



Selective Hypothesis Reporting in Psychology: Comparing Preregistrations and Corresponding Publications



Olmo R. van den Akker¹, Marcel A. L. M. van Assen^{1,2},
 Manon Enting¹, Myrthe de Jonge¹, How Hwee Ong³,
 Franziska Rüffer¹, Martijn Schoenmakers¹, Andrea H. Stoevenbelt^{1,4},
 Jelte M. Wicherts¹, and Marjan Bakker¹

¹Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands;

²Department of Sociology, Utrecht University, Utrecht, The Netherlands; ³Department of Social Psychology, Tilburg University, Tilburg, The Netherlands; and ⁴Department of Educational Science, University of Groningen, Groningen, The Netherlands

Advances in Methods and
 Practices in Psychological Science
 July-September 2023, Vol. 6, No. 3,
 pp. 1–15
 © The Author(s) 2023
 Article reuse guidelines:
 sagepub.com/journals-permissions
 DOI: 10.1177/25152459231187988
 www.psychologicalscience.org/AMPPS



Abstract

In this study, we assessed the extent of selective hypothesis reporting in psychological research by comparing the hypotheses found in a set of 459 preregistrations with the hypotheses found in the corresponding articles. We found that more than half of the preregistered studies we assessed contained omitted hypotheses ($N = 224$; 52%) or added hypotheses ($N = 227$; 57%), and about one-fifth of studies contained hypotheses with a direction change ($N = 79$; 18%). We found only a small number of studies with hypotheses that were demoted from primary to secondary importance ($N = 2$; 1%) and no studies with hypotheses that were promoted from secondary to primary importance. In all, 60% of studies included at least one hypothesis in one or more of these categories, indicating a substantial bias in presenting and selecting hypotheses by researchers and/or reviewers/editors. Contrary to our expectations, we did not find sufficient evidence that added hypotheses and changed hypotheses were more likely to be statistically significant than nonselectively reported hypotheses. For the other types of selective hypothesis reporting, we likely did not have sufficient statistical power to test for a relationship with statistical significance. Finally, we found that replication studies were less likely to include selectively reported hypotheses than original studies. In all, selective hypothesis reporting is problematically common in psychological research. We urge researchers, reviewers, and editors to ensure that hypotheses outlined in preregistrations are clearly formulated and accurately presented in the corresponding articles.

Keywords

hypotheses, bias, selective reporting, statistical significance, preregistration

Received 10/18/22; Revision accepted 6/15/23

Scientists should be open-minded and consider all new evidence, hypotheses, theories, and innovations when doing research, even those that challenge or contradict their own interests and beliefs (Anderson, 2000; Merton, 1942/1973). However, scientists do not always abide by this Mertonian norm. Studies have shown that scientists regularly add, drop, or alter study elements when preparing reports for publication (Dwan et al., 2013, 2014; Mazzola & Deuling, 2013; O'Boyle et al., 2017), a practice known as “selective reporting” (Cairo et al., 2020).

For example, researchers may fail to report study results that are not statistically significant and thus “not interesting” for publication (Chan et al., 2004), or they may alter hypotheses after seeing the data to make their article's narrative cleaner and more convincing (Giner-Sorolla, 2012; Kerr, 1998).

Corresponding Author:

Olmo van den Akker, Tilburg University, Warandelaan 2, 5037 AB
 Tilburg, The Netherlands.
 Email: ovdakker@gmail.com



Creative Commons NonCommercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits noncommercial use, reproduction, and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Selective reporting seems to be driven at least partly by a desire to publish work in prestigious, selective journals (Van der Steen et al., 2018) and biases the scientific literature toward articles with publishable (often statistically significant) results. Indeed, statistically significant results are so abundant in the scientific literature that it is unlikely that the literature represents all research that has been conducted (Scheel et al., 2021; Sterling, 1959; Sterling et al., 1995).

Selective-reporting practices have been identified in many scientific fields, but studies on this issue have been especially prevalent in biomedicine (see DeVito et al., 2020; Thibault et al., 2021; Vinkers et al., 2021). The reason for this is that clinical trials in this field are generally required to be registered in a formal and publicly accessible registry (DeAngelis et al., 2005; European Commission, 2012; Food and Drug Administration Amendments Act of 2007, 2018). This requirement enables comparing the registered protocol and the actual scientific publication to assess whether the authors of the publication changed, omitted, or added results, outcomes, or hypotheses. A systematic review of dozens of such metastudies by Thibault et al. (2021) found that between 10% and 68% (95% prediction interval) of articles contain at least one primary-outcome discrepancy.

The social sciences do not have an extensive registration infrastructure, so selective reporting has mainly been studied by comparing publications with dissertations (Cairo et al., 2020; Mazzola & Deuling, 2013; O'Boyle et al., 2017) and archived research proposals (Franco et al., 2016). Only a handful of studies compared publications with their corresponding preregistration, and all of them found that these publications often contained undisclosed deviations (psychology: Claesen et al., 2021; gambling: Heirene et al., 2021; economics and political science: Ofosu & Posner, 2023). In our study, we make use of the increased popularity of preregistration in psychological research in recent years (Hardwicke et al., 2022; Nosek & Lindsay, 2018) and check a large sample of preregistered psychology publications to assess the prevalence of one form of selective reporting: the selective reporting of hypotheses.

Selective hypothesis reporting can take on different types. We derived the terminology for these types from the biomedical literature, more specifically from Chan et al. (2004) and Thibault et al. (2021). One major difference between our study and earlier biomedical studies, however, is that we focus on hypotheses, whereas biomedical studies typically focus on outcomes (i.e., dependent variables). This may be because outcomes take a prominent place in the clinicaltrials.gov registration template used for many clinical trials. In the current study, we distinguish five types of selective hypothesis reporting.

First, the number of hypotheses can change from the preregistration to the publication, which includes

hypotheses that were present in the preregistration but did not appear in the publication (“omitted hypotheses”) and hypotheses that were not present in the preregistration but did appear in the publication (“added hypotheses”). Second, the status of hypotheses can change between the preregistration and the publication, which includes hypotheses that were labeled as “primary” in the publication but as “secondary” in the preregistration (“promoted hypotheses”) and hypotheses that were labeled as “secondary” in the publication but as “primary” in the preregistration (“demoted hypotheses”). Third, the direction of hypotheses (i.e., a positive, negative, null, or nondirectional effect; or $A > B$, $A < B$, $A = B$, or $A \neq B$ when comparing groups) can change between the preregistration and the publication (“changed hypotheses”). Note that hypotheses can also differ in other ways between preregistration and article. For example, sometimes authors alter the names of certain variables in the article compared with the preregistration, or sometimes authors change the hypothesis from passive to active tense or vice versa. We do not consider such changes in this study because they change a hypothesis only superficially rather than structurally. We thus use the adjective “changed” only for hypotheses with a direction change.

