Abstract

Study preregistration has become increasingly popular in psychology, but its effectiveness in restricting potentially biasing researcher degrees of freedom remains unclear. We used an extensive protocol to assess the producibility (i.e., the degree to which a study can be properly conducted based on the available information) of preregistrations and the consistency between preregistration and their corresponding papers for 300 psychology studies. We found that preregistrations often lack methodological details and that undisclosed deviations from preregistered plans are frequent. Combining the producibility and consistency results highlights that biases due to researcher degrees of freedom are likely in many preregistered studies. More comprehensive registration templates typically yielded more producible and hence better preregistrations. We did not find that effectiveness of preregistrations differed over time or between original and replication studies. Furthermore, we found that operationalizations of variables were generally more effectively preregistered than other study parts. Inconsistencies between preregistrations and published studies were mainly encountered for data collection procedures, statistical models, and exclusion criteria. Our results indicate that, to unlock the full potential of preregistration, researchers in psychology should aim to write more producible preregistrations, adhere to these preregistrations more faithfully, and more transparently report any deviations from their preregistrations. This could be facilitated by training and education to improve preregistration skills, as well as the development of more comprehensive templates.

Keywords: preregistration, preregistration producibility, preregistration-study consistency, preregistration deviation, preregistration template, open science, meta-research

Introduction

Hypothesis testing research involves making a lot of decisions. Such decisions include choosing a statistical model, the construction of outcome measures, and data handling strategies like dealing with missing data and outliers (Wicherts et al., 2016). These decisions are commonly known as 'researcher degrees of freedom' (Simmons, Nelson, & Simonsohn, 2011). The more decisions a researcher needs to make from the start of a project to its conclusion, the more degrees of freedom a study is said to have. In contrast to popular belief, researchers do not always make such decisions in a rational and objective manner (see Veldkamp, Hartgerink, Van Assen, & Wicherts, 2017). One reason for this is that researchers are susceptible to cognitive biases like confirmation bias and motivated reasoning bias (Bishop, 2020; Munafò, Chambers, Collins, Fortunato, & Macleod, 2020). In recent years, these biases have been highlighted as one of the main reasons for the replication crisis, the phenomenon that many studies fail to replicate in psychology and beyond (Malich & Munafò, 2022). One of the most common research biases involves a strong preference for research results that are easier to publish and hence beneficial to one's career because of similarly biased systematic incentives (Nosek, Spies, & Motyl, 2012). Because results involving *p*-values lower than .05 are deemed easier to publish, the label *p*-hacking has been used for the phenomenon of making research decisions to achieve a desired result (Parsons et al., 2022), although these decisions are typically neither explicitly intentional nor malicious (Smaldino & McElreath, 2016).

Following the replication crisis, several solutions have been proposed to combat questionable research practices such as *p*-hacking (see overview by Pennington, 2023). One particularly promising solution is preregistration (Nosek, Ebersole, DeHaven, & Mellor, 2018; Wagenmakers, Wetzels, Borsboom, Van der Maas, & Kievit, 2012), where researchers openly

publish their hypotheses, study design, and analysis plan before collecting or analyzing the research data. Because researchers publish their decisions beforehand, preregistration can restrict researcher degrees of freedom and lower the possibility for *p*-hacking (Wicherts, et al., 2016), thereby diminishing the potential for biased outcomes to appear in the literature. The effectiveness of preregistration in achieving this goal depends on at least two aspects: (1) the *producibility* of the preregistration (i.e., whether the information provided in the preregistration is comprehensive enough to properly conduct the study)[1], and (2) the *consistency* between the preregistration and the published 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 largely adhere to the preregistered plan, preregistration is less effective (i.e., fewer researcher degrees of freedom are restricted and there is more room for *p*-hacking and other biased decision-making).

Empirical evidence on the effectiveness of preregistration in the social sciences is limited but the available studies from different fields show that preregistrations do not typically restrict most relevant researcher degrees of freedom (economics and political science: Ofosu & Posner, 2021; gambling studies: Heirene et al., 2021; multiple fields: Bakker et al., 2020). Specifically, Ofosu and Posner noted that independent variables, dependent variables, and statistical models were clearly outlined in most preregistrations, but that only a small proportion of preregistrations specified how missing data and outliers were to be handled. Heirene et al. and Bakker et al. found similar results: decisions relating to study design were relatively well-restricted compared to decisions regarding data collection and statistical analysis. This is problematic because the

[1] We called this aspect 'strictness' in our preregistration but changed this based on a reviewer's comment.

many decisions in analyzing data could still create sizeable variation in outcomes that researchers could selectively report (Olsson-Collentine, Van Aert, Bakker, & Wicherts, 2023).

In studies examining preregistration-study consistency, estimates of undisclosed deviations range from approximately two-thirds in a sample of gambling studies (Heirene et al., 2021) to about 90% in the journal *Psychological Science* (Claesen, Gomes, Tuerlinckx, & Vanpaemel, 2021). This is in line with earlier studies from biomedicine that also identified many inconsistencies between study registrations and papers (Li et al., 2018; Thibault et al., 2021). In the field of economics and political science, Ofosu and Posner (2021) focused on inconsistencies with regard to hypotheses and found that preregistered hypotheses could be retrieved in only two-thirds of the corresponding papers. Finally, in a sample of psychology studies, Van den Akker et al. (2023) found that about half of preregistered hypotheses could not be identified in the published paper and about one-fifth of preregistered hypotheses involved a change in the hypothesized direction of the effect. Consequently, although preregistrations could theoretically reduce questionable research practices, research suggests their implementation may not be as effective as initially hoped and thought.

It is important to note that deviations from a preregistration need not always be problematic (Nosek et al., 2019). Scientific research can be nonlinear and sometimes things change during the research process that could not have been foreseen. For example, the statistical assumptions of the preregistered model may not hold in practice, a subset of participants may need to be excluded because of a technical error, or the preregistration could simply have included a mistake. In situations like these, deviating from the preregistration may be the most reasonable way to still enable proper tests of the predetermined hypothesis. However, it is crucial to explain in the published work why any deviations were necessary, perhaps through

Preregistration Planning and Deviation Documentation (Van ‘t Veer et al., 2019). Only then can readers assess the rationale behind the deviations and calibrate their confidence in the claims being made.

The current project is the first to simultaneously investigate both the producibility of preregistrations and the consistency between preregistrations and published studies in psychology. We do so in a sample of published preregistrations and papers (N = 300 when assessing producibility and N = 57 when assessing consistency). Aside from this overall assessment of preregistration effectiveness, we also assess how effectively the following specific study parts are preregistered: the operationalizations of the variables, the data collection procedure, the statistical model, the inference criteria, the exclusion criteria, the treatment of missing data, and the treatment of violations of statistical assumptions. For the study parts with the most inconsistencies between preregistration and paper, we also assess the different types of inconsistencies, the frequency with which they occur, and any explanations the authors may have for them. This may help identify areas where preregistration practices require the biggest improvements. Finally, we test several novel hypotheses that illustrate what factors may influence preregistration effectiveness, like replication status, time, and the comprehensiveness of the preregistration template.

We preregistered (see https://osf.io/83ahg) hypotheses about the overall effectiveness of psychology preregistrations, expecting that preregistration effectiveness would vary between different preregistration and study types. Our first hypothesis was that replication studies would be preregistered more effectively than original studies. Preregistration producibility may be better for replication preregistrations because 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 more producible. Additionally, preregistration-study consistency might be better for replication preregistrations because the principal goal of a replication study is to mimic the primary study. Given that the details of the primary study are specified in the published replication study, researchers doing replication studies can be expected to adhere more to the preregistration than researchers doing original studies.

Our second hypothesis was that more comprehensive preregistration templates (i.e., those targeting a greater number of research decisions) would yield more effective preregistrations than less comprehensive templates. The reasoning underlying this hypothesis is that comprehensive templates nudge researchers to specify more study details, making the preregistrations more producible than preregistrations based on less comprehensive templates. Moreover, researchers using more comprehensive templates may value restricting researcher degrees of freedom more than researchers using less comprehensive templates and are therefore more likely to adhere to the preregistration. These predictions are in line with the finding that registrations using formats with detailed instructions restricted the opportunistic use of researcher degrees of freedom better than formats with minimal direct guidance (Bakker et al., 2020). A number of preregistration templates have been developed in recent years, some with a general purpose (e.g., Bowman et al., 2020; Preregistration Task Force, 2021), and some with a specific emphasis (e.g., for replication studies: Brandt et al., 2014; for secondary data analyses: Van den Akker et al., 2021; for systematic reviews: Van den Akker et al., 2022; for qualitative research, Haven & Van Grootel, 2019). In this study, we limited ourselves to general-purpose preregistration templates for hypothesis-testing research.

Our third hypothesis was that preregistration effectiveness has improved over time, something that was previously found by Heirene et al. (2021). We expected this to be likely as

researchers are preregistering more and more (Pfeiffer & Call, 2022) and should therefore be getting more familiar and experienced with the practice of preregistration. Intuitively, this would make them more effective at (a) making their preregistrations more producible and (b) ensuring higher preregistration-study consistency.

