TILBURG UNIVERSITY

We thank you for your time spent taking this survey.
Your response has been recorded.

Below is a summary of your responses

Download PDF

Assessing preregistration effectiveness

In this protocol, you will assess the strictness of a preregistration, and check whether the preregistration is consistent with the corresponding study. The links to the preregistration and the paper containing the study can be found in the Excel-file you have been provided.

Please select your initials and copy-paste the information requested below from the Excel-file. Study Label Prereg and Study Label Paper refer to the study within the preregistration and the study within the paper you need to code. An 'NA' response in the Excel-file means that there is only one study to code.

PSP ID

0

Coder initials

- (●) OA
- ( ) MB
- ( ) AS
- ( ) GN
- ( ) KH
- ( ) SS
- ( ) SA

- CP
- LV
- AC
- DL
- MS
- DD
- ME
- SF
- FH
- KY
- SG
- LS
- LF
- KK
- EH
- TE
- MP
- LK
- FD
- JC
- NAA
- FA
- YY
- HH
- BK
- SS2
- MK
- SG2
- SW
- SAJ
- BG
- FdO
- KK2
- BJB

Paper Title

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

Study Label Prereg

NA

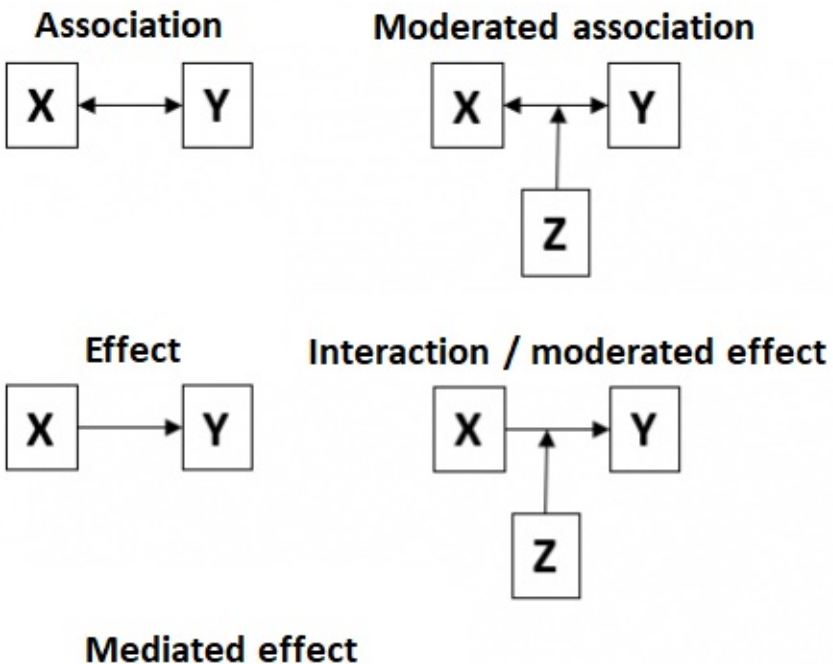
Study Label Paper

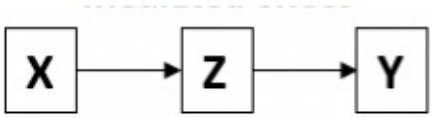
NA

You have been assigned a hypothesis from a preregistration-study pair (PSP). Please copy-paste rows M through U from the Excel-file into the empty text boxes below. This information will make it possible to effectively fill out the rest of the form.

Below you can fill our some basic information about the hypothesis you have been assigned. The figure lists the different hypothesis types that you can encounter.

An association consists of two independent variables (X and Y), a moderated association consists of two independent variables (X and Y) and a third variable (Z), an effect consists of an independent (X) and a dependent variable (Y), and both an interaction / moderated effect and a mediated effect consist of an independent (X), dependent (Y), and third variable (Z).





Hypothesis Text Prereg	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
Hypothesis Text Paper	Studies based on more comprehensive preregistration templates are more consistent with their preregistration than studies based on less comprehensive preregistration templates
Hypothesis Type	Effect

Hypothesis variables and (categories):

Independent Variable 1	Preregistration template comprehensiveness (OSF vs. Social Psychology)
Independent Variable 2	n
Third Variable	n
Dependent Variable	Preregistration-study consistency
First Control Variable	n
Other Control Variables	n

In this part of the protocol, you will be asked whether some parts of the PREREGISTRATION are specified in a 'producible manner'. Any one part of the preregistration is said to be PRODUCIBLE when the authors describe all steps that will be taken in that part (it should be specific) and each of the described steps allows only one interpretation or implementation (it should be precise). We use the term "producible" because you should be able to "produce" this part of the study based on the information in the preregistration.

Note: When the authors specify a part of the preregistration by referring to a supplementary document or another study within the preregistration, please also check for information there, but only include information from supplementary documents if the information can be clearly linked to the part of the preregistration you are trying to code. So, for example, if the authors state "the sampling plan is the same as in Study 1" please check Study 1 for information about the sampling plan but not anything else. And if the authors state "all items can be found in the supplementary materials" this should not be coded as producible if the

supplementary materials do not clearly indicate which items belong to which measure.