Note that the presence of statistical results related to added hypotheses in a publication is fine as long as they are labeled as “exploratory” (Logg & Dorison, 2021; Nosek et al., 2018). This is exemplified by the fact that both the CONSORT 2010 reporting guideline (Schulz et al., 2010) and the Journal Article Reporting Standards (JARS) reporting guideline (Appelbaum et al., 2018) explicitly encourage the reporting of exploratory analyses. Readers will then know that the hypotheses were drawn up a posteriori and that using hypothesis tests to make statistical inferences may be invalid (Wagenmakers et al., 2012). However, if the results of added hypotheses are labeled as “confirmatory” or not labeled at all, readers are unaware of the exploratory nature of the hypotheses and may inappropriately interpret the results using a hypothesis-testing framework. In these instances, undisclosed and statistically uncontrolled explorations could unjustly be perceived as solid confirmatory evidence. In this study, we therefore use the term “added hypotheses” only for nonpreregistered hypotheses with statistical results that are labeled as “confirmatory” or not labeled at all.

We investigate the different forms of selective hypothesis reporting in psychological research by identifying hypotheses in our sample of preregistrations and the accompanying publications. We distinguish between hypotheses that are part of direct replications and hypotheses that are part of original studies because we believe selective hypothesis reporting to be less of an issue for the former than for the latter (Hypothesis 1).

We also assess whether forms of selective hypothesis reporting are related to statistically significant results (Hypotheses 2a–2d). Our specific hypotheses and our rationale for these hypotheses are outlined below.

Hypotheses

We had no hypotheses about the exact proportion of studies involving selective hypothesis reporting, but we did expect that forms of selective hypothesis reporting would be less common among direct-replication hypotheses than among original hypotheses because direct-replication hypotheses need to adhere (both in the preregistration and the publication) to the hypotheses outlined in the original study. We also expected some forms of selective hypothesis reporting to be associated with statistical significance because results that are statistically significant are more likely to be published than results that are not statistically significant (Kerr, 1998; Scheel et al., 2021). Our hypotheses are listed more formally below and can also be found in our preregistration at <https://osf.io/z4awv>. Note that we originally uploaded our preregistration on OSF on January 21, 2021, before data collection. However, we formally entered it into the registry on March 5, 2023, to increase the findability of our preregistration (see <https://osf.io/nxgtv>). Aside from correcting the erroneous statement listed in Footnote 3, we did not make any changes.

Hypothesis 1: A hypothesis that is part of a direct replication is less likely to be selectively reported (omitted, promoted, demoted, or changed) than an original hypothesis.

Hypothesis 2a: The test result of an added hypothesis is more likely to be statistically significant than the test result of a preregistered hypothesis that is appropriately reported.

Hypothesis 2b: The test result of a promoted hypothesis is more likely to be statistically significant than the test result of a preregistered hypothesis that is appropriately reported.

Hypothesis 2c: The test result of a demoted hypothesis is less likely to be statistically significant than the test result of a preregistered hypothesis that is appropriately reported.

Hypothesis 2d: The test result of a changed hypothesis is more likely to be statistically significant than the test result of a preregistered hypothesis that is appropriately reported.

Because the statistical significance of omitted hypotheses is unknown, we did not formulate a hypothesis on the association between omitted hypotheses and statistical significance.

Method

Sample

We used two main sources to find published preregistrations. First, we looked at published articles that earned a Preregistration Challenge prize. The Preregistration Challenge was an educational campaign organized by the Center for Open Science (COS) in 2017 and 2018 in which researchers could earn \$1,000 if they published a study that was preregistered using a specific preregistration template (for more information, see “Preregistration Challenge,” n.d.). A full list of Preregistration Challenge prize-winning articles ($N = 180$) can be found in the OSF Zotero Library (<https://web.archive.org/web/20230305173614/https://www.zotero.org/groups/479248/osf/collections/D77RMN4N>).

Second, we looked at published articles that earned a Preregistration Badge in 2019 or before as part of the COS's Open Science Badges initiative (see “Open Science Badges Enhance Openness,” n.d.). Articles can earn a Preregistration Badge if the authors provide the URL, DOI, or other permanent paths to the preregistration in a public, open-access repository. We extracted 244 articles that earned a Preregistration Badge from a database with all articles that earned an Open Science Badge up until February 21, 2020 (Kambouris et al., 2020). After deleting duplicate articles, the total number of articles in our sample was $180 + 193 + 51 - 26 = 398$.

To assess whether these articles were from the field of psychology, we looked up their Research Areas as listed in the Web of Science Core Collection. If the article was not listed in that database, we categorized the Research Area ourselves on the basis of the publishing journal or the departmental affiliation of the authors. The articles in our sample often contained multiple preregistered studies. We considered a study separate from other studies in an article when that study was based on a different sample of participants. Each of these studies was coded separately. For the 329 psychology articles we included, we derived 613 preregistered studies.

Of these 613 preregistered studies, we omitted 48 studies because they were conducted in a registered-report framework (in which the studies are peer reviewed before data collection), 52 studies because they were part of a multilab article that did not focus on the individual studies but only on the bigger picture (e.g., Many Labs 2, Klein et al., 2018), five studies using nonhuman subjects, 14 studies because we were unable to locate a preregistration, and 14 studies because it was unclear which study was described in which (part of the) preregistration. Finally, we excluded 21 studies with preregistrations of secondary data analyses (i.e., data that already existed and were gathered to answer another research question from the one in the study) because such preregistrations qualitatively differ from those using