Overview of preregistered hypotheses

1) Replication studies are more effectively preregistered than original studies
    a. Preregistrations of replication studies are more producible than preregistrations of original studies
    b. Replication studies are more consistent with their preregistration than original studies
2) Studies based on more comprehensive preregistration templates are more effectively preregistered than studies based on less comprehensive preregistration templates
    a. Preregistrations based on more comprehensive templates are more producible than preregistrations based on less comprehensive templates
    b. Studies based on more comprehensive preregistration templates are more consistent with their preregistration than studies based on less comprehensive preregistration templates
3) Preregistration effectiveness has improved over time
    a. Preregistration producibility has improved over time
    b. Preregistration-study consistency has improved over time

Method

Selection of preregistered studies

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). This set of preregistrations has been previously used to assess whether hypotheses outlined in preregistrations matched those outlined in the corresponding papers (Van den Akker, et al., 2023). To search for hypotheses, Van den Akker et al. used the following keywords: “replicat”, “hypothes”, "investigat", “test”, “predict”, “examin”, and “expect”. Once they determined that the sentence with the keyword was indeed a hypothesis, they 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, 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. A hypothesis was labeled 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. The authors 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. Other than these exclusions, there were no unanticipated missing data.

An overview of our sample selection procedure can be found in the PRISMA flow diagram (Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009) in Figure 1. The protocols used by Van den Akker et al. (2023) to identify the hypotheses in preregistrations and their accompanying papers can be found at https://osf.io/fdmx4 and https://osf.io/uyrds, respectively.

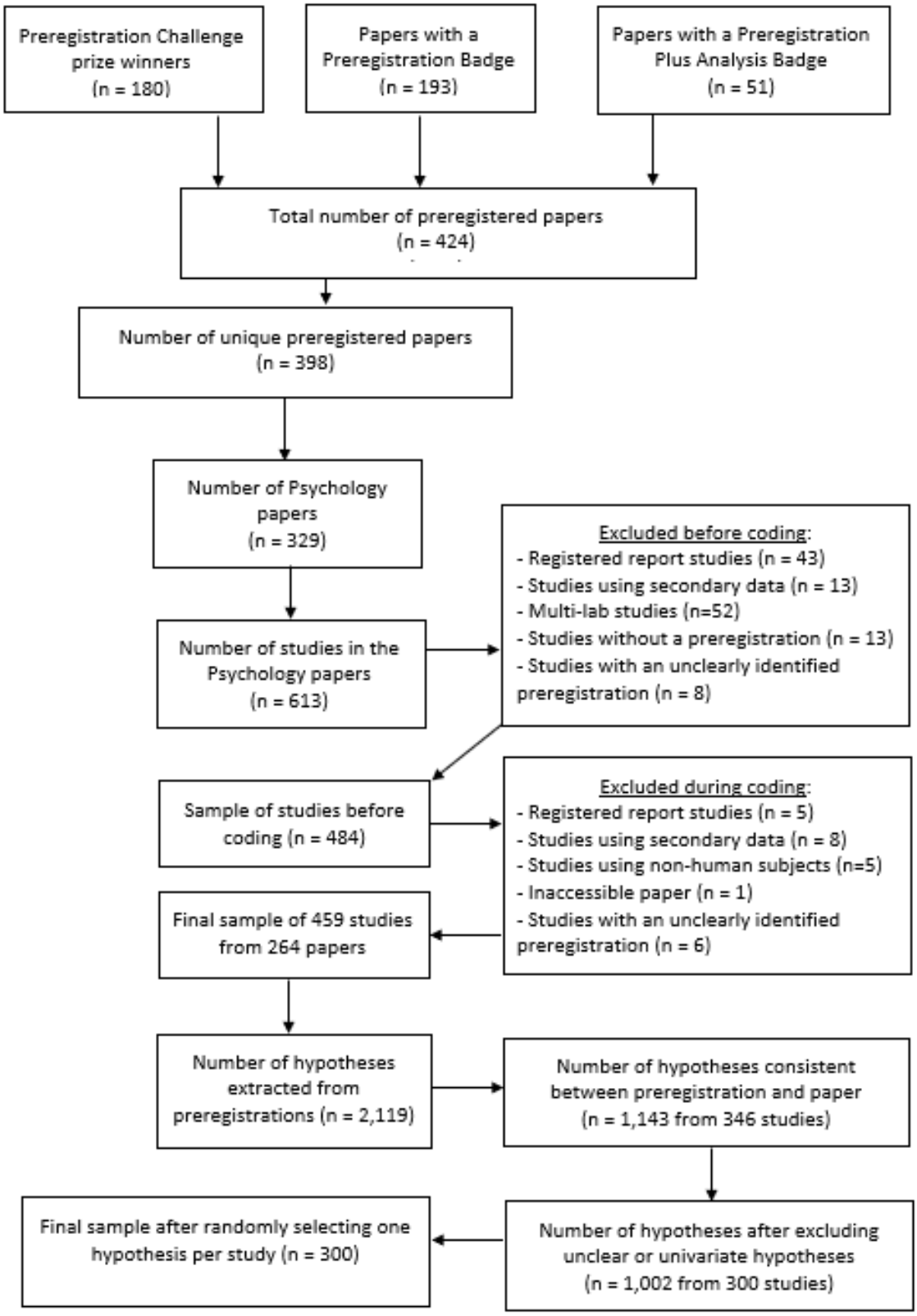


Figure 1. PRISMA flow diagram outlining the full sample selection procedure

Measuring preregistration effectiveness

We coded preregistration effectiveness using a protocol (adapted from Bakker et al., 2020) administered via Qualtrics that extracts information from the preregistration and the paper, and then helps assess preregistration producibility as well as preregistration-study consistency. The static version of this protocol can be found at https://osf.io/dpg3v. Filling out the protocol for one PSP typically took between 20 and 80 minutes, although particularly challenging pairs could take multiple hours. Each PSP was coded by two independent coders, who subsequently resolved any coding inconsistencies among each other. The 28 coders in this project were researchers interested in assessing the field of psychology from a meta-scientific perspective. They were trained using a set of ten example PSPs, and coded on average of 20.9 PSPs (min. = 4, max. = 33).

Assessing five major study parts

We extracted information about the preregistration and the paper by answering questions about five major study parts (denoted by numbers below), some of which we divided into smaller study elements (denoted by letters below):

1. the operationalization of the independent variable (in case the hypothesis implied a directional link between two or more variables) or the first variable (in case the hypothesis did not imply a directional link between two or more variables):[2]
    a. the procedure of measurement;
    b. the potential values;

[2] Because it proved to be impossible to determine whether authors intended for hypotheses to be directional, we used manipulation status as a demarcation criterion: hypotheses involving at least one manipulated variable were presumed to be directional (i.e., have an independent and dependent variable) whereas all other hypotheses were not presumed to be directional. Manipulated variables were not further divided into study elements, but measured variables were. A result of this change is that we now list five major study parts in both Table 2 and Table 3.

    c. how the variable was constructed from its components (e.g., a Likert scale based on item responses), if applicable

2. the operationalization of the dependent variable (in case the hypothesis implies a directional link between two or more variables) or the second variable (in case the hypothesis does not imply a directional link between two or more variables):
    a. the procedure of measurement;
    b. the potential values;
    c. how the variable was constructed from its components, if applicable;
3. the data collection procedure:
    a. sample size;
    b. sampling frame (i.e., the author's procedure for sampling participants);
4. the statistical model used:
    a. the model itself;
    b. the specification of the variables (e.g., whether a variable was added or changed);
    c. the manner in which the variables were used in the model (e.g., the contrasts or whether they were standardized);
5. the statistical 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 are thus crucial to restrict researcher degrees of freedom for.

Measuring preregistration producibility

We scored the five study parts on *preregistration producibility* by assessing whether they were described in a *specific* (all steps that will be taken were described) and *precise* (each of the

described steps allowed only one interpretation or implementation) manner (Bakker et al., 2020; Wicherts, et al., 2016) in the preregistration. When any part of a preregistration was described in a specific and precise manner, that part of the preregistration was scored with 2 points for producibility. When some but not all elements related to a part of the preregistration were described specifically and precisely, we awarded 1 point to that part. And, finally, when none of the elements was deemed specific and precise, we awarded 0 points.

An exception was the question about the data collection procedure, for which the protocol asked about two elements: sample size and sampling frame. If *either one of these two elements* was described specifically and precisely, the entire data collection procedure was 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, either of which would minimize researcher degrees of freedom. After taking the mean of all scores on the five major parts of the study, the preregistration could score between 0 (not producible at all) and 2 (optimally producible).

Measuring paper reproducibility

To be able to compare study parts between preregistration and paper properly, it is necessary that sufficient information about a study part is available in both the preregistration and the paper. For example, if the preregistration outlines in detail the statistical model that will be used, but the paper mentions the model only indirectly or not at all, it would be impossible to assess whether the model in the paper corresponds to the model in the preregistration. To assess whether sufficient information about a study part was provided in the paper, we also measured *paper reproducibility*. We measured this in exactly the same way as we measured preregistration producibility (see above). The term reproducibility was chosen for papers because studies

presented in papers are already carried out and thus can only be *re*produced. Studies planned in preregistrations, on the other hand, need to be carried out (produced) later and are therefore labeled 'producible'. We deemed study parts to be sufficiently comparable if a study part scored either a 1 or 2 on preregistration producibility (specifying the level of detail in the preregistration) *and* paper reproducibility (specifying the level of detail in the paper). For the parts where this was not the case, we did not compute preregistration-study consistency.

