

Psychological Methods

The Potential of Preregistration in Psychology: Assessing Preregistration Producibility and Preregistration-Study Consistency

Olmo R. van den Akker, Marjan Bakker, Marcel A. L. M. van Assen, Charlotte R. Pennington, Leone Verweij, Mahmoud M. Elsherif, Aline Claesen, Stefan D. M. Gaillard, Siu Kit Yeung, Jan-Luca Frankenberger, Kai Krautter, Jamie P. Cockcroft, Katharina S. Kreuer, Thomas Rhys Evans, Frédérique M. Heppel, Sarah F. Schoch, Max Korbmacher, Yuki Yamada, Nihan Albayrak-Aydemir, Shilaan Alzahawi, Alexandra Sarafoglou, Maksim M. Sitnikov, Filip Děchtěrenko, Sophia Wingen, Sandra Grinschgl, Helena Hartmann, Suzanne L. K. Stewart, Cátia M. F. de Oliveira, Sarah Ashcroft-Jones, Bradley J. Baker, and Jelte M. Wicherts

Online First Publication, October 10, 2024. <https://dx.doi.org/10.1037/met0000687>

CITATION

van den Akker, O. R., Bakker, M., van Assen, M. A. L. M., Pennington, C. R., Verweij, L., Elsherif, M. M., Claesen, A., Gaillard, S. D. M., Yeung, S. K., Frankenberger, J.-L., Krautter, K., Cockcroft, J. P., Kreuer, K. S., Evans, T. R., Heppel, F. M., Schoch, S. F., Korbmacher, M., Yamada, Y., Albayrak-Aydemir, N., Alzahawi, S., Sarafoglou, A., Sitnikov, M. M., Děchtěrenko, F., Wingen, S., Grinschgl, S., Hartmann, H., Stewart, S. L. K., de Oliveira, C. M. F., Ashcroft-Jones, S., Baker, B. J., & Wicherts, J. M. (2024). The potential of preregistration in psychology: Assessing preregistration producibility and preregistration-study consistency. *Psychological Methods*. Advance online publication. <https://dx.doi.org/10.1037/met0000687>

The Potential of Preregistration in Psychology: Assessing Preregistration Producibility and Preregistration-Study Consistency

Olmo R. van den Akker^{1, 2}, Marjan Bakker¹, Marcel A. L. M. van Assen^{1, 3}, Charlotte R. Pennington⁴, Leone Verweij^{1, 5}, Mahmoud M. Elsherif⁶, Aline Claesen⁷, Stefan D. M. Gaillard^{8, 9}, Siu Kit Yeung¹⁰, Jan-Luca Frankenberger¹¹, Kai Krautter¹², Jamie P. Cockcroft¹³, Katharina S. Kreuer⁸, Thomas Rhys Evans^{14, 15}, Frédérique M. Heppel¹⁶, Sarah F. Schoch¹⁷, Max Korbmacher¹⁸, Yuki Yamada¹⁹, Nihan Albayrak-Aydemir^{20, 21}, Shilaan Alzahawi²², Alexandra Sarafoglou¹⁶, Maksim M. Sitnikov²³, Filip Děchtěrenko²⁴, Sophia Wingen²⁵, Sandra Grinschgl²⁶, Helena Hartmann^{27, 28}, Suzanne L. K. Stewart²⁹, Cátia M. F. de Oliveira¹³, Sarah Ashcroft-Jones³⁰, Bradley J. Baker³¹, and Jelte M. Wicherts¹

¹ Department of Methodology and Statistics, Tilburg University

² QUEST Center for Responsible Research, Berlin Institute of Health at Charité, Berlin, Germany

³ Department of Sociology, Utrecht University

⁴ School of Psychology, Aston University

⁵ Department of Public Administration and Sociology, Erasmus University

⁶ Department of Vision Sciences, Psychology and Behaviour, University of Leicester

⁷ Faculty of Psychology and Educational Sciences, KU Leuven

⁸ Institute for Science in Society, Radboud University

⁹ Center of Trial and Error, Utrecht, The Netherlands

¹⁰ Department of Psychology, Chinese University of Hong Kong

¹¹ Behavioral Science, Radboud University

¹² Organizational Behavior Unit, Harvard Business School

¹³ Department of Psychology, University of York

¹⁴ School for Human Sciences, University of Greenwich

¹⁵ Institute for Lifecourse Development, University of Greenwich

¹⁶ Department of Psychological Methods, University of Amsterdam

¹⁷ Donders Institute, Radboud University Medical Center

¹⁸ Department of Health and Functioning, Western Norway University of Applied Sciences

¹⁹ Faculty of Arts and Science, Kyushu University

²⁰ Open Psychology Research Centre, Open University

²¹ Department of Psychological and Behavioural Science, London School of Economics and Political Science

²² Graduate School of Business, Stanford University

²³ Department of Organization Studies, Tilburg University

²⁴ Department of Cognitive Psychology, Institute of Psychology, Czech Academy of Sciences, Prague, Czechia

²⁵ Faculty of Management, Economics and Social Sciences, University of Cologne

²⁶ Institute for Psychology, University Graz

²⁷ Clinical Neurosciences, Department for Neurology, University Hospital Essen


²⁸ Center for Translational Neuro- and Behavioral Sciences (C-TNBS), University Hospital Essen

²⁹ School of Psychology, University of Chester

³⁰ Department of Experimental Psychology, University of Oxford

³¹ Department of Sport and Recreation Management, Temple University

Sacha Epskamp served as action editor.