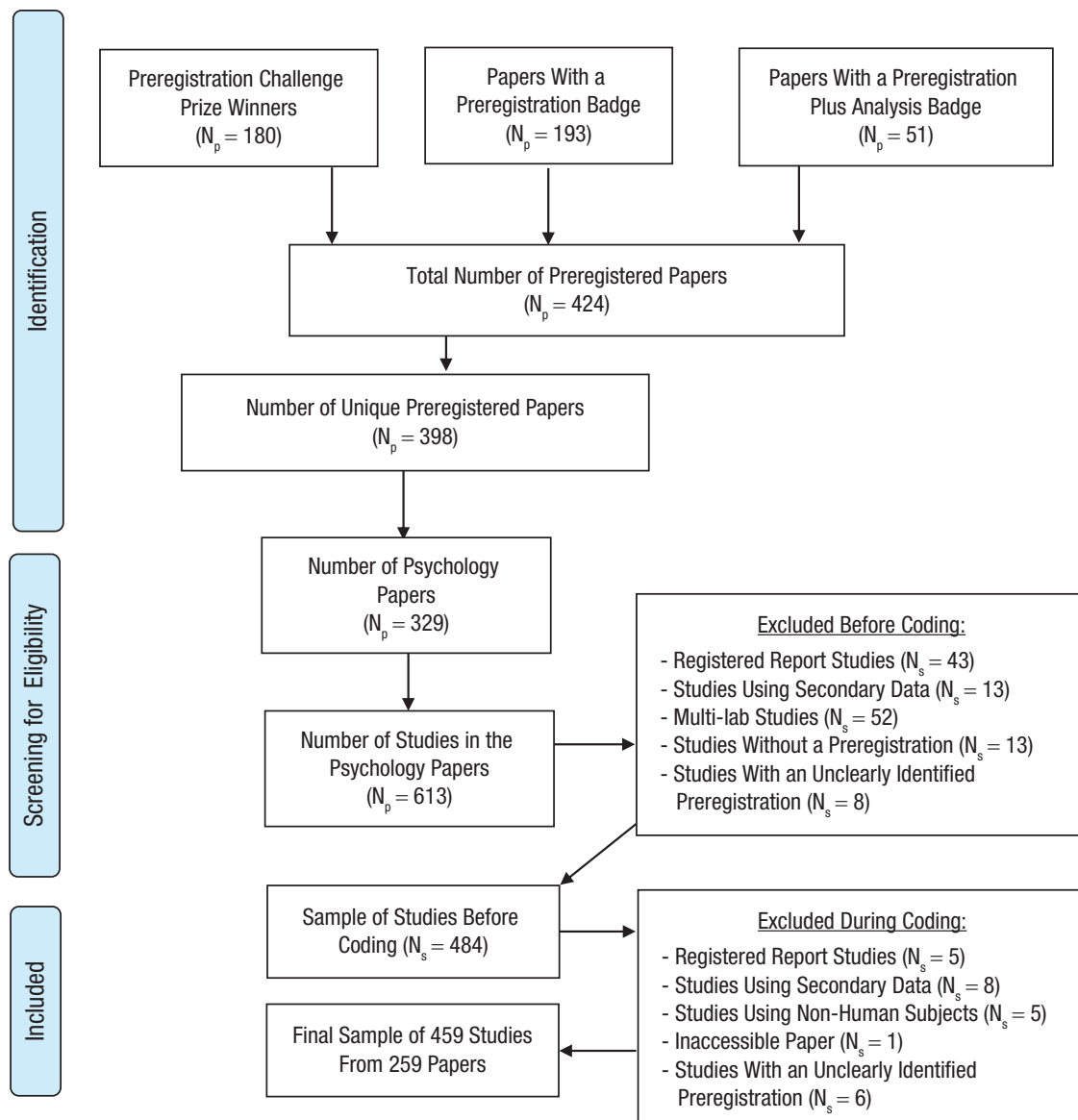


Fig. 1. PRISMA flow diagram outlining the full sample-selection procedure.

primary data (van den Akker et al., 2021; Weston et al., 2019) and would therefore have required different coding procedures. All exclusions left us with a final sample of 459 studies from 259 articles, yielding an average of 1.8 studies per article. Screening for eligible studies was done by O. R. van den Akker before coding started, although 25 exclusions (5% of the total) were made during coding. These later exclusions were made by O. R. van den Akker following advice from coders who noticed that a certain study did not match the inclusion criteria after all. A PRISMA flow diagram (Moher et al., 2009) outlining the full sample-selection procedure (including exclusions during coding) can be found in Figure 1.

Identifying hypotheses

Because we could not find a validated procedure to systematically and manually extract hypotheses from a scientific article,¹ we developed two new Qualtrics protocols: one for preregistrations (<https://osf.io/fdmx4>) and one for their accompanying publications (<https://osf.io/uyrds>). These protocols were created after a series of meetings (involving O. R. van den Akker, M. A. L. M. van Assen, M. Bakker, and J. M. Wicherts) and a series of pilots using articles not included in the eventual sample (involving all authors except for J. M. Wicherts). The protocols were preregistered before data collection.

Coding was carried out by all authors except J. M. Wicherts and consisted of four phases: (a) Two coders independently identified hypotheses in the preregistration, (b) the coders discussed any inconsistencies in their coding and resolved these together, (c) the same two coders independently identified hypotheses in the publication, and (d) the coders discussed any inconsistencies in their coding and resolved these together. Coders were trained before coding by O. R. van den Akker, who instructed them about the protocol and assessed how they coded a trial run. O. R. van den Akker provided guidance throughout this trial run until both he and the coder were satisfied about the coders' grasp of the protocol.

We identified hypotheses in preregistrations and publications by first checking whether any hypotheses were listed in a separate section. If not, we searched the running text for the following keywords (chosen based on Scheel et al.'s, 2021, analysis of hypothesis introduction phrases): "replicat," "hypothes," "investigat," "test," "predict," "examin," and "expect." We included a hypothesis if the authors hypothesized a relationship between two or more variables using any of these keywords.

If we found a hypothesis that was phrased in a conceptual way (e.g., "We expect an association between extraversion and IQ") and an operational way (e.g., "We expect an association between scores on the Multidimensional Introversion-Extraversion Scale and scores on the Wechsler Intelligence Scale for Children"), we counted only the more specific operational hypothesis because we did not want to count equivalent hypotheses twice. Moreover, we reasoned that it would be easier to identify operational hypotheses in scientific articles than it would be to identify conceptual hypotheses. If we found multiple operational hypotheses (e.g., one using the Wechsler Intelligence Scale for Children and one using the RAKIT Intelligence Test), we counted each one as a different hypothesis. Because there could be additional measures in other sections than the section in which we found the hypothesis (e.g., in the methods/measures/variables sections), we checked the entire preregistration for additional measures. The same principle holds for additional control variables (e.g., in the variables/analysis sections), so we checked the other sections for control variables as well.

