025 is based on a Bonferroni correction ($\frac{.05}{2} = .025$) where we assume two independent analyses (the regressions involving the variables strictness and consistency). The analysis involving effectiveness is not independent from the other analyses because effectiveness is computed based on the strictness and consistency scores of the different study parts (see the section 'Measuring preregistration effectiveness').

Research question 7 (RQ7)

To answer RQ7 we will assess for each paper in our sample whether they fulfilled the four criteria for obtaining a preregistration badge. For the first criterion (RQ7a) we will code a paper with a preregistration badge with a 1 ('criterion fulfilled') if the preregistration was categorized as 'frozen and accessible' or 'frozen and inaccessible', and with a 0 ('criterion not fulfilled') if the preregistration was categorized as 'non-frozen', or 'inaccessible'. This will yield a proportion of papers that have fulfilled the first criterion, but we will also present the proportion of papers in each of the four categories. See the section 'Measuring whether preregistrations are registered in line with the criteria for earning a preregistration badge' for more information about these categorizations.

For the second criterion (RQ7b) we will code the paper with a 1 if the preregistration's registration date precedes the month of data collection, and with a 0 if the month of data collection precedes the paper's registration date. If no month of data collection is presented in the paper, we will code this as 'Not applicable'. See the section 'Measuring registration dates' for more information.

For the third criterion (RQ7c) we will code the paper with a 1 if it has a preregistration-study consistency score of 4 or more out of 5, and with a 0 if it has a lower score.

For the fourth criterion (RQ7d) we will code the paper with a 1 if the results of all preregistered hypotheses can be retrieved from the paper (i.e., if there are no 'omitted hypotheses'). See RQ1 in the preregistration of the study on selective hypothesis reporting (https://osf.io/z4awv) for more information.

In case a paper involves multiple preregistered studies, we will include in our results the preregistration that fulfills the most criteria. In case multiple preregistrations fulfill the same number of criteria, we will include the preregistration that is linked to the first study in our database related to the paper. We will separately present the proportion of papers adhering to the first, second, third, and fourth criterion, and we will present a frequency table including the number of papers that adhere to one, two, three, and four of the criteria. Note that this is the only research question for which the analysis is done on the paper level (because only papers get badges, not the individual studies within the papers).

Outliers, missing data, and robustness analyses

Because we include categorical variables, and the ranges of the variables 'months', strictness, consistency, and effectiveness are restricted we do not have to define or deal with statistical outliers, and because we force responses in our Qualtrics protocol we do not anticipate having to define or deal with missing data. However, because our protocol pilots indicated that the hypotheses in some preregistrations are very difficult to identify, we will also run our models excluding preregistrations that were assessed by at least one of the coders as "very difficult" (on a 5-point scale ranging from "very easy" to "very difficult").

Coders

All coders hold at least an undergraduate degree in psychology or a related field. Coders were sought out at the Metascience Symposium at Stanford University in September 2019 and at the REWARD EQUATOR Conference in Berlin in February 2020. Additionally, we posted a call on Twitter and made inquiries within our peer network. Those interested were encouraged to send a message to the first author and are contacted when this preregistration goes live, or before to help out with piloting the protocols.

All preregistration-study pairs will be coded by two coders who do so independently. Any inconsistencies between these two coders will be resolved among themselves afterward. If an inconsistency cannot be resolved a third coder will make the final call. To assess interrater reliability, we will use Cohen's kappa (Cohen, 1960).

References

Bakker, M., Veldkamp, C. L., van Assen, M. A., Crompvoets, E. A., Ong, H. H., Nosek, B. A., ... & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLoS Biology*, *18*(12), e3000937.

Claesen, A., Gomes, S. L. B. T., Tuerlinckx, F., & Vanpaemel, W. (2020, June 25). Preregistration: Comparing dream to reality. https://doi.org/10.31234/osf.io/d8wex

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ..., & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490.

Kambouris, S., Singleton Thorn, F., Van den Akker, O., De Jonge, M., Rüffer, F., Head, A., & Fidler, F. (2020). Database of Articles with Open Science Badges: 2020-02-21 Snapshot. https://doi.org/10.17605/osf.io/q46r5

Van den Akker, O. R., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., … Bakker, M. (2021, February 22). Preregistration of secondary data analysis: A template and tutorial. https://doi.org/10.31234/osf.io/hvfmr

Van Zant, A. B., & Moore, D. A. (2015). Leaders' use of moral justifications increases policy support. *Psychological Science*, *26*(6), 934-943. https://doi.org/10.1177/0956797615572909

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*, *2*(3), 214-227.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology*, *7*, 1832.