Olmo R. van den Akker  <https://orcid.org/0000-0002-0712-3746>

This research was funded in whole, or in part, by the European Commission (Grant 726361). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

This work was supported by a Consolidator Grant (Innovative Methods for Psychology: Reproducible, Open, Valid, and Efficient, Grant 726361) from the European Research Council. The authors would like to thank Sarahanne Field, Gustav Nilsson, Max Primbs, Robert Thibault, Simon Schwab, and

Samuel Westwood for their contributions earlier in the project. The authors declare that they have no conflicts of interest. Preliminary data related to this study were presented online on September 18, 2021, as part of the Metascience Symposium 2021. The final results related to this study were presented at the Perspectives on Scientific Error Event on February 29, 2024, and The Royal Society's "The Promises and Pitfalls of Preregistration" Meeting on March 4, 2024.

This work is licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International License (CC BY-NC-ND 4.0; <https://creativecommons.org/licenses/by-nc-nd/4.0>). This license permits copying and redistributing the work in any medium or format for

Abstract

Study preregistration has become increasingly popular in psychology, but its potential to restrict researcher degrees of freedom has not yet been empirically verified. 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 preregistrations 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. These results highlight that biases due to researcher degrees of freedom remain possible in many preregistered studies. More comprehensive registration templates typically yielded more producible preregistrations. We did not find that the producibility and consistency of preregistrations differed over time or between original and replication studies. Furthermore, we found that operationalizations of variables were generally preregistered more producible and consistently 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.

Translational Abstract

This article is about study preregistration, the process of publicly sharing the design, the hypotheses, and the analysis plans of a scientific study before conducting that study. This is said to be important because of two reasons. First, researchers are prone to confirmation bias, the phenomenon that people interpret new information as confirmation of what they already believe. Second, researchers make a career out of publishing in scientific journals and the chance of that is higher if you have found something interesting in your study. These two reasons lead to a situation where researchers are eager to find evidence in favor of a theory that they think is true or interesting. In turn, this could lead them to make decisions during the research process, either consciously or subconsciously, that will lead them to the result that they want. This is problematic because these decisions might not be the optimal decisions to get to the truth. If researchers already preregister their study plan before doing the study, this would limit their freedom to make these suboptimal decisions. In this article, we assess whether preregistration actually helps in preventing such decisions. We found that researchers are not yet sufficiently skilled at preregistering their studies and that they often provide too little information in their preregistration. We also found that they sometimes conduct their study in a different way than they planned in their preregistration. A solution to improve preregistration skills could be specific training and education, as well as more guidance in the form of preregistration templates that prompt researchers what they should include in their preregistrations.

Keywords: preregistration, preregistration producibility, preregistration-study consistency, preregistration deviations, metaresearch

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 et al., 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 et al., 2017). One reason for this is that researchers are susceptible to cognitive biases like confirmation bias and motivated reasoning bias (Bishop, 2020; Munafò et al., 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 &

noncommercial use provided the original authors and source are credited and a link to the license is included in attribution. No derivative works are permitted under this license.

Olmo R. van den Akker served as lead for formal analysis, project administration, visualization, and writing—original draft. Marjan Bakker served as lead for writing—review and editing and served in a supporting role for formal analysis. Marcel A. L. M. van Assen served as lead for writing—review and editing. Charlotte R. Pennington, Leone Verweij, Mahmoud M. Elsherif, Aline Claesen, Stefan D. M. Gaillard, Siu Kit Yeung, Jan-Luca Frankenberger, Kai Krautter, Jamie P. Cockcroft, Katharina S. Kreuer, Thomas Rhys Evans, Frédérique M. Heppel, Sarah F. Schoch, Max Korbmacher, Yuki Yamada, Nihan Albayrak-Aydemir, Shilaa Alzahawi, Alexandra Sarafoglou, Maksim M. Sitnikov, Filip Děchtěrenko, Sophia Wingen, Sandra Grinschgl, Helena Hartmann, Suzanne L. K. Stewart, Cátia M. F. de Oliveira, Sarah Ashcroft-Jones, and Bradley J. Baker served in a supporting role for writing—review and editing. Jelte M. Wicherts served as lead for funding acquisition and writing—