Measuring 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).

Combining producibility and consistency

To compute preregistration effectiveness for a given preregistration, we first multiplied the score for preregistration producibility with the score for preregistration-study consistency for each part separately. These multiplied scores signify how effectively each individual study part was preregistered. The highest possible score per part was 2, and could be achieved with a producibility score of 2 and a consistency score of 1. The lowest possible score was 0 and could be achieved if the producibility score and/or the consistency score were 0. We then took the mean of all of these partial effectiveness scores to get a total score that indicates how effectively

a given study was preregistered as a whole (with scores varying from 0 to 2, where higher values indicate higher effectiveness). For example, let us suppose a PSP scored on preregistration producibility 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 on preregistration-study consistency 1 point for the operationalizations of the independent and dependent variable, and 0 points for the data collection protocol, the statistical model and the inference criteria. The preregistration effectiveness score of that study would then be $(1\times1 + 2\times1 + 1\times0 + 0\times0 + 0\times0)/10 = 0.3$. We took the mean of the partial effectiveness scores, instead of the sum like we preregistered, because we believe the resulting score range of [0, 2] is more interpretable than a sum score range of [0, 10].

Assessing minor study parts

Aside from the five 'major' parts of a study outlined above, we also scored four 'minor' study parts. Note that with the term 'minor' we do not mean that these study parts are less important to preregister well, but merely that these study parts may not apply to each study design. For example, if the analysis of a study does not involve a control variable, the first minor study part below is no longer applicable. Similarly, the second minor study part is not applicable if study participants were forced to respond to all items in a questionnaire, thereby circumventing missing data other than from attrition. The minor study parts are listed below using numbers, and the elements that constitute those parts are listed using letters.

1. the operationalization of the control variable:
    a. the procedure of measurement;
    b. the potential values;

    c. how the variable was constructed from its components, if applicable;

2. how missing data was handled:
    a. the definition of missing data;
    b. how missing data were dealt with;
3. how violations of statistical assumptions were handled:
    a. which assumptions were checked;
    b. how the assumptions were checked;
    c. how violations of assumptions were dealt with;
4. exclusion criteria.[3]

We scored the minor parts in the same way as the major parts, but the scores for these parts were not used to calculate a score for the preregistration/study overall. As such, they only provide information about preregistration producibility, preregistration-study consistency, and preregistration effectiveness of the individual study parts.

Assessing whether a hypothesis is part of a replication

Information about the replication status of hypotheses was taken directly from Van den Akker et al. (2023). They assessed whether a hypothesis was part of a replication or an original study by first searching the preregistration and paper for the string "replic" and assessing whether the authors referred to the hypothesis as being part of a replication attempt. If the authors did, in either the preregistration or the paper, Van den Akker et al. coded the hypothesis as a replication hypothesis. If the authors did not, Van den Akker et al. coded the hypothesis as an

[3] In our own preregistration, we divided the exclusion criteria into two elements: the definition of the criteria and the procedure of exclusion. When inspecting the data, however, we noticed that the types of inconsistencies listed for the definition were equivalent to the types of inconsistencies listed for the procedure: they all mentioned that one or more exclusion criteria were not mentioned, added, or changed in the paper compared to the preregistration. After some discussion among coders, we realized that authors typically did not describe the procedure of exclusion (e.g., whether the criteria were determined before or after data collection, or whether exclusion was listwise or pairwise) in a preregistration or paper. We suspect that most authors assumed that listwise exclusion was self-evident and other information was superfluous. Because of this, we decided to disregard the procedure of exclusion as a study element and only regard the definition of the criteria (to assess preregistration producibility).

original hypothesis. The protocols used to assess whether a hypothesis was part of a replication can be found at https://osf.io/fdmx4 (for preregistrations) and https://osf.io/uyrds (for published papers).

Determining the comprehensiveness of preregistration templates

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.

Table 1. *Frequencies and Comprehensive Scores of the Preregistration Templates used to draft the Preregistrations in our Sample.*

Template	Freq.	Comprehensiveness
OSF Prereg template (Bowman et al., 2020)	122	24
AsPredicted (https://aspredicted.org)	112	10
Pre-Registration in Social Psychology (Van 't Veer & Giner-Sorolla, 2016)	21	14
OSF's Open Templates (https://osf.io/9j6d7; https://osf.io/haadc)	7	-
Happy Lab Pre-Registration Template (https://osf.io/yvsj8)	7	-
Replication Recipe (Brandt et al., 2014) (https://osf.io/4jd46)	1	-
Unknown	45	-
Total	315	-

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.

Determining registration dates

To assess whether preregistration effectiveness increased over time we coded the date that the preregistration was formally registered. For frozen registrations (i.e., dated registrations that cannot be altered after the registration date) on the Open Science Framework, this information is clearly listed on the right-side of the preregistration document next to the word "registered". For frozen registrations on AsPredicted, this information is clearly listed on the top of the preregistration document next to the word "public". 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 date of the first preregistration in the sample, which was 14 April 2014 (Van Zant & Moore, 2015).

Determining the type of deviations and authors' explanations

We used an open-ended question to elicit the deviations between preregistrations and papers. For example, coders could state that the sample size was higher in the paper than in the preregistration, or could state which exclusion criteria differed between preregistration and paper. We also used an open question to elicit the authors' explanations for inconsistencies between the preregistration and the actual study in the published paper, if any. Both questions are listed in the static version of the protocol, which can be found at https://osf.io/dpg3v.

Results

Descriptive statistics

Of the 300 PSPs in our sample, we classified 138 (46%) as replication studies, and the remaining 162 (54%) as original studies. Registration time, as measured by the number of months since the registration date of the first preregistration in our sample, had a mean of 38.7 months (SD = 12.4), a median of 40, and a maximum of 67.

The data used in our analyses are publicly available on the Open Science Framework (https://osf.io/vwgak). The R-code we used is also publicly available, at https://osf.io/2yzsr (for analyses regarding producibility and effectiveness) and https://osf.io/g3fra (for analyses regarding consistency).

Table 2 presents the mean scores for preregistration producibility, consistency, and effectiveness for all the separate study parts as well as the total mean scores for the five major

study parts. Table 2 also provides the frequency of the individual scores (0, 1, and 2 for producibility and effectiveness; 0, 1, and NA for consistency). The overall mean producibility score of the preregistrations in our sample was 1.33 out of 2 (N = 300, *SD* = 0.41, min. = 0, max. = 2), and the overall mean consistency score was 0.71 out of 1 (N = 57, *SD* = 0.20, min. = 0, max. = 1). The mean effectiveness score per PSP was 0.79 out of 2 (N = 300, *SD* = 0.43, min. = 0, max. = 2).[4] The correlation between the producibility scores and consistency scores was r = -.11, $t(55) = -0.82$, $p = .418$).

[4] We also assessed the preregistration effectiveness for the current study and arrived at a score of 2.0 for preregistration producibility, a score of 0.8 for preregistration-study consistency (because our sample size was not consistent), and therefore a score of 0.8 for preregistration effectiveness. For the non-essential elements, we scored 2 points for the producibility of the exclusion criteria and the handling of missing data but 0 points for the handling of violations of statistical assumptions. Finally, the exclusion criteria were not consistent because we added two criteria, whereas the missing data were inconsistent because we did not mention them in the paper at all. After making this assessment, we included a sentence about handling missing data to the paper. This shows that our assessment protocol is not only useful to assess producibility and consistency *post hoc* but also when writing up your preregistration or paper. Our (obviously biased) assessment can be found at https://osf.io/byacg.