Note: When the authors specify a part of the preregistration by referring to a different paper, please do not code this as producible. The information should be contained within the preregistration itself (and possibly the supplementary materials).

You will assess the producibility of the following study and hypothesis. That is, all questions need to be answered with this hypothesis in mind.

Study Label Prereg:
NA

Hypothesis text:
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

Type of hypothesis:
Effect

Hypothesis variables and categories:
IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)
IV2: n
3V: n
DV: Preregistration-study consistency
CV1: n
CVs: n

Is Independent Variable 1 manipulated as part of an experiment?

IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)

- ○ Yes
- ◉ No

Does the PREREGISTRATION specify which measure(e.g., test, scale, question set, physical measurement) is used as INDEPENDENT VARIABLE 1?

Note: You will be asked whether this variable is specified producibly in a later

question.

IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)

- (●) Yes
- ( ) No

For INDEPENDENT VARIABLE 1, we distinguish between a NON-COMPOSITE MEASURE (one measurement, e.g., age, gender, or a single item) and a COMPOSITE MEASURE (several measurements or items are combined to one scale or measurement by using a sum, linear combination, SEM, or other method).

IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)

- ( ) Non-composite
- (●) Composite

Please copy-paste the text from the PREREGISTRATION that is about the operationalization of INDEPENDENT VARIABLE 1. Please include all information that can help score the producibility of this part of the preregistration (and particularly information about the elements listed in the next question).

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)

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. [The scores are presented in Table 1, for both OSF and Social Psychology]

Is the protocol to measure the individual components of INDEPENDENT VARIABLE 1 described in a producible manner? That is, are the following elements described in a producible manner?

producible manner?

1. The procedure of measurement (e.g., information about the administration of an EEG, IQ test, or personality scale) [procedure]
2. The potential values of each component (e.g., the response options of individual items in a questionnaire) [values]
3. The procedure to construct the composite from its components (e.g., arithmetic mean, weighted mean, sum) [construction]

IV1: **Preregistration template comprehensiveness (OSF vs. Social Psychology)**

- [x] Yes, procedure
- [x] Yes, values
- [x] Yes, construction
- [ ] No, none of the elements

Does the PREREGISTRATION specify which measure (e.g., test, scale, question set, physical measurement) is used as the DEPENDENT VARIABLE?

Note: You will be asked whether this variable is specified producibly in a later question.

DV: **Preregistration-study consistency**

- (●) Yes
- ( ) No

For the DEPENDENT VARIABLE, we distinguish between a NON-COMPOSITE MEASURE (one measurement) and a COMPOSITE MEASURE (several measurements or items are combined to one scale or measurement by using a sum, linear combination, SEM, or other method).

DV: **Preregistration-study consistency**

- ( ) Non-composite
- (●) Composite

Please copy-paste the text from the PREREGISTRATION that is about the operationalization of the DEPENDENT VARIABLE. Please include all information that can help score the producibility of this part of the preregistration (and particularly

information about the elements listed in the next question).

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

DV: Preregistration-study consistency

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.

Is the protocol to measure the individual components of the DEPENDENT VARIABLE described in a producible manner? That is, are the following elements described in a producible manner?

1. The procedure of measurement (e.g., information about the administration of an EEG, IQ test, or personality scale) [procedure]
2. The potential values of each component (e.g., the response options of individual items in a questionnaire) [values]
3. The procedure to construct the composite from its components (e.g., arithmetic mean, weighted mean, sum) [construction]

DV: Preregistration-study consistency

- [x] Yes, procedure
- [x] Yes, values
- [x] Yes, construction
- [ ] No, none of the elements

Please copy-paste the text from the PREREGISTRATION that is about the DATA COLLECTION PROCEDURE. Please include all information that can help score the producibility of this part of the preregistration (and particularly information about the elements listed in the next question). If the preregistration does not mention a data collection procedure at all, please fill out NA.

collection procedure at all, please fill out N.A.

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

Note: You do not have to copy-paste information about the power analysis, if the authors did one.

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.

Does the PREREGISTRATION describe the DATA COLLECTION PROCEDURE in a producible manner? That is, are the following elements clearly specified?

- **The exact number of participants (not a minimum) the authors want to include in the study [sample size]**
- **The exact time frame and situation (i.e., period, not exact dates) in which participants will be**

invited [sampling frame]

If [sample size] is specified, please also fill out the exact number of participants. Note that this relates to the effective sample size (i.e., the sample size that will be used in the analysis to draw conclusions about the hypothesis).

- [x] Yes, sample size

484

- [ ] Yes, sampling frame
- [ ] No, none of the elements

Did the authors use a POWER ANALYSIS to determine the sample size?

- ○ Yes
- ● No

Please copy-paste the text from the PREREGISTRATION that is about the INCLUSION / EXCLUSION CRITERIA for PARTICIPANTS / DATA. Please include all information that can help score the producibility of this part of the preregistration. If the preregistration does not mention inclusion and exclusion criteria at all, please fill out NA.

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

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.

Does the PREREGISTRATION specify producible INCLUSION / EXCLUSION CRITERIA that will be used to select PARTICIPANTS / DATA? That is, does the preregistration specify the following elements in a producible manner?