review and editing. Olmo R. van den Akker, Marjan Bakker, Marcel A. L. M. van Assen, and Jelte M. Wicherts contributed equally to conceptualization and methodology. Marjan Bakker, Marcel A. L. M. van Assen, and Jelte M. Wicherts contributed equally to supervision. Olmo R. van den Akker, Charlotte R. Pennington, Leone Verweij, Mahmoud M. Elsherif, Aline Claesen, Stefan D. M. Gaillard, Siu Kit Yeung, Jan-Luca Frankenberger, Kai Krautter, Jamie P. Cockcroft, Katharina S. Kreuer, Thomas Rhys Evans, Frédérique M. Heppel, Sarah F. Schoch, Max Korbmacher, Yuki Yamada, Nihan Albayrak-Aydemir, Shilaa Alzahawi, Alexandra Sarafoglou, Maksim M. Sitnikov, Filip Děchtěrenko, Sophia Wingen, Sandra Grinschgl, Helena Hartmann, Suzanne L. K. Stewart, Cátia M. F. de Oliveira, Sarah Ashcroft-Jones, and Bradley J. Baker contributed equally to investigation.

Correspondence concerning this article should be addressed to Olmo R. van den Akker, Department of Methodology and Statistics, Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands. Email: ovdakker@gmail.com

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 et al., 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 et al., 2018; Wagenmakers et al., 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: (a) The producibility of the preregistration (i.e., whether the information provided in the preregistration is comprehensive enough to properly conduct the study),¹ and (b) 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). As such, both the producibility criterion and the consistency criterion must be met for a preregistration to successfully combat questionable research practices.

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, 2023; 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 many decisions in analyzing data could still create sizeable variation in outcomes that researchers could selectively report (Olsson-Collentine et al., 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 et al., 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 (2023) 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, van Assen, 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 producibility and consistency, we also assess the following specific study parts: 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 producibility and consistency, like replication status, time, and the comprehensiveness of the preregistration template. We preregistered our hypotheses at <https://osf.io/83ahg>.

Our first hypothesis was that replication studies would be preregistered more producible and consistently 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. 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 producible and consistent 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

¹ We called this aspect "strictness" in our preregistration.

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, Peters, et al., 2023; 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 producibility and consistency 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 better 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, van Assen, 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 method sections. A hypothesis was labeled as a “match” if the hypothesis in the article 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 producibility and consistency 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 producibility and consistency 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 Preferred Reporting Items for Systematic reviews and Meta-Analyses flow diagram (Moher et al., 2009) in Figure 1. The protocols used by Van den Akker, van Assen, 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/uuyrds>, respectively.

Measuring Preregistration Producibility and Consistency

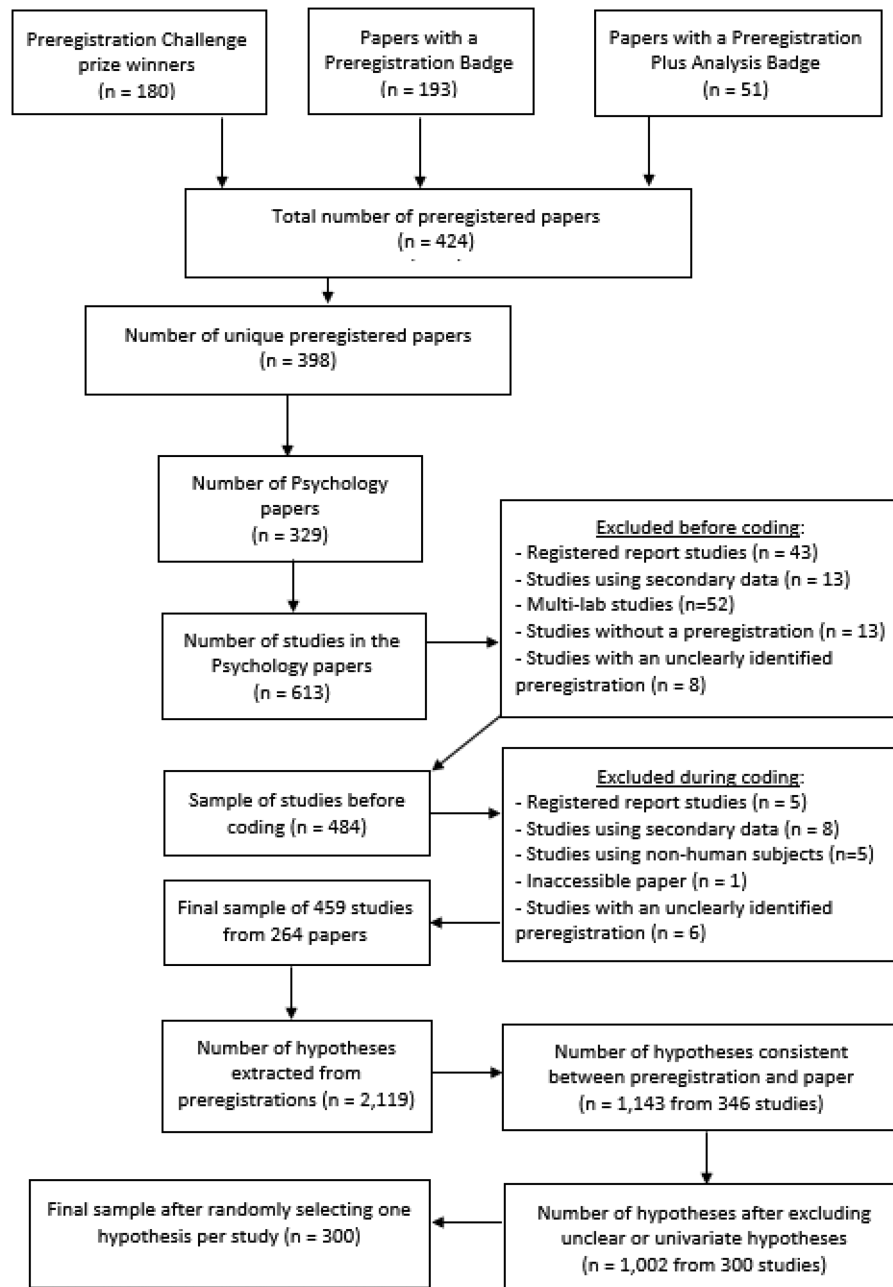
We coded preregistration producibility and consistency using a protocol (adapted from Bakker et al., 2020) administered via Qualtrics that extracts information from the preregistration and the paper and then helps assess preregistration producibility as well as preregistration-study consistency. The 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 min, 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 metascientific perspective. They were trained using a set of 10 example PSPs and coded an 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