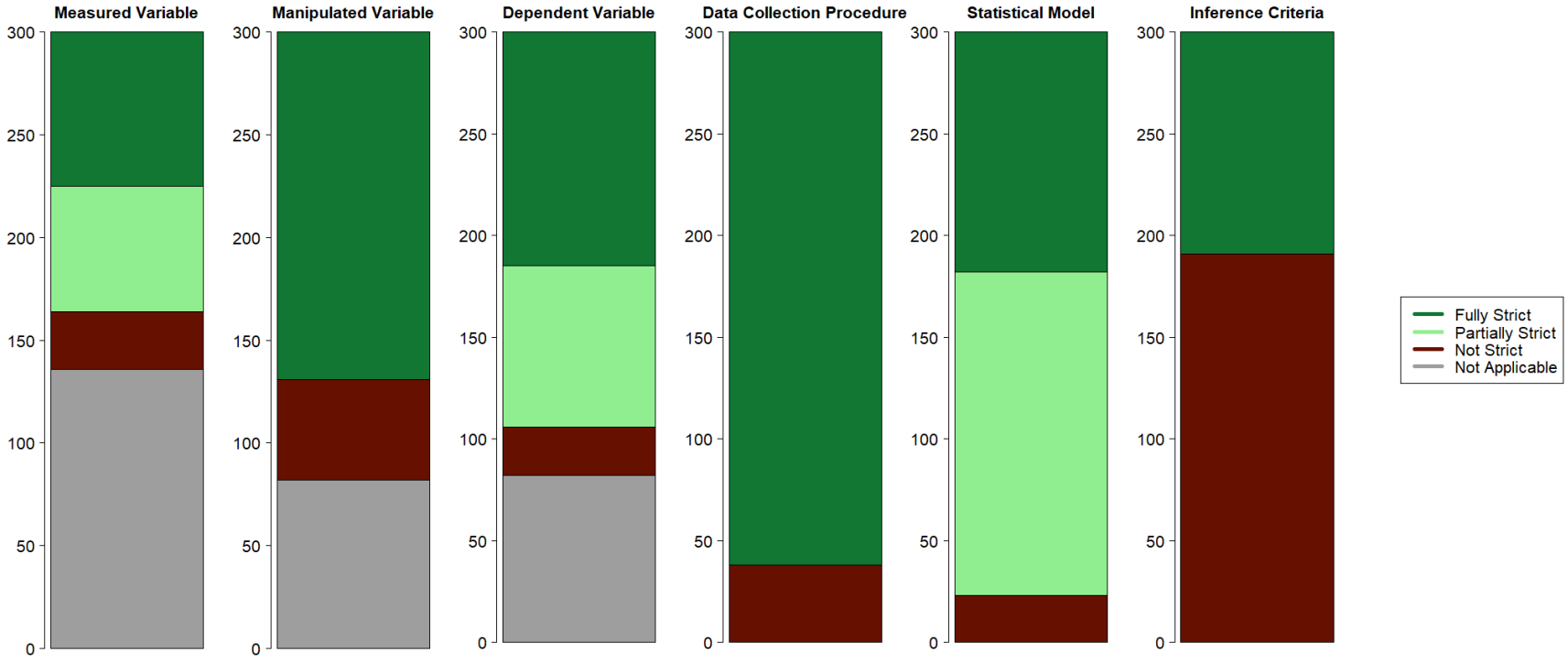
Table 2. *Overview of Producibility, Consistency, and Effectiveness Scores (with Standard Deviations) for the five Major and five Minor Study Parts, as well as Total Scores for the Major Study Parts.*

	Producibility				**Consistency**				**Effectiveness**			
Major study parts	**0**	**1**	**2**	**Score (SD)**	**0**	**1**	**NA**	**Score (SD)**	**0**	**1**	**2**	**Score (SD)**
Measured variable (*N*=164) *	28 (17%)	61 (37%)	75 (46%)	**1.29 (0.74)**	45 (27%)	91 (55%)	28 (17%)	**0.67 (0.47)**	73 (45%)	43 (26%)	48 (29%)	**0.85 (0.85)**
Manipulated variable (*N*=218) *	49 (22%)	NA	169 (78%)	**1.55 (0.84)**	8 (4%)	146 (67%)	64 (29%)	**0.95 (0.22)**	63 (33%)	NA	155 (67%)	**1.42 (0.91)**
Dependent variable (*N*=218) *	24 (11%)	79 (36%)	115 (53%)	**1.42 (0.68)**	56 (26%)	131 (60%)	31 (14%)	**0.70 (0.46)**	87 (40%)	59 (27%)	72 (33%)	**0.93 (0.85)**
Data collection procedure (*N*=300)	38 (13%)	NA	262 (87%)	**1.75 (0.67)**	173 (58%)	87 (29%)	40 (13%)	**0.33 (0.47)**	213 (71%)	NA	87 (29%)	**0.58 (0.91)**
Statistical model (*N*=300)	23 (8%)	159 (53%)	118 (39%)	**1.32 (0.61)**	121 (40%)	135 (45%)	44 (15%)	**0.53 (0.50)**	165 (55%)	89 (30%)	46 (15%)	**0.60 (0.74)**
Inference criteria (*N*=300)	191 (64%)	NA	109 (36%)	**0.73 (0.96)**	10 (3%)	94 (31%)	196 (65%)	**0.90 (0.30)**	206 (69%)	NA	94 (31%)	**0.63 (0.93)**
Total scores				**6.65 (2.04)**				**3.53 (1.00)**				**3.96 (2.12)**
	Producibility				**Consistency**				**Effectiveness**			
Minor study parts	**0**	**1**	**2**	**Score (SD)**	**0**	**1**	**NA**	**Score (SD)**	**0**	**1**	**2**	**Score (SD)**
Measured control variable (*N*=20) **	7 (35%)	5 (25%)	8 (40%)	**1.05 (0.89)**	9 (45%)	4 (20%)	7 (35%)	**0.31 (0.48)**	16 (80%)	1 (5%)	3 (15%)	**0.35 (0.75)**
Manipulated control variable (*N*=23) **	5 (22%)	NA	18 (78%)	**1.57 (0.84)**	0 (0%)	18 (78%)	5 (22%)	**1.00 (0.00)**	5 (22%)	NA	18 (78%)	**1.57 (0.84)**
Exclusion criteria (*N*=300)	68 (23%)	NA	232 (77%)	**1.55 (0.84)**	86 (29%)	106 (35%)	108 (36%)	**0.55 (0.50)**	194 (65%)	NA	106 (35%)	**0.71 (0.96)**
Missing data (*N*=300)	159 (53%)	21 (7%)	120 (40%)	**0.87 (0.96)**	9 (3%)	37 (12%)	254 (85%)	**0.80 (0.40)**	263 (88%)	2 (1%)	35 (12%)	**0.24 (0.65)**
Statistical assumptions (*N*=300)	279 (93%)	17 (6%)	4 (1%)	**0.08 (0.32)**	2 (1%)	6 (2%)	292 (97%)	**0.75 (0.46)**	294 (98%)	5 (2%)	1 (0%)	**0.02 (0.17)**

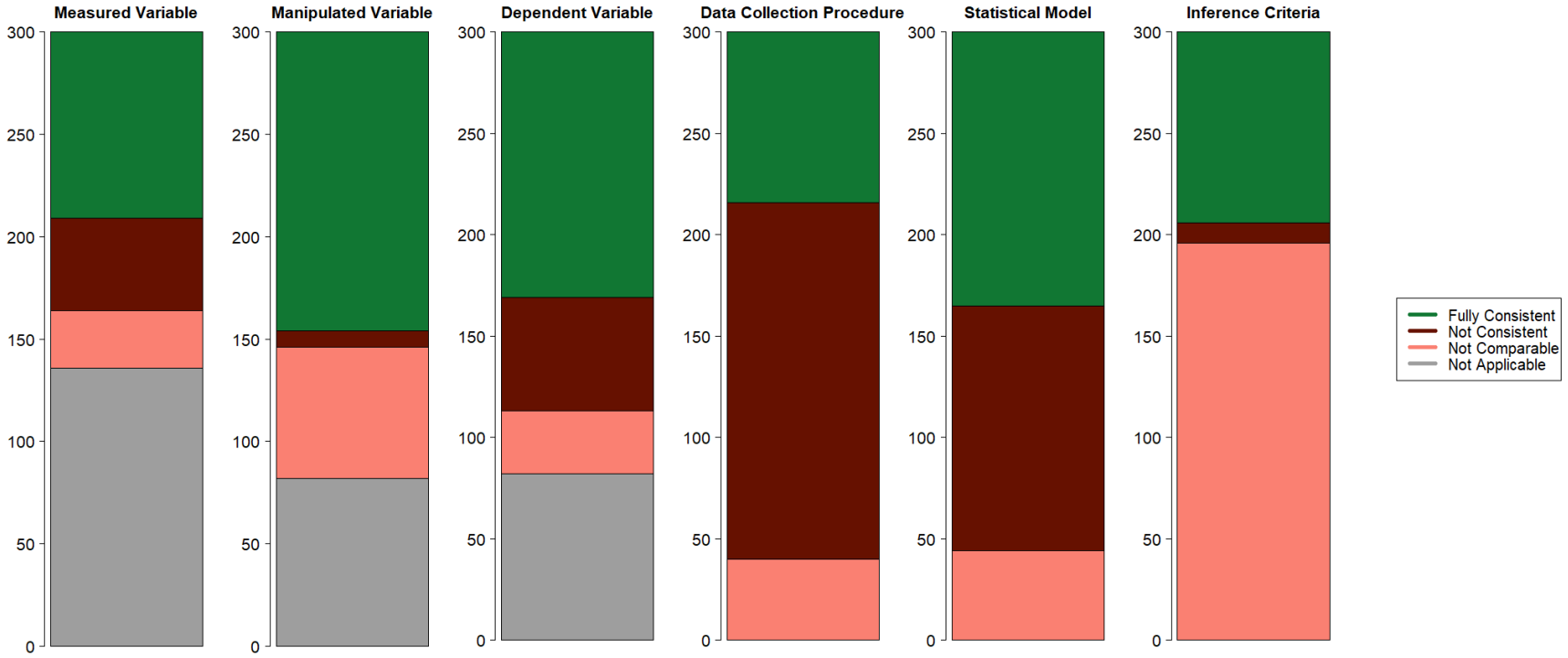
Note Table 2. The single asterisk in Table 2 highlights that we had 82 hypotheses without a directional relationship and therefore two measured variables, and 218 hypotheses with a directional relationship and therefore a manipulated variable and a dependent variable. The double asterisk in Table 2 highlights that we only had 46 hypotheses with one or more control variables. Twenty of those were part of a non-directional hypothesis, and 26 were part of a directional hypothesis.