1. **The definitions underlying participant / data selection (e.g., how demographic information is assessed, what constitutes an outlier, what it means for a participant to not participate seriously)? [definition]**
2. **The method to exclude participants / data (e.g., exclusion before or after data collection, the use of nonparametric tests, bootstrapping)? [method]**

Notes:

- **When the authors explicitly state that they will analyze *all* the data, please consider this as producible for both elements.**
- **You can disregard statements about how *incomplete or missing data* are handled as you will be asked about that in another question.**

- [x] Yes, definition
- [x] Yes, method
- [ ] No, none of the elements

Please copy-paste the text from the PREREGISTRATION that is about how the study deals with INCOMPLETE OR MISSING DATA. Please include all information that can help score the producibility of this part of the preregistration (and particularly information about the elements listed in the next question). If the preregistration does not mention handling incomplete or missing data, please fill out NA.

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

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. 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.

Does the PREREGISTRATION indicate in a producible manner how the study deals with INCOMPLETE

OR MISSING DATA? That is, are the following elements clearly specified?

1. **The definition of a missing case [definition]**
2. **The procedure to handle missing cases (e.g., pairwise deletion, listwise deletion, imputation method, intention-to-treat method, full information method) [method]**

- [x] Yes, definition
- [x] Yes, method
- [ ] No, none of the elements

Please copy-paste the text from the PREREGISTRATION that is about the FIRST STATISTICAL MODEL TESTING THE HYPOTHESIS. The model should include the variables listed below (and exclude any non-listed variables).

Important:

- **Only the first specification of the statistical model counts. For example, if the authors first present a model without robust standard errors and then a model with robust standard errors (both testing the same hypothesis), select the one without robust standard errors. Limit yourself to copy-pasting information about that model because you will need to retrieve that particular model from the paper based on this information.**
- **Please include all information that can help score the producibility of this part of the preregistration (and particularly information about the elements listed in the next question). If the preregistration does not mention a statistical model testing the hypothesis, please fill out NA.**
- **In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].**

Hypothesis text:
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

Type of hypothesis:
Effect

Hypothesis variables and categories:
IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)

IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)
IV2: n
3V: n
DV: Preregistration-study consistency
CV1: n
CVs: n

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

Does the PREREGISTRATION specify THE FIRST STATISTICAL MODEL TESTING THE HYPOTHESIS in a producible manner? That is, are the following elements clearly specified?

1. The statistical model used (e.g., t-test, chi-squared test, linear / logistic regression, two-way ANOVA) [model]
2. The relevant variables and their factor levels (including mediating, moderating, interacting, and control variables) [variables]
3. The manner in which the variables are used in the analysis (e.g., mean centered, SEM model specification including potential residual covariances, robust standard errors) [details]

If the script for the statistical analysis is provided, please score this question as 'Yes'.

- [x] Yes, model
- [x] Yes, variables
- [ ] Yes, details
- [ ] No, none of the elements

Please copy-paste the text from the PREREGISTRATION with information about how the authors test for VIOLATIONS OF STATISTICAL ASSUMPTIONS. Please include all information that can help score the producibility of this part of the preregistration (and particularly information about the elements listed in the next question). If the preregistration does not mention how violations of statistical assumptions are handled, please fill out NA.

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

NA

Does the PREREGISTRATION indicate in a producible manner how the authors test for VIOLATIONS OF STATISTICAL ASSUMPTIONS and what they do with violations? That is, are the following elements clearly specified?

1. Which assumptions are checked (e.g., normality, homoscedascity, linearity, homogeneity of variances, sphericity)? [which]
2. How the assumptions are checked (e.g., type of test like Levene's test, alpha level)? [how]
3. What is done in cases of violations (e.g., transformations, non-parametric tests)? [deal]

- [ ] Yes, which
- [ ] Yes, how
- [ ] Yes, deal
- [x] No, none of the elements

Please copy-paste the text from the PREREGISTRATION that is about the INFERENCE CRITERIA. Please include all information that can help score the producibility of this part of the preregistration. If the preregistration does not mention inference criteria at all, please fill out NA.

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

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 ($.05/4 \approx .01$) 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').

Does the PREREGISTRATION indicate the INFERENCE CRITERIA that will be used in a producible manner (e.g., statistical significance, sidedness of the test, corrections for multiple testing, Bayesian criteria)? Note that the authors need to be explicit in what the sidedness of a significance test is (i.e., one-sided vs. two-sided) and what the cut-off criterion for their statistical decision is (e.g., Bayes factor, alpha value).

- [x] Yes
- [ ] No

Please write down any comments you have about coding this part of the protocol.

Technically a secondary data analysis because the data were already collected before preregistering.

In this part of the protocol, you will be asked whether some parts of the PAPER are specified in a 'reproducible manner'. Any one part of the paper is said to be REPRODUCIBLE when the authors describe all steps that were taken in that part (it should be specific) and each of the described steps allows only one interpretation or implementation (it should be precise). We use the term "reproducible" because you should be able to "reproduce" this part of the study based on the information in the paper.