Figure 1
PRISMA Flow Diagram Outlining the Full Sample Selection Procedure



Note. PRISMA = Preferred Reporting Items for Systematic reviews and Meta-Analyses.

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, and
 - (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,

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 six major study parts in both Tables 2 and 3.

- (b) the potential values, and
- (c) how the variable was constructed from its components, if applicable.
- 3. The data collection procedure:
 - (a) sample size, and
 - (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), and
 - (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. We selected these study elements because the coding in our pilot study indicated that they were relatively common across study designs. While we considered other study elements like the estimation technique in a statistical model or corrections for multiple testing, these were not always applicable across studies. We felt that the chosen elements were a good balance between informational value, applicability, and feasibility of coding.

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*). We took the mean of the partial producibility scores instead of the sum as we preregistered, because the resulting score range of (0, 2) is in line with the scores of the separate study parts and is more interpretable than a sum score range of (0, 10).

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 article, 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 reproduced. 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 because there simply was not enough information available to make a preregistration-study comparison.

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*). After taking the mean of all scores on the five major parts of the study, the preregistration could score between 0 (*not consistent at all*) and 1 (*optimally consistent*). We took the mean of the partial consistency scores instead of the sum as we preregistered, because the resulting score range of (0, 1) is in line with the scores of the separate study parts and is more interpretable than a sum score range of (0, 5).

Combining Producibility and Consistency. Aside from preregistration producibility and consistency, we also computed an overall effectiveness score for each preregistration, as preregistered. To that end, 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 because it is crucial that a study part is both producible and consistent to prevent questionable research practices in that part. 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).

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, and
 - (c) how the variable was constructed from its components, if applicable.
2. How missing data was handled:
 - (a) the definition of missing data, and
 - (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, and
 - (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 and preregistration-study consistency 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, van Assen, 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 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 point was given. Scoring was done by two independent coders (Olmo R. van den Akker and Charlotte R. Pennington) who together resolved three initial coding discrepancies. For one discrepancy, an independent third coder (Marjan Bakker) 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 Preregistration in Social Psychology template scored substantially less well, with 10 and 14 out of 27 points, respectively.

Determining Registration Dates

To assess whether preregistration producibility and consistency 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 nonfrozen 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 April 14, 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

Overview of Deviations From Our Preregistration

1. We changed our original sample by excluding studies that involved (a) only hypotheses for which we could not clearly determine the type of hypothesis or (b) only hypotheses with

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 article 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).

Table 1

Frequencies and Comprehensive Scores (Separated for Included Items, Explanations, and Examples) of the Preregistration Templates Used to Draft the Preregistrations in Our Sample

Template	Frequency	Comprehensiveness scores items	Comprehensiveness scores explanations	Comprehensiveness score examples	Total comprehensiveness scores
OSF Prereg template (Bowman et al., 2020)	122	8	8	8	24
AsPredicted (https://aspredicted.org)	112	5	0	5	10
Preregistration in Social Psychology (van 't Veer & Giner-Sorolla, 2016)	21	8	6	0	14
OSF's Open Templates (https://osf.io/9j6d7 ; https://osf.io/haadc)	7				
Happy Lab Pre-Registration Template (https://osf.io/yvvsj8)	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. OSF = Open Science Framework.