To investigate whether hypotheses were omitted, we used two approaches. In the first approach, we checked whether the preregistered hypothesis was referred to as a hypothesis in the introduction or methods section of the article and, if so, concluded that the hypothesis was not omitted. In the second approach, we checked whether we could find a statistical result related to the preregistered hypothesis in the results section of the article and if so, concluded that the hypothesis was not

omitted. In this second approach, the result should have been reported in the main text, not tucked away in a table, figure, or appendix. We decided to be strict in this regard because we believe that testing the preregistered hypotheses is the reason for conducting the confirmatory study in the first place, and thus, we believe all of them should be mentioned in the main body of the article. We include both the first and second approach when we present statistics about the prevalence of selective hypothesis reporting, but we use only hypotheses omitted from the results sections for our hypothesis tests.

Of the preregistered hypotheses identified as hypotheses somewhere in the article, we checked whether they were labeled as equally important as in the preregistration. To this end, we used the keywords "key," "leading," "main," "major," "primary," and "principle" for "primary hypotheses" and "additional," "auxiliary," "minor," and "secondary" for "secondary hypotheses." If none of these words could be associated with the hypothesis, we categorized its importance as "nonspecified." We had to rely on these keywords because unlike study outcomes in biomedicine, hypotheses are typically not labeled as "primary" or "secondary" in psychology. We assessed the directionality of hypotheses in the preregistrations and articles by giving coders both a concrete indication (directional: "Men will score higher on the Verbal Aggression Scale than women"; nondirectional: "Men and women score differently on the Verbal Aggression Scale"; null: "Men and women will not differ in their scores on the Verbal Aggression Scale") and an abstract indication (directional: " $M > W$ "; nondirectional: " $M \neq W$ "; null: " $M = W$ ") of what to look for. We also assessed whether these categorizations were consistent between the article and the preregistration. These assessments gave us the necessary information to establish the prevalence of promoted hypotheses (Hypothesis 2b), demoted hypotheses (Hypothesis 2c), and changed hypotheses (Hypothesis 2d).

Because several coders indicated they were unsure about their responses related to the directionality of the hypotheses, O. R. van den Akker manually checked (and corrected) all hypotheses for which the directionality was originally coded as inconsistent between preregistration and article. The corrections can be found in the Excel file with the data (see the ManualChanges columns in <https://osf.io/8y2dv>) and were discussed with and accepted by the original coders.

We also assessed how many statistical results were presented in the article that were not related to a preregistered hypothesis and not explicitly stated as exploratory or nonpreregistered. Such added hypotheses (Hypothesis 2a) should involve a different relationship between the variables than in a preregistered hypothesis or involve a different variable or measure altogether and

should be reported in the main text rather than being tucked away in a table, figure, or appendix. In our assessment of added hypotheses, we included only studies with at least one preregistered hypothesis because we inadvertently failed to present the coders with questions about added hypotheses for studies with zero hypotheses in Qualtrics.

Because of time constraints, we assessed only selective hypothesis reporting for the first 16 preregistered hypotheses of a study even if more than 16 preregistered hypotheses were identified. In those instances, we also did not check for added hypotheses. Finally, note that the categories omitted, added, promoted, and demoted hypotheses are mutually exclusive but not exhaustive categories. In the present article, we state that a study does not include selective hypothesis reporting when it does not include any omitted, added, promoted, demoted, or changed hypotheses.

Assessing whether a hypothesis is part of a direct replication

We operationalized the replication status of hypotheses (see Hypothesis 1) in three ways. In line with Scheel et al. (2021), we assessed whether a hypothesis was part of a replication study or an original study by searching the preregistration and article for the string “replic” and assessing whether the authors referred to the hypothesis as being part of a replication attempt. If they did in either the preregistration or the article, we coded the hypothesis as a “replication hypothesis.” If they did not, we coded the hypothesis as an “original hypothesis.”

Second, we checked the articles to see whether hypotheses were part of a “direct replication” or “conceptual replication.” We coded hypotheses as part of a direct replication when the authors used the same methods (materials and procedure) to test the hypothesis as in a prior study. The methods had to be truly identical except that the replication study used a different sample and except for any translations of study materials. If the methods were not identical in this way, we coded the hypothesis as part of a conceptual replication.

Third, we logged the way the authors themselves labeled the hypotheses in the articles and coded hypotheses as part of a direct replication if the authors referred to them using any of the words “direct,” “directly,” “exact,” “exactly,” “identical,” or “direct & very close” and as part of a conceptual replication if the authors referred to them using other words (e.g., “conceptual,” “similar, except,” “close”).

We preregistered (see <https://osf.io/z4awv>) that we would use the second operationalization of replication status to test Hypothesis 1 if more than 20% of the replication hypotheses found in the articles were categorized

as direct as opposed to conceptual. However, direct-replication hypotheses constituted only 19.4% of the replication hypotheses. As preregistered, we therefore used the first operationalization of replication status for the main test of Hypothesis 1 and used the second and third operationalizations as robustness checks.

Assessing whether a hypothesis is supported

For every preregistered hypothesis for which we found a statistical result, we coded whether the result was statistically significant (see Hypotheses 2a–2d). We did this by comparing the reported p value with .05 unless the authors specifically mentioned that they used a significance level lower than .05 (e.g., because they used a Bonferroni correction). In case of the latter, we concluded that the result was significant if the p value was smaller than the authors’ significance level. If the authors reported a Bayes factor instead of a p value, we concluded that the hypothesis was supported if the Bayes factor was larger than 3. We used a threshold value of 3 because it has long been used as the value above which evidence for a hypothesis is deemed substantial (Jeffreys, 1961).² If authors specifically mentioned that they used another Bayes factor threshold other than 3, we concluded that the hypothesis was supported if the Bayes factor was larger than the authors’ Bayes factor. Given our hypothesis tests, we consider a supported Bayesian hypothesis as equivalent to a statistically significant result.

Results

Descriptive statistics

We identified 2,119 hypotheses in 459 preregistered studies from 259 articles. The number of hypotheses per study (article) is thus 4.6 (8.2); 30 studies had zero hypotheses, and 29 studies had more than 16 hypotheses. When two coders counted and coded the number of hypotheses in a preregistration, they agreed about the number of hypotheses in only 53.7% of the cases. Regarding assessing study difficulty, we found medium consistency between coders: Kendall’s $\tau = .21$, $z = 5.03$, $p < .001$.