Note: When the authors specify a part of the paper by referring to a supplementary document or another study within the paper, please also check for information there, but only include information from supplementary documents if the information can be clearly linked to the part of the paper you are trying to code. So, for example, if the authors state "the sampling plan is the same as in Study 1" please check Study 1 for information about the sampling plan but not anything else. And if the authors state "all items can be found in the supplementary materials" this should not be coded as reproducible if the supplementary materials do not clearly indicate which items belong to which measure.

Note: When the authors specify a part of the paper by referring to a different paper, please do not code this as reproducible. The information should be contained within the paper itself (and possibly supplementary materials).

You will assess the producibility of the following hypothesis. That is, all questions need to be answered with this hypothesis in mind.

Study Label Paper:
NA

Hypothesis text:
Studies based on more comprehensive preregistration templates are more consistent with their preregistration than studies based on less comprehensive preregistration templates

Type of hypothesis:
Effect

Hypothesis variables and categories:
IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)
IV2: n
3V: n
DV: Preregistration-study consistency
CV1: n
CVs: n

Is Independent Variable 1 manipulated as part of the experiment?

IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)

- ○ Yes
- ● No

Does the PAPER specify which measure (e.g., test, scale, question set, physical measurement) is used as the INDEPENDENT VARIABLE 1?

IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)

- ● Yes
- ○ No

For INDEPENDENT VARIABLE 1, we distinguish between a NON-COMPOSITE MEASURE (one measurement) and a COMPOSITE MEASURE (several measurements or items are combined to one scale or measurement by using a sum, linear combination, SEM, or other method).

IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)

- ○ Non-composite
- ● Composite

Please copy-paste the text from the PAPER that is about the operationalization of INDEPENDENT VARIABLE 1. Please include all information that can help score the reproducibility of this part of the paper (and particularly information about the

reproducibility of this part of the paper (and particularly information about the elements listed in the next question).

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)

To identify the preregistration template used for a specific study we searched the paper presenting that study 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. We scored the three preregistration templates with the highest frequency on their comprehensiveness (i.e., their potential to restrict researcher degrees of freedom) using a newly developed protocol. Using that protocol, we assessed whether the template included a prompt, additional instructions, and an example for the nine major and minor study parts (see https://osf.io/rtuvb for the filled-out protocol). The maximum possible comprehensiveness score using this protocol was 27 (very comprehensive), which each of the five major and four minor study parts receiving a maximum of 3 points. We gave 1 point if the study part was included in the template without additional instructions and an example, 2 points if it was included with either additional instructions or an example, and 3 points if it was included with both additional instructions and an example. When the study part was not included in the template, 0 points were given. Scoring was done by two independent coders (ORA and CRP) who together resolved three initial coding discrepancies. 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 their comprehensiveness score. We observed large differences in comprehensiveness between the templates. While the OSF Prereg template scored almost the maximum number of points (24/27), the AsPredicted template and the Pre-Registration in Social Psychology template scored substantially less well, with 10 and 14 out of 27 points, respectively.

Is the protocol to measure the individual components of INDEPENDENT VARIABLE 1 described in a reproducible manner? That is, are the following elements described in a reproducible manner?

1. The procedure of measurement (e.g., information about the administration of an EEG, IQ test, or personality scale) [procedure]
2. The potential values of each component (e.g., the response options of individual items in a questionnaire) [values]
3. The procedure to construct the composite from its components (e.g., arithmetic mean, weighted mean, sum) [construction]

IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)

- [x] Yes, procedure
- [x] Yes, values
- [x] Yes, construction

- [x] Yes, construction
- [ ] No, none of the elements

Does the PAPER specify which measure (e.g., test, scale, question set, physical measurement) is used as the DEPENDENT VARIABLE?

DV: Preregistration-study consistency

- (●) Yes
- ( ) No

For the DEPENDENT VARIABLE, we distinguish between a NON-COMPOSITE MEASURE (one measurement) and a COMPOSITE MEASURE (several measurements or items are combined to one scale or measurement by using a sum, linear combination, SEM, or other method).

DV: Preregistration-study consistency

- ( ) Non-composite
- (●) Composite

Please copy-paste the text from the PAPER that is about the operationalization of the DEPENDENT VARIABLE. Please include all information that can help score the reproducibility of this part of the paper (and particularly information about the elements listed in the next question).

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

DV: Preregistration-study consistency

To assess the consistency between a preregistration and the actual study, we scored whether the description of a study part in the preregistration and the corresponding paper were consistent. A preregistration and a study were considered ‘consistent’ when the researcher adhered to the action described in the preregistration within the published paper. In the preregistration-study consistency part of the protocol, any part could earn 1 point (consistent) or 0 points (inconsistent). This meant that the total consistency score could be between 0 (not consistent at all) and 5 (very consistent).

Is the protocol to measure the individual components of the DEPENDENT

VARIABLE described in a reproducible manner? That is, are the following elements described in a reproducible manner?

1. The procedure of measurement (e.g., information about the administration of an EEG, IQ test, or personality scale) [procedure]
2. The potential values of each component (e.g., the response options of individual items in a questionnaire) [values]
3. The procedure to construct the composite from its components (e.g., arithmetic mean, weighted mean, sum) [construction]

DV: **Preregistration-study consistency**

- [x] Yes, procedure
- [x] Yes, values
- [x] Yes, construction
- [ ] No, none of the elements

Please copy-paste the text from the PAPER that is about the DATA COLLECTION PROCEDURE. Please include all information that can help score the reproducibility of this part of the paper (and particularly information about the elements listed in the next question). If the paper does not mention a data collection procedure at all, please fill out NA.