- one variable. See the section Selection of Preregistered Studies section for more details.
- We used manipulation status to assess directional hypotheses instead of directionality as coded in Van den Akker, van Assen, et al. (2023). Because of this, we now distinguish between six major study parts in Tables 2 and 3 instead of five. See Footnote 2 for more information.
 - We took the mean instead of the sum of the partial producibility and consistency scores to get the overall producibility and consistency scores. See the section Measuring Preregistration Producibility and Consistency section for more information.
 - We only regarded the definition of the exclusion criteria and not the procedure of exclusion. See Footnote 3 for more information.

- Because the inference criteria were almost never explicitly stated in the article, we used implicit consistency instead to assess the consistency of inference criteria. See the paragraph following Figure 2 for more details.
- We did not include the overall effectiveness scores in Table 2 and Figure 2, as reviewers preferred us to focus on producibility and consistency scores. However, the table and figure including effectiveness scores can be found at <https://osf.io/ky8u5>.

Descriptive Statistics

Of the 300 PSPs in our sample, we classified 138 (46%) as replication studies, and the remaining 162 (54%) as original studies.

Table 2

Overview of Producibility and Consistency Scores (With Standard Deviations) for the Five Major and Five Minor Study Parts, as Well as Mean Scores for the Major Study Parts

Study parts	Producibility				Consistency			
	0	1	2	Score (SD)	0	1	NA	Score (SD)
Major study parts								
Measured variable ($N = 164$) ^a	28 (17%)	61 (37%)	75 (46%)	1.29 (0.74)	45 (27%)	91 (55%)	28 (17%)	0.67 (0.47)
Manipulated variable ($N = 218$) ^a	49 (22%)	NA	169 (78%)	1.55 (0.84)	8 (4%)	146 (67%)	64 (29%)	0.95 (0.22)
Dependent variable ($N = 218$) ^a	24 (11%)	79 (36%)	115 (53%)	1.42 (0.68)	56 (26%)	131 (60%)	31 (14%)	0.70 (0.46)
Data collection procedure ($N = 300$)	38 (13%)	NA	262 (87%)	1.75 (0.67)	176 (59%)	84 (28%)	40 (13%)	0.32 (0.47)
Statistical model ($N = 300$)	23 (8%)	159 (53%)	118 (39%)	1.32 (0.61)	121 (40%)	135 (45%)	44 (15%)	0.53 (0.50)
Inference criteria ($N = 300$)	191 (64%)	NA	109 (36%)	0.73 (0.96)	10 (3%)	94 (31%)	196 (65%)	0.90 (0.30)
Mean scores				1.33 (0.41)				0.71 (0.20)
Minor study parts								
Measured control variable ($N = 20$) ^b	7 (35%)	5 (25%)	8 (40%)	1.05 (0.89)	9 (45%)	4 (20%)	7 (35%)	0.31 (0.48)
Manipulated control variable ($N = 23$) ^b	5 (22%)	NA	18 (78%)	1.57 (0.84)	0 (0%)	18 (78%)	5 (22%)	1.00 (0.00)
Exclusion criteria ($N = 300$)	68 (23%)	NA	232 (77%)	1.55 (0.84)	86 (29%)	106 (35%)	108 (36%)	0.55 (0.50)
Missing data ($N = 300$)	159 (53%)	21 (7%)	120 (40%)	0.87 (0.96)	9 (3%)	37 (12%)	254 (85%)	0.80 (0.40)
Statistical assumptions ($N = 300$)	279 (93%)	17 (6%)	4 (1%)	0.08 (0.32)	2 (1%)	6 (2%)	292 (97%)	0.75 (0.46)

Note. NA = not available.

^a Highlighting 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. ^b Highlighting that we only had 46 hypotheses with one or more control variables. Twenty of those were part of a nondirectional hypothesis, and 26 were part of a directional hypothesis.

Table 3

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

Study elements	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%)

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/vwagak>). 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 and consistency for the separate study parts and 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; 0, 1, and not applicable 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).⁴ The correlation of producibility and consistency scores was $r = -.11$, $t(55) = -0.82$, $p = .418$. A visualization of the scores for producibility and consistency can be found in Figure 2.