Figure 2. *Preregistration producibility scores (a), preregistration-study consistency scores (b), and effectiveness scores (c)*

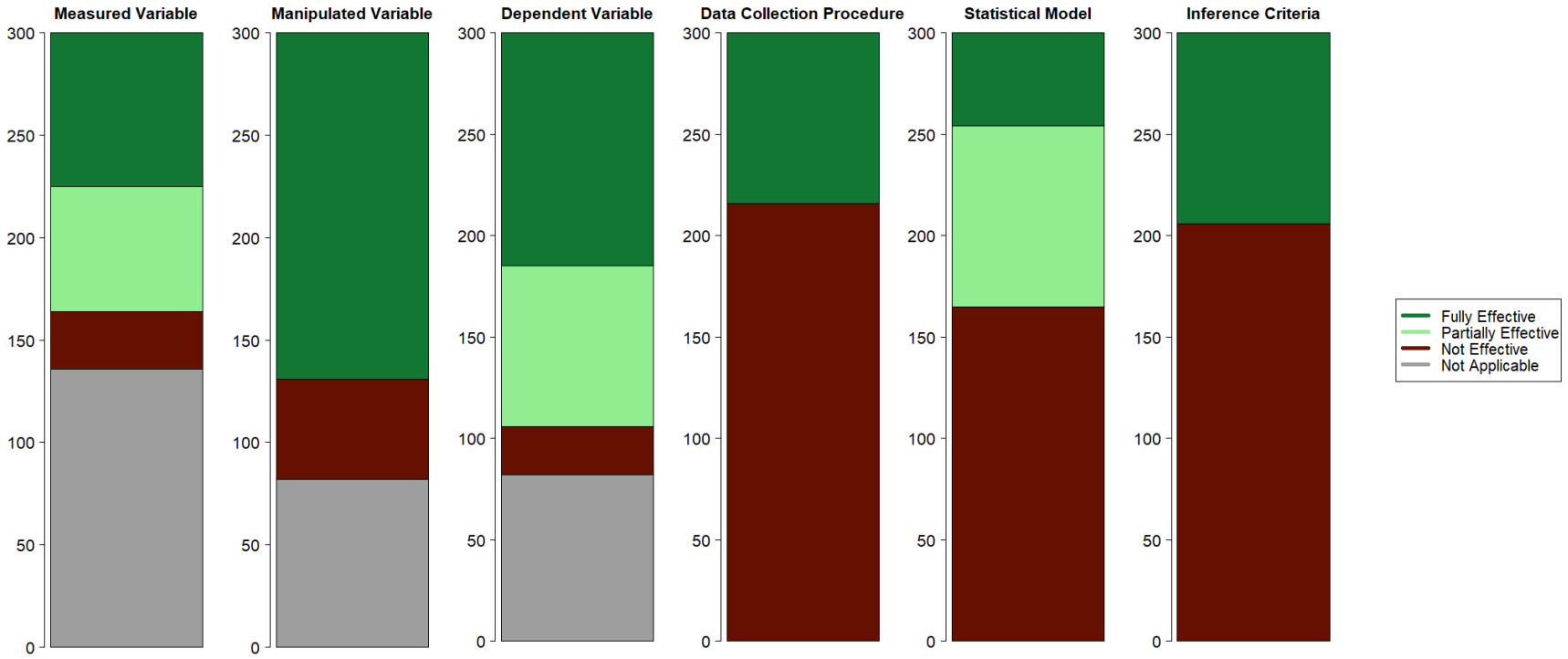
a)



b)


Measured Variable
Manipulated Variable
Dependent Variable
Data Collection Procedure
Statistical Model
Inference Criteria
300
250
200
150
100
50
0
Fully Consistent
Not Consistent
Not Comparable
Not Applicable

c)


Measured Variable
Manipulated Variable
Dependent Variable
Data Collection Procedure
Statistical Model
Inference Criteria
300
250
200
150
100
50
0
Fully Effective
Partially Effective
Not Effective
Not Applicable

Note that the consistency scores in Table 2 indicate the proportion of PSPs for which that study part was consistent out of all PSPs for which the study parts could be compared between preregistration and paper. For example, the statistical model could be compared 256 times, of which 121 (47%) were consistent. Because the inference criteria were almost never explicitly stated in the paper, we used implicit consistency instead. That is, we checked whether the authors' conclusion about the statistical result was in line with their preregistered inference criterion. For example, if the preregistration specified $\alpha = .01$ and the paper drew a conclusion in the form of "we found an effect of X on Y, $p = .007$" we would consider this as consistent and score the consistency of inference criteria with 1 point. However, if the paper specified $\alpha = .01$ and stated "we found an effect of X on Y, $p = .023$" we would consider this as inconsistent (and allocate 0 points) as a different criterion seems to be used. Note that this was a deviation from our preregistration, but that this deviation did not influence our measurement of preregistration producibility and paper reproducibility.

To allow the calculation of preregistration effectiveness for each individual PSP, all 'NA' responses for consistency were recoded to scores of 0. As can be seen in Table 2, we found mean efficiency scores below 1 (out of 2) for all study parts except for manipulated variables (1.42). Generally, the operationalizations of the variables (measured, manipulated, and dependent) were more effectively preregistered than the other study parts. A visualization of the scores for producibility, consistency, and effectiveness can be found in Figures 2-4.

We also collected data about the study elements that constitute the different study parts. In Table 3, we see that within each study part, the elements were often more or less equally producible (see the column 'Prereg producibility'). Consistency between preregistration and paper with regard to study elements (computed only for elements that were at least partially

producible in the preregistration *and* reproducible in the paper) is outlined in the column 'Consistency' in Table 3. In the final column of Table 3 ('Explanations'), we provide information about the presence of authors' explanations for preregistration deviations in the final paper. Explanations of deviations were rarely provided, especially for study elements where inconsistencies were rare. We also assessed what kind of inconsistencies were most common by exploring the three study parts with the most inconsistencies: the data collection procedure, the exclusion criteria, and the statistical model. Our categorization of inconsistencies for these three study parts can be found at https://osf.io/crd3u.

Table 3

Overview of the Preregistration Producibility, Paper Reproducibility, Consistency and Authors' Explanations for Preregistration Deviations for Each Study Element.

	Prereg Producibility	Paper Reproducibility	Consistency	Explanations
Measured variable (*N*=164)				
Procedure of measurement	102 (62%)	125 (76%)	89 / 92 (97%)	0 / 3 (0%)
Potential values	87 (53%)	108 (66%)	69 / 73 (95%)	1 / 4 (25%)
Procedure to construct composite (*N*=73)	37 (51%)	45 (62%)	20 / 23 (87%)	0 / 3 (0%)
Manipulated variable (*N*=218)	169 (78%)	202 (93%)	146 / 154 (95%)	0 / 8 (0%)
Dependent variable (*N*=218)				
Procedure of measurement	184 (84%)	199 (91%)	150 / 163 (92%)	1 / 13 (8%)
Potential values	150 (69%)	177 (81%)	115 / 120 (96%)	0 / 5 (0%)
Procedure to construct composite (*N*=134)	83 (62%)	76 (57%)	54 / 57 (95%)	0 / 3 (0%)
Measured control variable (*N*=20)				
Procedure of measurement	13 (65%)	12 (60%)	8 / 8 (100%)	0 / 0
Potential values	12 (60%)	8 (40%)	5 / 7 (71%)	0 / 2 (0%)
Procedure to construct composite (*N*=8)	3 (38%)	2 (25%)	1 / 1 (100%)	0 / 0
Manipulated control variable (*N*=23)	18 (78%)	22 (96%)	18 / 23 (78%)	0 / 5 (0%)
Data collection procedure (*N*=300)				
Exact sample size	178 (59%)	176 / 178 (99%)	49 / 176 (28%)	26 / 120 (22%)
Sampling frame	84 (28%)	68 / 84 (81%)	35 / 52 (67%)	6/17 (35%)
Exclusion criteria (*N*=300)	232 (77%)	225 (75%)	106 / 192 (55%)	13 / 86 (15%)
Missing data (*N*=300)				
Definition of criteria	123 (41%)	55 (18%)	35 /42 (83%)	0 / 7 (0%)
Method of handling	138 (46%)	53 (18%)	39 / 42 (93%)	0 / 3 (0%)
Statistical model (*N*=300)				
Which model was used	256 (85%)	244 (81%)	162 / 216 (75%)	11 / 54 (20%)
Specification of variables	254 (85%)	263 (88%)	207 / 226 (92%)	5 / 22 (23%)
How the variables are used in the model	128 (43%)	110 (37%)	66 / 75 (88%)	5 / 9 (56%)
Statistical assumptions (*N*=300)				
Which assumptions are checked	20 (7%)	19 (6%)	8 / 8 (100%)	0 / 0
How assumptions are checked	4 (1%)	8 (3%)	1 / 1 (100%)	0 / 0
What is done in case of violations	19 (6%)	18 (6%)	6 / 6 (100%)	0 / 0
Inference criteria (*N*=300)	109 (36%)	37 (12%)	94 / 104 (90%)	1 / 10 (10%)

Hypothesis tests

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 producibility (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). In contrast to our hypothesis, we found no evidence that replication studies were preregistered more producibly (M = 1.31) than original studies (M = 1.35), B_1 = -0.001, t(234,7) = -0.01, 99% CI = [-0.10, 0.10], p = .988, nor that preregistration-study consistency was higher for replication studies (M = 0.72) compared to original studies (M = 0.70), B_1 = -0.001, t(49.2) = -0.016, 99% CI = [-0.14, 0.14], p = .988. Consequently, the effectiveness of preregistration was not higher for replication studies (M = 0.79) than for original studies (M = 0.80), B_1 = 0.03, t(294.1) = 0.55, 99% CI = [-0.10, 0.15], p = .582. The regressions related to all hypotheses are presented in Table 4. We computed unstandardized coefficients for all preregistered hypothesis tests and used an alpha level of .01, as preregistered.