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

Note: You do not have to copy-paste information about the power analysis, if the authors did one.

Our selection of preregistered studies was derived from a population of 459 preregistered psychology studies that had either won a Preregistration Challenge prize via the Center for Open Science initiative (see https://cos.io/our-services/prereg-more-information) or earned a Preregistration Badge before 2020 (see https://cos.io/our-services/open-science-badges). We previously used this set of preregistrations to assess whether hypotheses outlined in preregistrations matched those outlined in the corresponding papers (Van den Akker, et al., 2022). To search for hypotheses, we used the following keywords: “replicat”, “hypothes”, "investigat", “test”, “predict”, “examin”, and “expect”. Once we determined that the sentence with the keyword was indeed a hypothesis, we copy-pasted the text from the preregistration and separately extracted the variables (independent variables, dependent variables, mediating variables, and control variables). In the second stage of the project of Van den Akker et al., coders were presented with the texts and the variables of all hypotheses and were asked to try to match the hypotheses to the hypotheses in the corresponding papers’ introduction or methods sections. We labeled a hypothesis as a ‘match’ if the hypothesis in the paper involved the same

variables and the same relationship between the variables as detailed in the preregistration. We ended up with a total of 1,143 matching hypotheses from 346 preregistration-study pairs (PSPs). For the current project, we randomly selected one hypothesis per PSP. We did this because assessing more than one matching hypothesis in a given study would have led to dependencies in our data. Moreover, we wanted to assess preregistration effectiveness for study elements that are typically constrained to one particular hypothesis (e.g., the operationalization of the variables, and the statistical model). During the selection process, we excluded 46 studies that only involved hypotheses for which we could not clearly determine the type of hypothesis (i.e., an association, effect, moderation, or mediation) or hypotheses involving only one variable. We did so because our method for computing preregistration effectiveness required one clear hypothesis with at least two variables. Note that we did not explicitly preregister these exclusions. As a result, our final sample consisted of 300 hypotheses from 300 PSPs.

Does the PAPER describe the DATA COLLECTION PROCEDURE in a reproducible manner? That is, are the following elements clearly specified?

1. **The exact number of participants (not a minimum) the authors included in the study [sample size]**
2. **The exact time frame (i.e., period, not exact dates) and situation in which participants were invited [sampling frame]**

If [sample size] is specified, please also fill out the exact number of participants. Note that this relates to the effective sample size (i.e., the sample size was used in the analysis to draw conclusions about the hypothesis).

- [x] Yes, sample size

300

- [ ] Yes, sampling frame
- [ ] No, none of the elements

Did the authors use a POWER ANALYSIS to determine the sample size, power, or effect size related to the hypothesis test? Note that a post hoc power analysis suffices as well.

- ○ Yes
- ● No

Did the authors describe when the data was collected? If so, please add the month and year in the open text box using the following notation: MM-YYYY.

- ○ Yes

- ● No

Please copy-paste the text from the PAPER that is about the authors' INCLUSION / EXCLUSION CRITERIA to select PARTICIPANTS / DATA. Please include all information that can help score the reproducibility of this part of the paper (and particularly information about the elements listed in the next question). If the paper does not mention inclusion / exclusion criteria at all, please fill out NA.

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

During the selection process, we excluded 46 studies that only involved hypotheses for which we could not clearly determine the type of hypothesis (i.e., an association, effect, moderation, or mediation) or hypotheses involving only one variable. We did so because our method for computing preregistration effectiveness required one clear hypothesis with at least two variables. Note that we did not explicitly preregister these exclusions. As a result, our final sample consisted of 300 hypotheses from 300 PSPs.

Does the PAPER specify reproducible INCLUSION / EXCLUSION CRITERIA that will be used to select PARTICIPANTS / DATA? That is, does the paper specify the following elements in a reproducible manner?

1. **The definitions underlying participant / data selection (e.g., how demographic information is assessed, what constitutes an outlier, what it means for a participant to not participate seriously)? [definition]**
2. **The method to exclude participants / data (e.g., exclusion before or after data collection, the use of nonparametric tests, bootstrapping)? [method]**

Notes:

- **When the authors explicitly state that they analyze *all* the data, please consider this as reproducible for both elements.**
- **You can disregard statements about how incomplete or missing data are handled as you will be asked about that in another question.**

- [x] Yes, definition
- [x] Yes, method
- [ ] No, none of the elements

Please copy-paste the text from the PAPER with information about how the study deals with INCOMPLETE OR MISSING DATA. Please include all information that can

help score the reproducibility of this part of the paper (and particularly information about the elements listed in the next question). If the paper does not mention incomplete or missing data at all, please fill out NA.

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

NA

Does the PAPER indicate in a reproducible manner how the study deals with INCOMPLETE OR MISSING DATA? That is, are the following elements clearly specified?