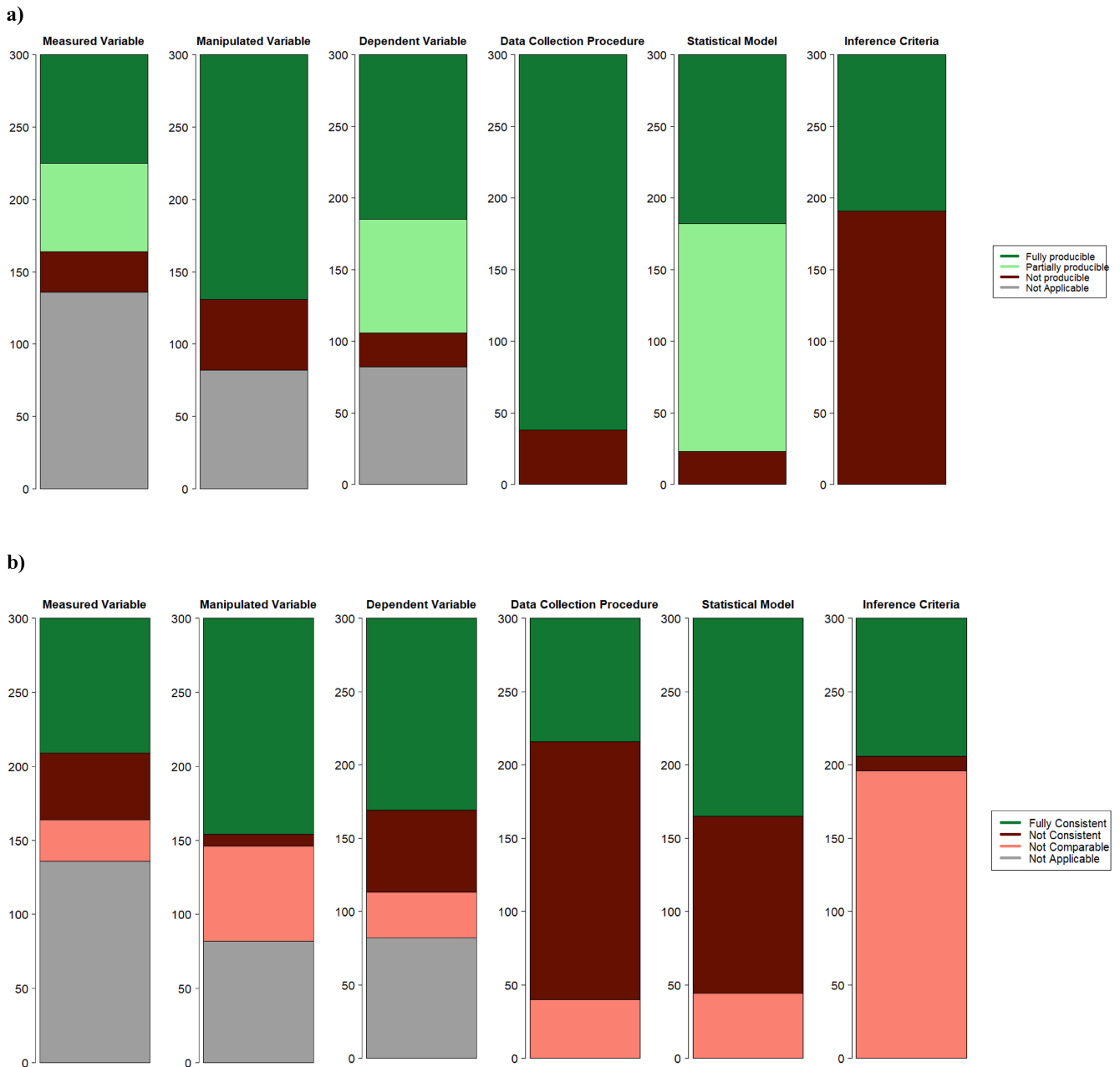
Note that the consistency scores in Table 2 indicate the proportion of PSPs for which that study part was consistent out of all PSPs for which the study parts could be compared between preregistration and paper. For example, the statistical model could be compared 256 times, of which 121 (47%) were consistent. Because the inference criteria were almost never explicitly stated in the article, 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.

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

⁴ We also assessed the preregistration producibility and consistency for the current study and arrived at a score of 2.0 for preregistration producibility, and a score of 0.8 for preregistration-study consistency (because our sample size was not consistent). For the nonessential 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 is 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>.

Figure 2
Preregistration Producibility Scores (a) and Preregistration-Study Consistency Scores (b)



Note. See the online article for the color version of this figure.

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>.