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). In line with our hypothesis, we found that preregistrations based on the OSF template were more producible (M

= 1.62) than preregistrations based on the AsPredicted template (M = 1.15), $B_1 = 0.44$, $t(170.4) = 8.30$, 99% CI = [0.30, 0.59], $p < .001$, and the Social Psychology template (M = 1.31), $B_1 = 0.30$, $t(97.2) = 3.32$, 99% CI = [0.07, 0.54], $p = .001$. Similarly, OSF preregistrations (M = 0.98) were more effective than both AsPredicted preregistrations (M = 0.69), $B_1 = 0.30$, $t(144.3) = 4.62$, 99% CI = [0.16, 0.44], $p < .001$, and the Social Psychology preregistrations (M = 3.21), $B_1 = 0.34$, $t(90.6) = 2.79$, 99% CI = [0.03, 0.66], $p = .006$. The higher effectiveness in OSF templates related to AsPredicted templates was likely due to differences in producibility, as there was no significant difference in preregistration-study consistency between OSF templates (M = 0.71) and AsPredicted templates (M = 0.67), $B_1 = 0.04$, $t(48.2) = 0.35$, 99% CI = [-0.28, 0.36], $p = .730$. We could not test for a difference in preregistration-study consistency between OSF templates and Social Psychology templates because preregistration-study consistency could not be assessed for any of the Social Psychology templates as insufficient information was present to compare preregistrations and papers. We computed unstandardized coefficients for all preregistered hypothesis tests and used an alpha level of .01, as preregistered.

Finally, to test whether preregistration effectiveness improved over time (Hypothesis 3), we again ran the same three multilevel regressions as for Hypothesis 1, with the addition of a continuous variable denoting the number of months between the registration date of the preregistration and the registration date of the first preregistration in our sample (see M3a, M3b, M3c in Table 4). As no effect of time was observed in any of the three analyses, we conclude that there was not sufficient evidence that the quality of preregistration improved over time (producibility: $B_1 = 0.001$, $t(296.8) = 0.38$, 99% CI = [-0.004, 0.006], $p = .704$; preregistration-study consistency: $B_1 = 0.02$, $t(50.2) = 1.46$, 99% CI = [-0.003, 0.01], $p = .151$; and effectiveness: $B_1 = -0.001$, $t(266.0) = -0.61$, 99% CI = [-0.01, 0.004], $p = .545$). We computed

unstandardized coefficients for all preregistered hypothesis tests and used an alpha level of .025, as preregistered.

Exploratory analyses

The results outlined above indicate whether our sample of studies were preregistered sufficiently producible, consistent, and consequently, effective. While these results indicate the potential for *p*-hacking in a certain study, they do not speak to whether *p*-hacking actually took place. Because the research process largely takes place behind the closed doors of offices, direct evidence for *p*-hacking is almost impossible to attain. However, we can use the proxy of statistical significance to explore whether more producible, more consistent, and more effective preregistrations are associated with a lower rate of statistical significance, which would suggest less *p*-hacking in these studies. To test this, we linked each study's producibility scores, consistency scores, and effectiveness scores to whether the assessed hypothesis (see the section 'Selection of preregistered studies' for a description of how we selected hypotheses) yielded a statistically significant result. We used multilevel analyses with study as Level 1 and paper as Level 2. Data about statistical significance was derived from Van den Akker et al. (2023). The analysis for consistency did not converge because of the low number of data points (24 consistency scores with a statistically significant result, and 24 consistency scores with a non-significant result). Furthermore, we found no evidence of an association of preregistration producibility and effectiveness with statistical significance (producibility: $B_1 = -0.14$, $t(200.4) = -1.71$, 99% CI = [-0.29, 0.02], $p = .088$; effectiveness: $B_1 = -0.12$, $t(227.5) = -1.72$, 99% CI = [-0.26, 0.02], $p = .088$).

Table 4. Results of our tests of Hypothesis 1, Hypothesis 2, and Hypothesis 3 (M3a, M3b, and M3c).

Parameters	M1a	M1b	M1c	M2a1	M2b1	M2c1	M2a2	M2b2	M2c2	M3a	M3b	M3c
Fixed effects (standard errors)												
Intercept	**1.34* (0.03)**	**0.71* (0.04)**	**0.80* (0.03)**	**1.16* (0.05)**	**0.66* (0.12)**	**0.67* (0.06)**	**1.30* (0.08)**	-	**0.65* (0.11)**	**1.32* (0.08)**	**0.50* (0.15)**	**0.85* (0.09)**
Level 1												
Replication	-0.001 (0.04)	-0.001 (0.05)	0.03 (0.05)	0.01 (0.04)	0.01 (0.06)	0.05 (0.05)	0.01 (0.04)	-	0.01 (0.07)	-0.002 (0.04)	-0.02 (0.05)	0.03 (0.05)
OSF vs. AP	-	-	-	**0.44* (0.05)**	0.04 (0.12)	**0.30* (0.07)**	-	-	-	-	-	-
OSF vs. SP	-	-	-	-	-	-	**0.30* (0.09)**	-	**0.34* (0.12)**	-	-	-
Months	-	-	-	-	-	-	-	-	-	0.001 (0.002)	0.004 (0.003)	-0.001 (0.002)
Random effects												
Paper-level	0.14	0.01	0.10	0.08	0.02	0.09	0.08	-	0.09	0.14	0.01	0.10

Note. Model 1a refers to the model testing the first part of Hypothesis 1 (producibility), while Model 1b and 1c test the second (consistency) and third (effectiveness) part, respectively. The same holds for the models M2 and M3, which test Hypothesis 2 and Hypothesis 3. * indicates $p < .01$.

Discussion

The number of preregistrations has greatly increased in recent years (Pennington, 2023). However, empirical evidence has been lacking as to whether preregistration achieves its goal of restricting researcher degrees of freedom. In this study, we assessed 300 preregistered psychology studies on how producible the preregistrations were and how consistent the preregistrations were with their corresponding papers. We found a mean producibility score of 1.33 out of 2 and a mean consistency score of 0.71 out of 1. Combining producibility and consistency, we found a mean score for preregistration effectiveness of 0.79 out of 2. These scores indicate that over the years 2014-2020, the practice of preregistration was not as effective as it could have been, either because preregistrations were not producible enough and/or because researchers generally deviated substantially from the preregistration. As such, the possibility for the opportunistic use of researcher degrees of freedom remained after preregistration. This finding is in line with earlier studies that assessed preregistration in economics and political science (Ofosu & Posner, 2021), in gambling (Heirene et al., 2021), and in a cross-disciplinary sample (Bakker et al., 2020).

When focusing on different study parts, we found that the operationalizations of the variables were preregistered more producibly than other study parts and that the data collection procedure, the statistical model, and the exclusion criteria were the least consistent between preregistration and paper. Moreover, we rarely encountered any concrete explanations by the authors for inconsistencies between preregistrations and papers. These results replicate previous findings that study parts that are more effectively preregistered tend to be tied to the operationalization of variables (however, see Sarafoglou, Hoogeveen, & Wagenmakers, 2023). This may be the case because the variables are the foundation of a scientific study, and

researchers are more invested in properly preregistering them. More cynically, it could be argued that it is easier to *p*-hack during the statistical analysis than in the operationalization of the variables, simply because there are more researcher degrees of freedom related to the statistical analysis (Wicherts, et al., 2016). Future meta-scientific research could investigate the research process in detail to comprehensively identify the different ways a researcher could steer a study in a certain direction, and which of these ways generally biases the results most (see Stefan & Schönbrodt, 2023). Such research would shed light on which study parts to give priority when preregistering a study.

We also carried out three novel hypothesis tests. Hypothesis 1, stating an association between replication status and preregistration producibility and consistency, was not supported. Our rationale for expecting more producible preregistrations for replication studies than for original studies was that information about the to-be-replicated study should be readily available in the paper, meaning that authors could simply include that information in their preregistration. However, this study found that study designs were often not comprehensively reported in papers with prereregistered studies, and the same issue likely holds for papers with non-preregistered studies. The vast number of reporting guidelines designed to help researchers report study details more comprehensively (see the EQUATOR Network, Simera et al., 2010) confirms this.