1. The definition of a missing case [definition]
2. The procedure to handle missing cases (e.g., pairwise deletion, listwise deletion, imputation method, intention-to-treat method, full information method) [method]

- [ ] Yes, definition
- [ ] Yes, method
- [x] No, none of the elements

Please copy-paste the text from the PAPER that is about the STATISTICAL MODEL YOU IDENTIFIED IN THE PREREGISTRATION. The information you copy-pasted about the model from the preregistration can be found below.

Please include all information that can help to score the reproducibility of this part of the paper (and particularly information about the elements listed in the next question). If the paper does not mention a statistical model testing the hypothesis, please fill out NA.

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

Statistical model identified in the preregistration:
consistency.OSF1vsSP <- lmer(consistency ~ OSF1vsSP + replic + (1 | paper), data =

consistency.OSFvsSP <- lmer(consistency ~ OSFvsSP + replic + (1 | paper), data = PPP)

To compare preregistration templates in line with Hypothesis 2, we ran the same three multilevel regressions as for Hypothesis 1 twice: once with the addition of a dummy variable with value 1 if the preregistration was produced using the Open Science Framework template, and value 0 if the preregistration was produced using the AsPredicted template (M2a1, M2b1, and M2c1), and once with the addition of a dummy variable with value 1 if the preregistration was produced using the Open Science Framework template, and value 0 if the preregistration was produced using the Social Psychology template (M2a2, M2b2, and M2c2). [Text for H1: To test whether replication studies were more effectively preregistered than original studies (Hypothesis 1) we ran three multilevel regressions (with study as the first level, and paper as the second level): one with preregistration strictness (M1a), one with preregistration-study consistency (M1b), and one with preregistration effectiveness as the dependent variable (M1c). The main independent variable replic was a dummy (replication vs. original study).]

Does the PAPER describe the STATISTICAL MODEL in a reproducible manner? That is, are the following elements clearly specified?

1. The statistical model used (e.g., t-test, chi-squared test, linear / logistic regression, two-way ANOVA) [model]
2. The relevant variables and their factor levels (including mediating, moderating, interacting, and control variables) [variables]
3. The manner in which the variables are used in the analysis (e.g., mean centered, SEM model specification including potential residual covariances, robust standard errors) [details]

If the script for the statistical analysis is provided, please score this question as 'Yes'.

- [x] Yes, model
- [x] Yes, variables
- [x] Yes, details
- [ ] No, none of the elements

Please copy-paste the text from the PAPER that is about how the authors test for VIOLATIONS OF STATISTICAL ASSUMPTIONS and how they deal with them. Please include all information that can help score the reproducibility of this part of the paper (and particularly information about the elements listed in the next question). If the paper does not mention violations of statistical assumptions at all, please fill out NA.

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this

reconciliation process you can add notes in the text box at your discretion using [square brackets].

NA

Does the PAPER indicate how the authors test for VIOLATIONS OF STATISTICAL ASSUMPTIONS and how they deal with them? That is, are the following elements in a reproducible manner?

1. Which assumptions are checked (e.g., normality, homoscedascity, linearity, homogeneity of variances, sphericity)? [which]
2. How the assumptions are checked (e.g., type of test like Levene's test, alpha level)? [how]
3. What is done in cases of violations (e.g., transformations, non-parametric tests)? [deal]

- [ ] Yes, which
- [ ] Yes, how
- [ ] Yes, deal
- [x] No, none of the elements

Please copy-paste the text from the PAPER that is about the INFERENCE CRITERIA. If the authors are not explicit about the inference criteria, please copy-paste the statistical conclusion because this often indicates the inference criteria implicitly (e.g., "extraversion was significantly associated with physical strength, t(90) = 2, p < .05" indicates an alpha level of .05). If you can't find the statistical conclusion pertaining to the hypothesis, please fill out NA.

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

We computed unstandardized coefficients for all preregistered hypothesis tests and used an alpha level of .01.

Does the PAPER indicate the INFERENCE CRITERIA in a reproducible manner (e.g., statistical significance, sidedness of the test, corrections for multiple testing, Bayesian criteria)? Note that the authors need to *explicitly* indicate the alpha-level, Bayes factor, or sidedness for this to be reproducible. An implicit mention in the statistical conclusion is thus not consistent with reproducible inference criteria.

- [x] Yes
- [ ] No

Please write down any comments you have about coding this part of the protocol.

In this part of the protocol, you will be asked to assess whether the preregistration and the paper are CONSISTENT. The preregistration and the published paper are 'consistent' on a given element only when that element 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.

For example, "We will use items 1 through 4 of the 6-item EMP scale to measure empathy" and "To measure empathy we used the EMP scale excluding items 5 and 6" are equivalent because, even though the wording differs, the researcher's action (using the first four items of the scale) is the same.

You will assess the preregistration-paper consistency of the following hypothesis. That is, all questions need to be answered with this hypothesis in mind.