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

Table 4
Results of the Tests of Hypothesis 1, Hypothesis 2, and Hypothesis 3, With Confidence Intervals Within Square Brackets

Parameter	M1a	M1b	M1c	M2a1	M2b1	M2c1	M2a2	M2b2	M2c2	M3a	M3b	M3c
Fixed effects (standard errors)												
Intercept	1.34 [1.25, 1.43]	0.71 [0.62, 0.80]	0.80 [0.70, 0.90]	1.16 [1.04, 1.28]	0.66 [0.35, 0.98]	0.67 [0.52, 0.82]	1.30 [1.09, 1.52]		0.65 [0.36, 0.95]	1.32 [1.16, 1.48]	0.50 [0.21, 0.79]	0.85 [0.67, 1.03]
Level 1												
Replication	-0.001 [-0.10, 0.10]	-0.001 [-0.13, 0.14]	0.03 [-0.10, 0.15]	0.01 [-0.09, 0.10]	0.01 [-0.14, 0.15]	0.05 [-0.09, 0.19]	0.01 [-0.11, 0.12]		0.01 [-0.19, 0.20]	-0.002 [-0.08, 0.08]	-0.02 [-0.12, 0.09]	0.03 [-0.07, 0.13]
OSF vs. AP			0.44 [0.30, 0.58]		0.04 [-0.28, 0.36]	0.30 [0.13, 0.47]						
OSF vs. SP							0.30 [0.07, 0.54]		0.34 [0.03, 0.66]			
Months										0.001 [-0.003, 0.005]	0.004 [-0.002, 0.01]	-0.001 [-0.01, 0.003]
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 (M1a) refers to the model testing the first part of Hypothesis 1 (producibility), while Models 1b and 1c test the second (consistency) and third (effectiveness) parts, respectively. The same holds for the Models M2 and M3, which test Hypotheses 2 and 3. Bold indicates statistically significant. OSF = Open Science Framework; AP = AsPredicted template; SP = Preregistration in social psychology template.

were preregistered more productively ($M = 1.31$) than original studies ($M = 1.35$), $B_1 = -0.001$, $t(234.7) = -0.01$, 99% confidence interval (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 Open Science Framework (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, and 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: $B_1 = 0.001$, $t(296.8) = 0.38$, 95% CI = [-0.003, 0.005], $p = .704$; preregistration-study consistency: $B_1 = 0.04$, $t(50.2) = 1.46$, 95% CI = [-0.002, 0.01], $p = .151$; and effectiveness: $B_1 = -0.001$, $t(266.0) = -0.61$, 95% CI = [-0.01, 0.003], $p = .545$). 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 and consistent. 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 and more consistent 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 and consistency scores to whether the assessed hypothesis (see the Selection of Preregistered Studies section 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, van Assen, 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 nonsignificant result). Furthermore, we found no evidence of an association of preregistration producibility with statistical significance—producibility: $B_1 = -0.14$, $t(200.4) = -1.71$, 99% CI = $[-0.29, 0.02]$, $p = .088$.

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. These scores indicate that over the years 2014–2020, preregistrations were not producible and consistent enough to weed out all questionable research practices. Our findings are in line with earlier studies that assessed preregistration in economics and political science (Ofosu & Posner, 2023), 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 preregistered well tend to be tied to the operationalization of variables (however, see Sarafoglou et al., 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 metascientific 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 preregistered studies, and the same issue likely holds for papers with nonpreregistered 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 producible and consistent preregistrations than original studies.

In line with Hypothesis 2, preregistrations based on more comprehensive templates were generally more producible 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 conscientiously 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 producible and consistent 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 skilled 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 data set. 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 preregistration producibility and consistency 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 producibility and consistency and statistical significance did not provide sufficient evidence for the claim that more producible and consistent preregistrations better prevent *p*-hacking. One possible explanation for the absence of an association is that we investigated not only primary hypotheses but also hypotheses that were indicated by one of seven keywords (see the Selection of Preregistered Studies section). 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% CI = $[-0.26, 0.02]$), which raises the question of whether the added time and effort associated with a producible and consistent preregistration (Sarafoglou et al., 2023) is worth it.

Importantly, there could also be other reasons for why more producible 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 producible 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 & Dutilh, 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 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. 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 et al., 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. To correct for this, we did an exploratory analysis where we recalculated the producibility score for the statistical model, which became 1.70 (was 1.32). This updated score better aligns with Ofosu and Posner (2023), 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 partial producibility (score of 1) and full producibility (score of 2) was arbitrary. 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 productively 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 productively 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/3mxf5>.

Our results also mimic those of previous studies with regard to explanations for deviations. Like Claesen et al. (2021) 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). We also want to highlight that reviewer comments can be a reason for why papers differ from their corresponding preregistrations, which was the case when a reviewer advised us to exclude preregistration effectiveness scores from Table 2. The prevalence and constitution of preregistration deviations as a response to reviewer comments are currently unknown and warrant further study.

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 metaresearchers.