Additionally, we argued that preregistration-study consistency might be better for replication preregistrations because the principal goal of a replication study is to mimic the primary study. Authors of replication studies should therefore be more motivated to adhere to their preregistration than authors of original studies. However, there could be many other factors at play that influence preregistration-study consistency. It could be, for example, that the hypotheses or methodological designs of preregistered studies are simpler, which could have

counteracted any motivation effect in researchers as it should be easier to adhere to a simple preregistered plan than a difficult one. Alternatively, it could be that there is a difference in motivation between researchers who conduct a replication study and researchers who conduct original studies, but that this does not hold for researchers who preregister because their motivation to adhere to the preregistration is high regardless of study type. Finally, it could simply be that our initial intuition about (researchers conducting) replication studies was wrong. In any case, we did not find sufficient evidence to establish that replication studies involve more effective preregistrations than original studies.

In line with Hypothesis 2, preregistrations based on more comprehensive templates were generally more producible and more effective than preregistrations based on less comprehensive templates. However, consistency was not significantly higher. That more comprehensive templates did not yield more consistency between preregistrations and papers may be due to a faulty assumption. We assumed that people using more comprehensive templates would be more motivated to effectively preregister their study than people using less comprehensive templates, as comprehensive templates require more work. However, it may well be that the choice of preregistration template is determined by other factors like the specific field one is in, one's knowledge of the digital Open Science space, or simply random events.

In contrast with Hypothesis 3, we did not find evidence that preregistrations became more effective over time. One reason for this could be that the early adopters of preregistration (i.e., those who authored the earliest preregistrations in our sample) were already more effective at preregistration to begin with. This would make intuitive sense because their early uptake indicates an intrinsic interest in preregistration. Our data do not allow a test of this explanation because we do not know who the early adopters are in our dataset. It could for example be that a

researcher conducted preregistrations early on outside of the scope of the Preregistration Challenge or Preregistration Badge infrastructure. Building on our results with a survey about the adoption of preregistration practices could be informative to assess the plausibility of this explanation. Alternatively, it could be that our operationalization of time did not allow a valid test of the hypothesis. Ideally, one would assess the association between time and the effectiveness of preregistration *within* authors, but the short time period yielded almost no repeated first authors, thus ruling out this approach. Future studies that use a wider time period may be able to test this hypothesis more effectively. Finally, it may well be that preregistration skills have not improved over time because learning is difficult if one is not aware of one's mistakes. While preregistration templates can function as a building block of good preregistrations, these templates often do not specify common preregistration mistakes nor detailed examples of good preregistrations. The current study established common preregistration mistakes and identified a host of high-quality preregistrations. Hopefully, these will be used by researchers to improve their preregistration skills.

In general, we did not foresee that there would be so many situations (about 15% of cases) where we could not assess the consistency between preregistration and paper for a certain study part. This occurred when either the preregistration, the paper, or both did not provide sufficient information to allow a comparison. A consequence of this lack of proper reporting is that our statistical tests about preregistration-study consistency had less statistical power than anticipated, particularly for finding small true effect sizes. We urge researchers to explicitly mention in research papers all the study parts discussed in the preregistration, even if the information seems trivial or irrelevant. If there is insufficient information to compare

preregistration and paper, it is unclear whether researcher degrees of freedom were left open and readers are forced to conclude that *p*-hacking would have been possible.

Our exploratory analyses assessing the relationship between preregistration effectiveness and statistical significance did not provide sufficient evidence for the claim that more effective preregistrations better prevent *p*-hacking than less effective preregistrations. One possible explanation for the absence of an association is that we investigated not only primary hypotheses but hypotheses that were indicated by one of seven keywords (see the section 'Selection of preregistered studies'). It is plausible that primary hypotheses have a higher likelihood of being statistically significant because they were expected a priori to be supported (and that this was the reason to do the study in the first place), or because the hypothesis was selected a posteriori to become the primary hypothesis *because* it was statistically significant. In addition, the statistical power for our exploratory analyses was likely low because of the small number of studies we could assess (N=233). Our study suggests that if an association exists, it is likely small (95% confidence interval = [-0.26, 0.02]), which raises the question of whether the added time and effort associated with an effective preregistration (Sarafoglou et al., 2023) over a less effective preregistration is worth it.

Importantly, there could also be other reasons for why more effective preregistrations would be associated with a higher likelihood of statistically significant effects. For example, researchers who diligently and conscientiously write up a producible preregistration might also conduct a priori power analyses more diligently and conscientiously, leading to higher sample sizes and higher statistical power. In that case, more effective preregistrations would also be related to statistical significance, but the contributing factor would be the researchers, not preregistration itself. Because of the implications of finding an association between

preregistration and statistical significance of hypothesis tests and because of the possibility of confounding factors, we recommend conducting a confirmatory test of this hypothesis in a high-powered future study. In addition, similar confirmatory tests could be initiated to assess the validity of other benefits of preregistration (see Lakens, 2019; Sarafoglou et al., 2023; Wagenmakers & Dulith, 2016) that so far have remained largely theoretical.

Overall, our results suggest there is room for improvement in the practice of preregistration, but there are several limitations of our study that we need to consider. For example, the preregistration effectiveness scores for the data collection procedure and the statistical model may be low because our coding was quite strict. In the case of the data collection procedure, one could argue that our coding was *too* strict, specifically in the cases where an exact sample size was preregistered. As can be seen in Figure 2, many sample sizes only differed slightly between preregistration and study, sometimes by only one or two participants. As preregistered, we labeled each deviation, however small, as an inconsistency, yielding a consistency score and an effectiveness score of zero. However, slight deviations in sample size would yield only a limited potential for *p*-hacking as the addition or subtraction of one or two participants would probably not change a statistically nonsignificant ($p > .05$) to a statistically significant result ($p < .05$). Yet, such *p*-hacking is still possible. Indeed, optional stopping has been argued as one particularly potent way of getting a statistically significant result (Hartgerink, Van Aert, Nuijten, Wicherts, & Van Assen, 2016), especially in combination with other opportunistic uses of researcher degrees of freedom (Wicherts, 2017). As such, we maintain that any deviation from an explicitly stated sample size should be labeled as an inconsistency. We encourage readers to analyze our data (accessible at https://osf.io/vwgak) using their own definition of a sample size deviation to draw their own conclusions.

In the case of the statistical model, one issue is that the low scores on producibility could have arisen because we included the study element 'the way the variables were used in the model'. This element reflected factors such as mean-centering predictors or the use of robust standard errors. However, one might argue that proper preregistrations do not always require such detailed information. For example, some model specifications are so standard that mentioning them in a preregistration or paper would be seen as superfluous (e.g., the use of ordinary least squares estimation instead of weighted least squares estimation). The point here is that authors do not always need to specify detailed information about a statistical model other than the essential information captured by the other elements of the statistical model: the model itself, and the specification of the variables. However, if the authors did not specify any additional information, we did score the element 'the way the variables were used in the model' with zero points for producibility, and thus effectiveness. To correct for this, we did an exploratory analysis where we recalculated the producibility score, consistency score, and effectiveness score for the statistical model, which became 1.70 (was 1.32), 0.65 (was 0.53), and 1.00 (was 0.60), respectively. These updated scores better align with Ofosu and Posner (2021), who found that not only the variables, but also the statistical model was generally well-preregistered.

Furthermore, it could also be argued that our scoring of producibility was arbitrary. Study parts could get a score of not producible (score of 0), partially producible (score of 1), or fully producible (score of 2). Alternative scoring methods may be just as valid. As a robustness check, we therefore rescored the producibility variable in two alternative ways to see whether that affected our inferences. First, we used a binary score in which a study part received a score of 1 if at least one of the study elements was deemed to be producibly described (and 0 otherwise).

Second, we used a binary score in which a study part received a score of 1 if all study elements were deemed to be producibly described (and 0 otherwise). Note that both ways correspond to most extreme scoring rules, with the difference between 'not producible' and 'partially producible' being infinitely larger than the difference between 'partially producible' and 'fully producible' (0,1,1), or infinitely smaller (0,0,1). For both these scoring methods, the results that were (not) statistically significant in the original analyses were (not) statistically significant in the new analyses, with the coefficients being in the same direction. The detailed results of the robustness analyses can be found at https://osf.io/3mxfs.

Our results also mimic those of previous studies with regard to explanations for deviations. Like Claesen et al. (2020) and Heirene et al. (2021), we found that authors rarely explain inconsistencies between preregistrations and papers. This is problematic because such omissions mean that readers cannot assess whether the deviations were reasonable and the severity of the test may be compromised (Lakens, 2019). We recommend researchers to document the deviations from a preregistration explicitly, comprehensively, and transparently, including a rationale for why the deviations occurred and how the deviation could impact the results, perhaps employing Preregistration Planning and Deviation Documentation (Van 't Veer et al., 2019).