Of all hypotheses identified in the preregistrations, we categorized 455 (21.5%) as part of a replication and 1,664 (78.5%) as original. Of all hypotheses identified in the articles, we categorized 143 (6.7%) as part of a direct replication, 595 (28.1%) as part of a conceptual replication, and 1,381 (65.2%) as original. The proportion of direct replications we found for preregistered studies (6.7%) is higher than estimates for nonpreregistered

Table 1. An Overview of the Prevalence of the Different Forms of Selective Hypothesis Reporting

	Percentage of studies ($N = 429^a$)	Percentage of articles ($N = 259^a$)	Average number per study	Average number per article
Selective hypothesis reporting	60	67	2.48	4.12
Omitted hypotheses (introduction)	56	62	2.28	3.77
Omitted hypotheses (results)	52	61	2.30	3.81
Added hypotheses ^b	57	69	4.09	6.92
Promoted hypotheses ^c	0	0	0	0
Demoted hypotheses ^d	1	2	0.09	0.16
Changed hypotheses	18	12	0.26	0.43

^aThe number of studies/articles with at least one preregistered hypothesis. ^bThe proportions were calculated using a denominator with the number of studies ($N = 400$) and the number of articles ($N = 236$) with at least one preregistered and at most 16 preregistered hypotheses.

^cThe proportions were calculated using a denominator with the number of studies ($N = 61$) and the number of articles ($N = 44$) with at least one secondary hypothesis. ^dThe proportions were calculated using a denominator with the number of studies ($N = 151$) and the number of articles ($N = 87$) with at least one primary hypothesis.

psychology studies, which range from 1.1% (Makel et al., 2012) to 2.6% (Scheel et al., 2021). At the same time, our estimate is substantially lower than the 57.8% estimate for registered reports (Scheel et al., 2021). In all, it appears that preregistered studies are more likely to be replications than nonpreregistered studies are.

The vast majority of hypotheses ($N = 1,475$; 69.6%) concerned associations/effects between two variables. The other hypothesis types were less common: interaction/moderation ($N = 326$; 15.4%), mediation ($N = 87$; 4.1%), univariate ($N = 57$; 2.7%), and other ($N = 174$; 8.2%). In the “other” category, we placed hypotheses that did not fit any of the types, such as predictions indicating atypical or complex relationships between variables (e.g., curvilinear associations or three-way interactions). Comparing hypothesis types between independent samples of preregistered and nonpreregistered studies could be an interesting follow-up project, especially if one would focus on the complexity or riskiness of hypotheses. Pham and Oh (2021) argued that the prestige premium of preregistration may result in “a bias toward studies that are easy to preregister . . . and a preference for research hypotheses that are obvious” (p. 185). In contrast, Scheel et al. (2021) proposed that researchers may deliberately preregister risky hypotheses because the negative effects of getting a small or negative result may be compensated by the credence received from preregistration.

Using our two approaches to assess omitted hypotheses, we were able to retrieve 1,143 of 2,119 preregistered hypotheses (53.9%) in the introduction or methods sections and 1,132 results of 2,119 preregistered hypotheses (53.4%) in the results section. Consequently, 976 hypotheses were missing from the introduction and methods sections (46.1%), and 987 hypothesis results were missing from the results section (46.6%). Of the 1,132 results we found in the results section, 743 (65.6%)

were statistically significant, and 389 (34.4%) were not. The number of omitted hypotheses per study and per article can be found in Table 1, as is the case for the other forms of selective hypothesis reporting we discuss below. The proportion of omitted hypotheses in the results (46.6%) is somewhat higher than earlier estimates by Ofosu and Posner (2023) and Heirene et al. (2021), who found that a little more than one-third of studies included omitted hypotheses. Based on a meta-analysis of 89 studies from mainly biomedicine, Thibault et al. (2021) estimated that 6% to 16% (95% confidence interval [CI]) of studies contain at least one omitted primary outcome, and 14% to 62% (95% CI) of studies contain at least one omitted secondary outcome. The results from this meta-analysis, which we believe is the most recent and comprehensive assessment of selective outcome reporting in biomedicine to date, are comparable with our results shown in Table 1.

Of the 401 studies with at least one and at most 16 hypotheses, we counted the number of added hypotheses (i.e., nonpreregistered statistical results). In studies with at least one added hypothesis ($N = 227$; 56.8%), the total number of added hypotheses was 1,634. The mean number of added hypotheses per study was 4.09 (see also Table 1), and the median number of added hypotheses per study was 1. The maximum number of added hypotheses in a single study was 48. Ofosu and Posner (2023) found that 18% of the studies in their sample included added hypotheses, of which 82% failed to mention that they were nonpreregistered, possibly suggesting that adding hypotheses is more common in psychology than in economics and political science. In their meta-analysis, Thibault et al. (2021) found that the number of studies with added primary outcomes in biomedicine (95% CI = [7%, 14%]) was somewhat lower than Ofosu and Posner’s estimate of 18%. The number of studies with added secondary outcomes was found to

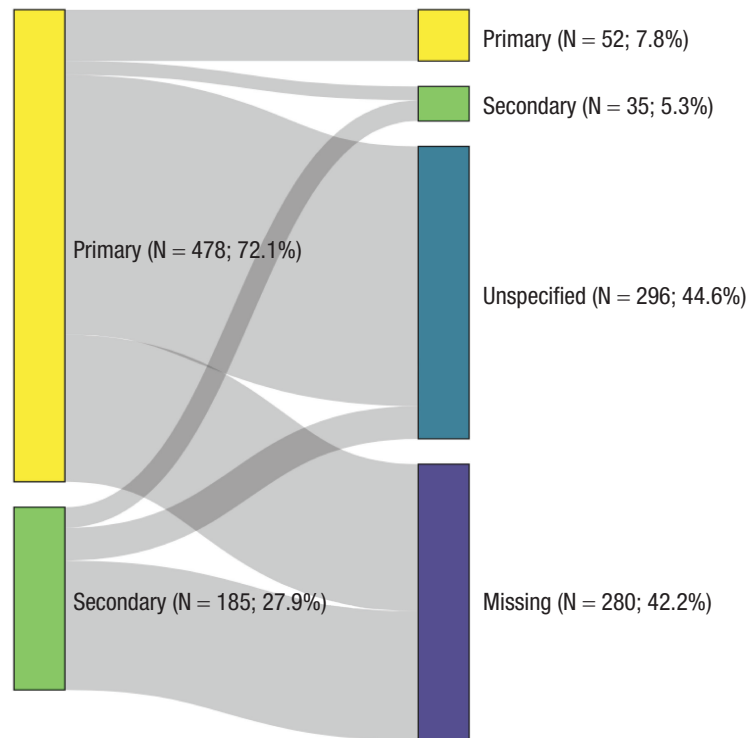


Fig. 2. Sankey diagram indicating how primary and secondary hypotheses changed from (left) preregistration to (right) article.

be 8% to 80% (95% CI), which is consistent with the estimate of both Ofosu and Posner and our estimate of 56.8%.