Study Label Prereg:
NA

Study Label Paper:
NA

Hypothesis text:
Studies based on more comprehensive preregistration templates are more consistent with their preregistration than studies based on less comprehensive preregistration templates

Type of hypothesis:
Effect

Hypothesis variables and categories:
IV1: Preregistration template comprehensiveness (OSF vs. Social Psychology)
IV2: n
3V: n
DV: Preregistration-study consistency

CV1: n
CVs: n

From the texts below, please assess whether the operationalization of INDEPENDENT VARIABLE 1 is consistent between the preregistration and the paper. That is, are all of the following elements consistent?
1. Which measure is used [specification]
2. The procedure of measurement (e.g., information about the administration of an EEG, IQ test, or personality scale) [procedure]
3. The potential values of each component (e.g., the response options of individual items in a questionnaire) [values]
4. The procedure how they will construct the composite from its elements (e.g., arithmetic mean, weighted mean, sum) [construction]

Preregistration text:
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. [The scores are presented in Table 1, for both OSF and Social Psychology]

Paper text:
To identify the preregistration template used for a specific study we searched the paper presenting that study 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. We scored the three preregistration templates with the highest frequency on their comprehensiveness (i.e., their potential to restrict researcher degrees of freedom) using a newly developed protocol. Using that protocol, we assessed whether the template included a prompt, additional instructions, and an example for the nine major and minor study parts (see https://osf.io/rtuvb for the filled-out protocol). The maximum possible comprehensiveness score using this protocol was 27 (very

comprehensive), which each of the five major and four minor study parts receiving a maximum of 3 points. We gave 1 point if the study part was included in the template without additional instructions and an example, 2 points if it was included with either additional instructions or an example, and 3 points if it was included with both additional instructions and an example. When the study part was not included in the template, 0 points were given. Scoring was done by two independent coders (ORA and CRP) who together resolved three initial coding discrepancies. 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 their comprehensiveness score. We observed large differences in comprehensiveness between the templates. While the OSF Prereg template scored almost the maximum number of points (24/27), the AsPredicted template and the Pre-Registration in Social Psychology template scored substantially less well, with 10 and 14 out of 27 points, respectively.

NOTE: the response options only cover the elements that were previously labeled as (partially) reproducible for both the preregistration *and* the paper.

- [x] Yes, specification
- [x] Yes, procedure
- [x] Yes, values
- [x] Yes, construction
- [ ] No, none of the elements are consistent

From the texts below, please assess whether the operationalization of the DEPENDENT VARIABLE is consistent between the preregistration and the paper. That is, are all of the following elements consistent?
1. Which measure is used [specification]
2. The procedure of measurement (e.g., information about the administration of an EEG, IQ test, or personality scale) [procedure]
3. The potential values of each component (e.g., the response options of individual items in a questionnaire) [values]
4. The procedure how they will construct the composite from its elements (e.g., arithmetic mean, weighted mean, sum) [construction]

Preregistration text:
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.

Paper text:
To assess the consistency between a preregistration and the actual study, we scored whether the description of a study part in the preregistration and the corresponding paper were consistent. A preregistration and a study were considered 'consistent' when the researcher adhered to the action described in the preregistration within the published paper. In the preregistration-study consistency part of the protocol, any part could earn 1 point (consistent) or 0 points (inconsistent). This meant that the total consistency score could be between 0 (not consistent at all) and 5 (very consistent).

NOTE: the response options only cover the elements that were previously labeled as (partially) reproducible for both the preregistration *and* the paper.

- [x] Yes, specification
- [x] Yes, procedure
- [x] Yes, values
- [x] Yes, construction
- [ ] No, none of the elements are consistent

From the texts below, please assess whether the DATA COLLECTION PROCEDURE is consistent between the preregistration and the paper. That is, are all of the following elements consistent?
1. The exact number of participants the authors want to include / included in the study [sample size]
2. The exact time frame (i.e., period, not exact dates) and situation in which participants will be/were invited [sampling frame]

Preregistration text:
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.

<u>Paper text</u>:

Our selection of preregistered studies was derived from a population of 459 preregistered psychology studies that had either won a Preregistration Challenge prize via the Center for Open Science initiative (see https://cos.io/our-services/prereg-more-information) or earned a Preregistration Badge before 2020 (see https://cos.io/our-services/open-science-badges). We previously used this set of

preregistrations to assess whether hypotheses outlined in preregistrations matched those outlined in the corresponding papers (Van den Akker, et al., 2022). To search for hypotheses, we used the following keywords: “replicat”, “hypothes”, "investigat", “test”, “predict”, “examin”, and “expect”. Once we determined that the sentence with the keyword was indeed a hypothesis, we copy-pasted the text from the preregistration and separately extracted the variables (independent variables, dependent variables, mediating variables, and control variables). In the second stage of the project of Van den Akker et al., coders were presented with the texts and the variables of all hypotheses and were asked to try to match the hypotheses to the hypotheses in the corresponding papers’ introduction or methods sections. We labeled a hypothesis as a ‘match’ if the hypothesis in the paper involved the same variables and the same relationship between the variables as detailed in the preregistration. We ended up with a total of 1,143 matching hypotheses from 346 preregistration-study pairs (PSPs). For the current project, we randomly selected one hypothesis per PSP. We did this because assessing more than one matching hypothesis in a given study would have led to dependencies in our data. Moreover, we wanted to assess preregistration effectiveness for study elements that are typically constrained to one particular hypothesis (e.g., the operationalization of the variables, and the statistical model). During the selection process, we excluded 46 studies that only involved hypotheses for which we could not clearly determine the type of hypothesis (i.e., an association, effect, moderation, or mediation) or hypotheses involving only one variable. We did so because our method for computing preregistration effectiveness required one clear hypothesis with at least two variables. Note that we did not explicitly preregister these exclusions. As a result, our final sample consisted of 300 hypotheses from 300 PSPs.