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 metaresearch on the topic of preregistration (often called registration in this discipline; Rice & 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 article but would be an interesting metaresearch 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 et al. (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 preregistrators, 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., Cromptvoets, 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), Article 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. <https://doi.org/10.1038/s41562-021-01193-7>

- 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), Article 211037. <https://doi.org/10.1098/rsos.211037>
- Dwan, K., Gamble, C., Williamson, P. R., Kirkham, J. J., & Group, R. B. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PLoS ONE*, 8(7), Article e66844. <https://doi.org/10.1371/journal.pone.0066844>
- Food and Drug Administration Modernization Act of 1997. (1997). *Public Law 105-15*. <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, Article e1935. <https://doi.org/10.7717/peerj.1935>
- Haven, T. L., & Van Grootel, D. L. (2019). Preregistering qualitative research. *Accountability in Research*, 26(3), 229–244. <https://doi.org/10.1080/08989621.2019.1580147>
- 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. <https://doi.org/10.31234/osf.io/nj4es>
- Jones, C. W., Keil, L. G., Holland, W. C., Caughey, M. C., & Platts-Mills, T. F. (2015). Comparison of registered and published outcomes in randomized controlled trials: A systematic review. *BMC Medicine*, 13(1), 1–12. <https://doi.org/10.1186/s12916-015-0520-3>
- Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, 62(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221
- 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., ... 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), Article 9. <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. <https://doi.org/10.1177/10892680221077997>
- 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. <https://doi.org/10.1016/j.tics.2019.12.002>
- Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, 23(10), 815–818. <https://doi.org/10.1016/j.tics.2019.07.009>
- 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. <https://doi.org/10.1073/pnas.1708274114>
- 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>
- Ofose, G. K., & Posner, D. N. (2023). Pre-analysis plans: An early stocktaking. *Perspectives on Politics*, 21(1), 174–190. <https://doi.org/10.1017/S1537592721000931>
- 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*. 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 Behaviour*, 6(3), 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]. <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). <https://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. <https://doi.org/10.1177/1745691619858427>
- Sarafoglou, A., Hoogveen, 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), Article 25152459221128319. <https://doi.org/10.1177/25152459221128319>
- 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), Article 24. <https://doi.org/10.1186/1741-7015-8-24>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9), Article 160384. <https://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), Article 220346. <https://doi.org/10.1098/rsos.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), Article 220142. <https://doi.org/10.1098/rsos.220142>
- Thibault, R. T., Clark, R., Pedder, H., van den Akker, O. R., Westwood, S., Thompson, J., & Munafò, M. (2021). *Estimating the prevalence of discrepancies between study registrations and publications: A systematic review and meta-analysis*. medRxiv. <https://doi.org/10.1101/2021.07.07.21259868>
- van den Akker, O. R., Peters, G. J. Y., Bakker, C. J., Carlsson, R., Coles, N. A., Corker, K. S., Feldman, G., Moreau, D., Nordström, T., Pickering, J. S., Riegelman, A., Topor, M. K., van Veggel, N., Yeung, S. K., Call, M., Mellor, D. T., & Pfeiffer, N. (2023). Increasing the transparency of systematic reviews: presenting a generalized registration form. *Systematic Reviews*, 12(1), 170. <https://doi.org/10.1186/s13643-023-02281-7>
- Van den Akker, O. R., van Assen, M. A. L. M., Enting, M., de Jonge, M., Ong, H., Rüffer, F. F., Schoenmakers, M., Stoevenbelt, A. H., Wicherts, J., Bakker, M. (2023). *Selective hypothesis reporting in psychology: Comparing preregistrations and corresponding publications*. MetaArxiv. <https://doi.org/10.31222/osf.io/nf6mq>
- Van den Akker, O. R., Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., Hall, A., Kosie, J., Kruse, E., Olsen, J., Ritchie, S., Valentine, K. D., Van 't Veer, A., & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, 5(4). <https://doi.org/10.15626/MP.2020.2625>
- 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. <https://doi.org/10.1016/j.jesp.2016.03.004>

- van 't Veer, A. E., Vazire, S., Campbell, L., Feldman, G., Etz, A., & Lindsay, D. S. (2019). *Preregistration planning and deviation documentation (PPDD)*. <https://osf.io/ywrqe>
- 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. <https://doi.org/10.1080/08989621.2016.1268922>
- Wagenmakers, E. J., & Dutilh, G. (2016). *Seven selfish reasons for preregistration*. APS Observer.
- 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. <https://doi.org/10.1177/1745691612463078>
- Wicherts, J. M. (2017). The weak spots in contemporary science (and how to fix them). *Animals*, *7*(12), Article 90. <https://doi.org/10.3390/ani7120090>
- Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, Article 1832. <https://doi.org/10.3389/fpsyg.2016.01832>
- Yamada, Y. (2018). How to crack pre-registration: Toward transparent and open science. *Frontiers in Psychology*, *9*, Article 1831. <https://doi.org/10.3389/fpsyg.2018.01831>

Received May 10, 2023

Revision received May 9, 2024

Accepted May 16, 2024 ■