From all preregistered primary hypotheses that were not omitted in the article ($N = 329$), we found that 52 (15.8%) were primary in both the preregistration and the article, 14 (4.3%) were demoted from primary to secondary, although the primacy of 263 hypotheses (80.0%) was not specified in the article. From all preregistered secondary hypotheses that were not omitted in the article ($N = 54$), we found that 21 (38.9%) were secondary in both the preregistration and the article, none were promoted from secondary to primary, and the importance of 33 (61.1%) was not specified in the article. A visual depiction of the hypotheses with a change in importance between preregistration and article can be found in Figure 2. Allocating the label of “primary” to one or more hypotheses was done in 151 out of 429 studies (35.2%). This practice appears to be less common in psychological research than in biomedical research (78.7% of studies; Thibault et al., 2021), in which the prevalence of promoted (95% CI = [3%, 9%]) and demoted (95% CI = [7%, 18%]) hypotheses seems to be higher. Psychological researchers may do well to take up the distinction between primary and secondary hypotheses given that Ofosu and Posner (2023) posited that this distinction may help to prevent researchers from

determining a hypothesis’s importance post hoc on the basis of statistical significance.

Finally, we assessed the number of changed hypotheses. Of the 958 preregistered directional hypotheses that were not omitted in the article, 882 had the same direction in the article (92.1%), four had a different direction (0.4%), 69 (7.2%) became nondirectional, and three became null hypotheses (0.3%). Of the 151 preregistered nondirectional hypotheses not omitted in the article, 131 remained nondirectional in the article (86.8%), 20 became directional (13.2%), and zero became null hypotheses. Of the 65 preregistered null hypotheses not omitted in the article, 49 remained null in the article (75.4%), 14 became directional (21.5%), and two became nondirectional (3.1%). A visual depiction of the hypotheses with a change in directionality between preregistration and article can be found in Figure 3. In sum, the vast majority of hypotheses did not involve a change in direction from preregistration to article, a result mimicked by Cairo et al. (2020), who found that the direction of only 3.4% of social-psychology hypotheses changed from dissertations to published articles.

When we excluded, per our preregistration, studies that were classified as “very difficult” by the coders ($N = 73$; 17.09%), the degree of selective hypothesis reporting decreased slightly compared with our results from the whole sample. However, it is still substantial (i.e., around

Table 2. Results of the Multilevel Regression Models Testing Hypothesis 1 (Model 1) and Hypotheses 2a and 2d (Model 2)

Parameters	Model 1	Model 2
Regression coefficients (fixed effects)		
Intercept	−0.15 (0.17)	0.53* (0.18)
Level 1		
Replication	−0.92* (0.23)	—
Added	—	0.75* (0.23)
Changed	—	−0.23 (0.38)
Variance components (random effects)		
Study level	4.87 (2.21)	1.73 (1.32)

Note: Standard errors are in parentheses. Replication is a binary variable that takes on the value of 1 if the hypothesis was scored as part of a replication in either the preregistration or the article and 0 otherwise. Added is a binary variable that takes on the value of 1 if the study including the preregistered hypothesis had added hypotheses and 0 if not. Changed is a binary variable that takes on the value of 1 if the preregistered hypothesis had a direction change from preregistration to article and 0 if not.

* $p < .01$.

50% of the studies have omitted hypotheses and added hypotheses, and around 20% of studies have changed hypotheses; for the full results excluding very difficult studies, see <https://osf.io/geuxv>.

Selective hypothesis reporting and replication status (Hypothesis 1)

Preregistered analysis. To test whether selective hypothesis reporting is more common for replication hypotheses than for original hypotheses (Hypothesis 1), we employed a multilevel logistic regression with hypothesis as Level 1 and study as Level 2. The regression includes a binary dependent variable indicating whether a hypothesis is selectively reported in the publication (i.e., omitted, promoted, demoted, and/or changed) and a binary independent variable on Level 1 indicating whether a hypothesis is part of a replication. We tested Hypothesis 1 against $\alpha = .05$, as preregistered. The results indicate that hypotheses that are not part of a replication were more than twice less likely to be selectively reported than hypotheses that were part of a replication, $\beta_1 = -0.92$, $z = -4.08$, odds ratio (OR) = 0.40, 95% CI = [0.25, 0.62], $p = .00005$ (for the complete regression output, see Model 1 in Table 2). This supports our preregistered Hypothesis 1.³

Robustness analysis. As preregistered robustness checks, we ran two additional models. In the first model, we coded a hypothesis as part of a replication if the coders identified the hypothesis as a part of a direct replication using the information in the article only (6.7% of the 2,119 hypotheses described in the article: $\beta_1 = -0.92$, $z = -1.94$, OR = 0.40, 95% CI = [0.16, 1.01], $p = .052$). In the second model, we coded a hypothesis as part of a direct

replication if the authors themselves labeled the hypothesis as part of a “direct,” “exact,” “identical,” or “(very) close” replication (4.0% of the 2,119 hypotheses: $\beta_1 = -1.26$, $z = -2.46$, OR = 0.30, 95% CI = [0.10, 0.77], $p = .014$). The robustness checks showed mixed results when strictly looking at statistical significance, but the ORs were similar to or more extreme than the ORs from our main preregistered hypothesis. We therefore give precedence to the main analysis and conclude that hypotheses that are part of a replication are less often selectively reported than hypotheses that are not part of a replication. This constitutes new knowledge because earlier studies assessing selective hypothesis reporting in the social sciences did not consider replication status.