NOTE: the response options only cover the elements that were previously labeled as (partially) reproducible for both the preregistration *and* the paper.

- [ ] Yes, sample size
- [x] No, none of the elements are consistent

In what way is the sample size inconsistent?

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

Sample size is smaller due to unforeseen and non-preregistered exclusions

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".

During the selection process, we excluded 46 studies that only involved hypotheses for which we could not clearly determine the type of hypothesis (i.e., an association, effect, moderation, or mediation) or hypotheses involving only one variable. We did so because our method for computing preregistration effectiveness required one clear hypothesis with at least two variables. Note that we did not explicitly preregister these exclusions.

From the texts below, please assess whether the author's INCLUSION / EXCLUSION CRITERIA to select PARTICIPANTS / DATA are consistent between the preregistration and the paper. That is, are all of the following elements consistent?

1. **The definitions underlying participant / data selection (e.g., how demographic information is assessed, what constitutes an outlier, what it means for a participant to not participate seriously)? [definition]**
2. **The method to exclude participants / data (e.g., exclusion before or after data collection, the use of nonparametric test, bootstrapping)? [method]**

Preregistration text:

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.

Paper text:

During the selection process, we excluded 46 studies that only involved hypotheses for which we could not clearly determine the type of hypothesis (i.e., an association,

effect, moderation, or mediation) or hypotheses involving only one variable. We did so because our method for computing preregistration effectiveness required one clear hypothesis with at least two variables. Note that we did not explicitly preregister these exclusions. As a result, our final sample consisted of 300 hypotheses from 300 PSPs.

NOTE: the response options only cover the elements that were previously labeled as (partially) reproducible for both the preregistration *and* the paper.

- [ ] Yes, definition
- [x] Yes, method
- [ ] None of the elements are consistent

In what way are the definitions underlying inclusion and exclusion criteria inconsistent?

In a subsequent coding phase, you will compare your response to the response of another coder and reconcile any inconsistencies together. To facilitate this reconciliation process you can add notes in the text box at your discretion using [square brackets].

Added exclusions that were not mentioned in the prereg

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".

During the selection process, we excluded 46 studies that only involved hypotheses for which we could not clearly determine the type of hypothesis (i.e., an association, effect, moderation, or mediation) or hypotheses involving only one variable. We did so because our method for computing preregistration effectiveness required one clear hypothesis with at least two variables. Note that we did not explicitly preregister these exclusions.

From the texts below, please assess whether the <u>STATISTICAL MODEL</u> is consistent between the preregistration and the paper. That is, are all of the following elements consistent?
1. The statistical model used (e.g., t-test, chi-squared test, linear / logistic regression, two-way ANOVA) [model]
2. The relevant variables and their factor levels (including mediating, moderating, interacting, and control variables) [variables]
3. The manner in which the variables are used in the analysis (e.g., mean centered,

SEM model specification including potential residual covariances, robust standard errors) [details]

Preregistration text:
consistency.OSF1vsSP <- lmer(consistency ~ OSF1vsSP + replic + (1 | paper), data = PPP)

Paper text:
To compare preregistration templates in line with Hypothesis 2, we ran the same three multilevel regressions as for Hypothesis 1 twice: once with the addition of a dummy variable with value 1 if the preregistration was produced using the Open Science Framework template, and value 0 if the preregistration was produced using the AsPredicted template (M2a1, M2b1, and M2c1), and once with the addition of a dummy variable with value 1 if the preregistration was produced using the Open Science Framework template, and value 0 if the preregistration was produced using the Social Psychology template (M2a2, M2b2, and M2c2). [Text for H1: To test whether replication studies were more effectively preregistered than original studies (Hypothesis 1) we ran three multilevel regressions (with study as the first level, and paper as the second level): one with preregistration strictness (M1a), one with preregistration-study consistency (M1b), and one with preregistration effectiveness as the dependent variable (M1c). The main independent variable replic was a dummy (replication vs. original study).]

NOTE: the response options only cover the elements that were previously labeled as (partially) reproducible for both the preregistration *and* the paper.

- [x] Yes, model
- [x] Yes, variables
- [ ] No, none of the elements are consistent

From the texts below, please assess whether the INFERENCE CRITERIA are consistent between the preregistration and the paper.

Note: you can assume that the authors use a two-tailed test and an alpha of .05 to assess statistical significance, and a Bayes factor cut-off value of 3 (or 1/3) if they are not clear about this in the paper.

Preregistration text:
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 (.05/4≈.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').

Paper text:
We computed unstandardized coefficients for all preregistered hypothesis tests and used an alpha level of .01.

- (●) Yes
- ( ) No

Please write down any comments you have about coding this part of the protocol.

Please write down any comments you have about coding this protocol.

How difficult was it to code this preregistration-study pair?

- (●) Very easy
- ( ) Somewhat easy
- ( ) Neither easy nor difficult
- ( ) Somewhat difficult
- ( ) Very difficult

Don't forget to submit your answers by clicking on the right arrow!