While we did count the number of times that the authors explained a deviation from the preregistration, we did not report on whether these deviations were reasonable because we do not presume to have the expertise required to make that judgment for each individual study. However, we do have the wordings used by the authors to explain their deviations, so interested readers could do a deep dive into our data to assess the validity of preregistration deviation explanations in psychology. In general, our data is freely available for anyone to check our

coding efforts or to answer their own research questions. We believe the data we collected can be a valuable resource for meta-researchers.

Aside from comparisons with fields in the social sciences, it may also be informative to compare our results to studies in biomedicine, a field that has seen much meta-research on the topic of preregistration (often called registration in this discipline; Rice and Moher, 2019). Researchers in the United States have been mandated to register clinical trials as early as 1997 (Food and Drug Administration Modernization Act of 1997, 1997), making it possible to assess the producibility of these registrations and their consistency with the subsequent report. In general, these studies focus primarily on study outcomes and review studies show that a large proportion of clinical trial papers involves the addition, removal, or change of a primary outcome (Dwan et al., 2013: 40-62%; Jones et al., 2015: 65%; Li et al., 2018: 14% to 100%; Thibault et al., 2021: 10% to 68%). The reviews that also assess other study parts (Li et al., 2018; Thibault et al. 2021) find, like in the social sciences, that the exclusion criteria, sample size, and statistical analysis (including subgroup analyses) are the areas in which discrepancies occur most often. In review, the prevalence of discrepancies between preregistration and paper seems to be similar in the social sciences and biomedical sciences, also in terms of the types of discrepancies. A systematic comparison between the social sciences and biomedical sciences is outside the scope of this paper but would be an interesting meta-research pursuit to follow up on.

While preregistrations serve to lock in temporal relationships between planning and conducting, it should also be noted that the present study assumes that such temporal relationships were guaranteed in all the preregistrations we analyzed. However, preregistrations could be created after experiments have been carried out (Yamada, 2018). This is a problem inherent in preregistration itself, but this type of practice would be less likely to be observed in

registered reports, where experimental protocols are peer-reviewed and almost always revised before experiments are conducted (Chambers & Tzavella, 2022).

Similarly, an alternative to 'regular' preregistration could be analysis blinding, where researchers develop their analysis plan using data in which a third party removed any potentially biasing information. Sarafoglou, Hoogeveen, and Wagenmakers (2023) found that analysis blinding leads to higher consistency between preregistered and actual analysis than preregistration. For example, they found that the analysts in their study who practiced analysis blinding deviated with their exclusion criteria 2% of the time, while that was 16% for those who practiced preregistration. This practice thus seems to be a promising tool for researchers aiming to ensure the confirmatory status of their statistical analyses.

Finally, a way to improve consistency would be to have peer reviewers explicitly compare the preregistration and the actual study. Based on our experience, this comparison is often carried out haphazardly or is not carried out at all. A feasibility study on discrepancy review (TARG Meta-Research Group and Collaborators, 2022) showed that it can be effective and could feasibly be introduced as a regular practice. However, an important issue with this idea is that discrepancy review takes extra time, while reviewers already invest many unpaid hours in peer reviewing for scientific journals. If this burden increases, potential reviewers could become less tempted to accept peer review requests, leading to a potential breakdown of the system. While the feasibility study found that the extra time investment was not excessive, a full trial that looks at secondary effects of discrepancy review is desirable.

In sum, our results extend the results of other studies, making it increasingly clear that, while some researchers are good preregistrationers, much needs to be improved with regard to study preregistration. To unlock the full potential of preregistration, researchers in psychology

and likely other fields should aim to write more producible preregistrations, adhere to these preregistrations more faithfully, and in case of deviations, more transparently report them. The creation of more comprehensive templates, and specific training modules to improve preregistration skills would be beneficial in this regard.

References

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

Bishop, D. V. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, *73*(1), 1-19. https://doi.org/10.1177/1747021819886519

Bowman, S., DeHaven, A. C., Errington, T., Hardwicke, T. E., Mellor, D. T., Nosek, B. A., & Soderberg, C. K. (2020). OSF Prereg Template. https://doi.org/10.31222/osf.io/epgjd

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29-42.

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

Dwan, K., Gamble, C., Williamson, P. R., Kirkham, J. J., & Reporting Bias Group. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PloS one*, *8*(7), e66844.

Food and Drug Administration Modernization Act of 1997. (1997). Public Law 105-15. Retrieved from https://www.govinfo.gov/content/pkg/PLAW-105publ115/pdf/PLAW-105publ115.pdf

Hartgerink, C. H., Van Aert, R. C., Nuijten, M. B., Wicherts, J. M., & Van Assen, M. A. (2016). Distributions of p-values smaller than .05 in psychology: what is going on?. *PeerJ*, *4*, e1935.

Haven, T. L., & Van Grootel, D. L. (2019). Preregistering qualitative research. *Accountability in Research*, *26*(3), 229-244.

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

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, *62*(3), 221-230.

Li, G., Abbade, L. P. F., Nwosu, I., Jin, Y., Leenus, A., Maaz, M., Wang, M., Bhatt, M., Zielinski, L., Sanger, N., Bantoto, B., Luo, C., Shams, I., Shahid, H., Chang, Y., Sun, G., Mbuagbaw, L., Samaan, Z., Levine, M. A. H., Adachi, J. D., Thabane, L. (2018). A systematic review of comparisons between protocols or registrations and full reports in primary biomedical research. *BMC Medical Research Methodology, 18*(1), 1-20. https://doi.org/10.1186/s12874-017-0465-7

Malich, L., & Munafò, M. R. (2022). Introduction: Replication of Crises-Interdisciplinary Reflections on the Phenomenon of the Replication Crisis in Psychology. *Review of General Psychology*, *26*(2), 127-130.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264-269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Munafò, M. R., Chambers, C. D., Collins, A. M., Fortunato, L., & Macleod, M. R. (2020). Research culture and reproducibility. *Trends in Cognitive Sciences*, *24*(2), 91-93.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., ... & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815-818.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Ofosu, G. K., & Posner, D. N. (2021). Pre-analysis plans: An early stocktaking. *Perspectives on Politics*, 1-17. https://doi.org/10.1017/S1537592721000931.

Olsson-Collentine, A., van Aert, R. C. M., Bakker, M., & Wicherts, J. M. (2023). Meta-Analyzing the Multiverse: A Peek Under the Hood of Selective Reporting. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/43yae

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

Pennington, C. R. (2023). *A student's guide to open science: Using the replication crisis to reform psychology*. Open University Press.

Pfeiffer, N., & Call, M. (2022). Surpassing 100,000 Registrations on OSF: Strides in Adoption of Open and Reproducible Research [Blog Post]. Retrieved from https://www.cos.io/blog/surpassing-100000-registrations-on-osf

Preregistration Task Force. (2021). Preregistration Standards for Psychology - the Psychological Research Preregistration-Quantitative (aka PRP-QUANT) Template. ZPID (Leibniz Institute for Psychology). http://dx.doi.org/10.23668/psycharchives.4584

Rice, D. B., & Moher, D. (2019). Curtailing the use of preregistration: A misused term term. *Perspectives on Psychological Science*, *14*(6), 1105-1108.

Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E. J. (2023). Comparing analysis blinding with preregistration in the many-analysts religion project. *Advances in Methods and Practices in Psychological Science*, *6*(1), 25152459221128319.

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC medicine*, *8*(1), 1-6.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*: 160384. http://doi.org/10.1098/rsos.160384

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), 220346.

TARG Meta-Research Group and Collaborators. (2022). Discrepancy review: A feasibility study of a novel peer review intervention to reduce undisclosed discrepancies between registrations and publications. *Royal Society Open Science*, *9*(7), 220142.

Thibault, R. T., Clark, R., Pedder, H., van den Akker, O. R., Westwood, S., Thompson, J., & Munafo, M. (2021). Estimating the prevalence of discrepancies between study

registrations and publications: A systematic review and meta-analyses. *medRxiv*. https://doi.org/10.1101/2021.07.07.21259868

Van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology, 67,* 2-12.

Van ’t Veer, A. E., Vazire, S., Campbell, L., Feldman, G., Etz, A., & Lindsay, D. S. (2019). Preregistration Planning and Deviation Documentation (PPDD). Retrieved from https://osf.io/ywrqe

Van den Akker, O. R., van Assen, M. A. L. M., Enting, M., de Jonge, M., Ong, H., Rüffer, F. F., … Bakker, M. (2023). Selective Hypothesis Reporting in Psychology: Comparing Preregistrations and Corresponding Publications [MetaArxiv Preprint]. https://doi.org/10.31222/osf.io/nf6mq

Van den Akker, O. R., Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., ... & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, 5.

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

Veldkamp, C. L., Hartgerink, C. H., Van Assen, M. A., & Wicherts, J. M. (2017). Who believes in the storybook image of the scientist? *Accountability in Research*, *24*(3), 127-151.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., Van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on psychological science*, *7*(6), 632-638.

Wicherts, J. M. (2017). The weak spots in contemporary science (and how to fix them). *Animals*, *7*(12), 90.

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*, 1832.

Yamada, Y. (2018). How to crack pre-registration: Toward transparent and open science. *Frontiers in Psychology*, *9*, 1831. https://doi.org/10.3389/fpsyg.2018.01831