Exploratory analysis. Exploratively, we also compared whether studies in our sample that won a Preregistration Challenge prize ($N = 141$) and studies in our sample that earned a Preregistration Badge ($N = 305$) differed in the degree of selective hypothesis reporting. For this analysis, we excluded studies with both a Preregistration Challenge prize and a Preregistration Badge. We ran a multilevel model with study type (Challenge vs. Badge) as the independent variable and selective hypothesis reporting (the same variable as used in Model 1) as the dependent variable. We found that studies with a Preregistration Challenge prize less often involved selective hypothesis reporting (42%) than studies that earned a Preregistration Badge (54%), $\beta_1 = -0.97$, $z = -3.32$, OR = 0.38, 95% CI = [0.21, 0.67], $p = .001$. This difference may have come about because the Preregistration Challenge required researchers to fill out a detailed preregistration template, whereas there was no such requirement to earn a Preregistration Badge. A detailed template could have prompted

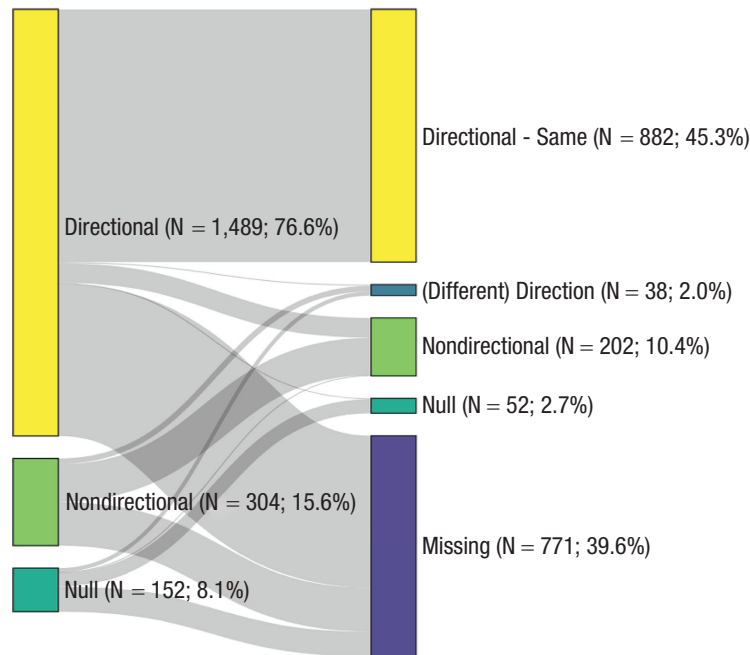


Fig. 3. Sankey diagram indicating how the directionality of hypotheses changed from (left) preregistration to (right) article.

researchers to more clearly lay out their hypotheses, which could in turn have increased researchers' sense of urgency in being consistent with their hypotheses. Alternatively, it could be that the researchers who participated in the Preregistration Challenge differed from researchers who earned a Preregistration Badge. For example, perhaps because of the added effort of filling out the template they could have been more motivated to preregister well and subsequently adhere to their preregistration. These speculative explanations would need to be tested in a confirmatory study.

Selective hypothesis reporting and statistical significance (Hypotheses 2a–2d)

Preregistered analysis. As tests of our Hypotheses 2a through 2d, we preregistered a multilevel logistic regression with hypothesis as Level 1, study as Level 2, and article as Level 3. The regression would include a binary dependent variable indicating whether the result is statistically significant and four Level 1 binary variables, each indicating whether a hypothesis is selectively reported in a certain way: added hypotheses (Hypothesis 2a), promoted hypotheses (Hypothesis 2b), demoted hypotheses (Hypothesis 2c), and changed hypotheses (Hypothesis 2d). We had to omit promoted hypotheses from our model because we did not encounter these. The remaining model did not converge when we included Level 3 or when we included demoted hypotheses. Therefore, we adjusted

our model to a two-level model that could test only Hypothesis 2a and 2d. We had preregistered the conditional move to a two-level model, but dropping the promoted and demoted hypotheses was unforeseen and thus not preregistered. We tested Hypotheses 2a and 2d against $\alpha = .01$, as was preregistered. We found that preregistered hypotheses in studies with added hypotheses were more likely to be statistically significant than preregistered hypotheses in studies without added hypotheses, $\beta_1 = 0.75$, $z = 3.18$, $OR = 2.11$, 99% CI = [1.15, 3.86], $p = .001$ (Hypothesis 2a; Model 2 in Table 2), but we did not find that changed hypotheses were more likely to be statistically significant than unchanged hypotheses,⁴ $\beta_2 = -0.23$, $z = -0.60$, $OR = 0.77$, 99% CI = [0.29, 2.05], $p = .547$ (Hypothesis 2d; Model 2 in Table 2).

Robustness analysis. As preregistered, we also ran our analyses without studies that we labeled as very difficult to code ($N = 73$; 15.9%). We still found support for our Hypothesis 1 that hypotheses that are part of a replication are less likely to be selectively reported (omitted, promoted, demoted, or changed) than original hypotheses ($\beta_1 = -0.76$, $z = -2.95$, $OR = 0.47$, 95% CI = [0.28, 0.78], $p = .003$). The robustness analysis for Hypothesis 2a was not in line with the original analysis: Preregistered hypotheses in studies with added hypotheses were not more likely to be statistically significant ($\beta_1 = 0.56$, $z = 2.35$, $OR = 1.76$, 99% CI = [0.95, 3.26], $p = .019$). The robustness analysis for

Hypothesis 2d was in line with the original analysis: Pre-registered hypotheses that were changed were not more likely to be statistically significant ($\beta_2 = -0.32$, $z = -0.82$, $OR = 0.72$, 99% CI = [0.26, 1.99], $p = .410$). We conclude that there is inconclusive evidence regarding Hypothesis 2a and a robust lack of evidence in favor of Hypothesis 2d. For an overview of the results without studies that were very difficult to code, see <https://osf.io/geuxv>.

Exploratory analysis. In hindsight, we realized that our preregistered test regarding added hypotheses was not entirely in line with our Hypothesis 2a. Although our test showed that studies with added hypotheses included more statistically significant preregistered hypotheses, we were more interested in whether added hypotheses themselves were more likely to be statistically significant than preregistered hypotheses. Therefore, we also tested this at the level of hypotheses rather than at the level of studies. Each study has a proportion of statistically significant preregistered hypotheses, p , and a proportion of statistically significant added hypotheses, a . We compared the means of these two sets of proportions using a nonpreregistered dependent t test. We found a statistically significant difference using an α of .05 but no statistically significant difference when using an α of .01, $M_{p-a} = -0.08$, $t(191) = -2.52$, $p = .013$, Cohen's $d = -0.18$. A nonparametric Wilcoxon rank sum test corroborated this result, $V = 3,844$, $p = .017$. When they compared dissertations and journal articles, Cairo et al. (2020) found that supported hypotheses were not more likely to be added than unsupported hypotheses. For biomedicine, the Thibault et al. (2021) meta-analysis indicated that 49% to 66% (95% CI) of outcome discrepancies involved a statistically significant result. Taken together, the results are not clear-cut about whether researchers in psychology and biomedicine add hypotheses primarily on the basis of statistical significance. If there is an effect, it is most likely small.

General Discussion

In this project, we assessed the prevalence of omitted, added, promoted, demoted, and changed hypotheses in psychological research. Moreover, we tested whether replication studies were more or less likely to involve these types of selective hypothesis reporting and whether these types of selective hypothesis reporting were associated with statistically significant results. We found that more than half of the preregistered studies we assessed contained omitted hypotheses ($N = 224$; 52%) or added hypotheses ($N = 227$; 57%), and about one-fifth of studies contained hypotheses with a direction change ($N = 79$; 18%). In addition, we found only a small number of studies with demoted hypotheses ($N = 2$; 1%) and no

promoted hypotheses. Replication studies were less likely to include selectively reported hypotheses than original studies, but we did not find that added and changed hypotheses were more likely to be statistically significant. We were not able to test whether promoted and demoted hypotheses were associated with statistical significance because of the low prevalence of such hypotheses.

When interpreting these results, it is important to consider the particularities of the sample we used. One consideration is that we limited our sample to studies from the Preregistration Challenge and studies that earned a Preregistration Badge, which may have negatively affected the representativeness of our results. Although we cannot substantiate the representativeness of our sample, we believe that we analyzed an important and relevant set of preregistrations in psychology for several reasons. The Preregistration Challenge and the Preregistration Badge initiatives are very well known in the psychological-science community and have fundamentally changed the preregistration infrastructure. Preregistration Badges are handed out by a large variety of psychology journals, including important journals in the field such as *Psychological Science*, *Advances in Methods and Practices of Psychological Science*, *Psychological Methods*, and the *Journal of Experimental Social Psychology*. Likewise, the Preregistration Challenge winners included articles published in a wide range of scientific journals and was paramount in the increased popularity of preregistration the field sees now (Pennington, 2023). Finally, our sample of 459 studies is the largest to date regarding both quantity and time range.

That being said, there are undoubtedly preregistrations that we overlooked by selecting our sample using these two sources. How this could have influenced our results is hard to say, but we contend that these “hidden” preregistrations might be of lower quality than the preregistrations we did select. The reason for that is that there were strict requirements for Preregistration Challenge prizes and Preregistration Badges. For example, to take part in the Preregistration Challenge, researchers were required to base their preregistrations on a detailed preregistration template. Likewise, Preregistration Badges were handed out only if several conditions were met, including that “the preregistered study design corresponds to the actual study design” and that “papers include a full disclosure of the results in accordance with the preregistration” (<https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/>). We believe these quality checks may have filtered out preregistrations of lower quality or publications with more selective reporting. The consequence of this is that the problems we identified with selective hypothesis reporting in this study may be an underestimate of issues in the wider psychological literature.

The particularities of the coding protocol are also important to discuss. While developing the protocol, we had to make many decisions to balance coding comprehensiveness and coding practicality. For example, to avoid spending a disproportionate time on single pre-registration-study pairs, we chose to assess selective hypothesis reporting for only the first 16 hypotheses we identified in a preregistration even though a preregistration could include more. Another example is that we selected the operational hypothesis when both a conceptual and an operational hypothesis were present in a preregistration. We did so because we believed that the specific nature of operational hypotheses would make them easier to retrieve in the article. Yet another example is that we tried to retrieve preregistered hypotheses only in the published article itself, not in any supplementary materials, because we believe all preregistered hypotheses should be correctly presented in the main text. Although some of these decisions may appear arbitrary, they could have substantively influenced our results and may make comparison with other studies on this topic difficult. Our coding protocols (<https://osf.io/fdmx4>, <https://osf.io/uyrds>), data (<https://osf.io/8y2dv>), and code (<https://osf.io/xjzre>) are openly available for everyone to scrutinize, and we strongly encourage readers to do this. Moreover, our protocols for identifying hypotheses and our data set could be valuable resources for metaresearchers who have research questions about hypotheses in preregistrations and/or articles or research questions about metaresearch projects like ours.

Despite our extensive protocol, the coders in our project often indicated that they struggled with identifying hypotheses in preregistrations and subsequently retrieving these hypotheses from published articles. These difficulties may be due to authors consciously or subconsciously omitting or changing hypotheses from preregistration to article. What could help to prevent this is a stricter adherence to existing reporting guidelines like CONSORT for biomedicine studies (Schulz et al., 2010) and JARS for psychology studies (Appelbaum et al., 2018). These guidelines typically emphasize that the results of all hypotheses should be reported and labeled as either “primary” or “secondary” and either “exploratory” or “confirmatory.” An alternative explanation is that hypotheses were phrased so vaguely in preregistrations, articles, or both that they could not effectively be identified or matched. This could have inflated the number of omitted hypotheses we found. Indeed, when two coders counted and coded the number of hypotheses in a preregistration, they agreed about the number of hypotheses in only 53.7% of the cases. Note that this is substantially higher than an earlier study by Bakker et al. (2020), who found agreement about the number of hypotheses in only 14.3% of cases. This difference may have come about because our more

expansive protocol left less room for the coders’ own interpretations.

On the basis of the results and the experience of the coders in this project, we believe that authors can improve the way they formulate their hypotheses. One simple recommendation would be to systematically put the hypotheses in a separate “hypotheses” section in both the preregistration and the eventual study and number all of them (possibly using letters to indicate hypotheses that are clustered together, as we did in our hypothesis section). This will help readers to quickly delineate what the hypotheses are in a (proposed) study and quickly assess whether they are selectively reported in the article. Maintaining consistency between preregistration and article is also important regarding variable names. Like Claesen et al. (2021), we frequently encountered cases in which the names of one or more of the variables in a hypothesis differed between article and preregistration, making our assessment of selective reporting challenging. Finally, it would help if all hypotheses were machine readable (Lakens & DeBruine, 2021). This would increase the reproducibility of research even more and with that, the ability to trail a hypothesis’s progress from preregistration to publication.

A more structural solution to improve the way hypotheses are phrased would be to push more strongly for the registered-reports format championed by, among others, Chris Chambers (Chambers & Tzavella, 2022; Nosek & Lakens, 2014). In the registered-report format, peer review takes place in two stages. In the first stage, the preregistration is peer reviewed, which has the advantage that ambiguously phrased or overly complex hypotheses can be identified and corrected before the study is actually carried out. In the second stage, the resulting article is peer reviewed, and reviewers explicitly compare the preregistration and the article. This explicit check might decrease the prevalence of selective hypothesis reporting in the final articles. Indeed, the first studies on the effectiveness of registered reports found that the proportion of positive results in registered-report studies was substantially lower than in nonpre-registered studies, indicating less selective reporting (Allen & Mehler, 2019; Scheel et al., 2021).

Because registered reports are not yet commonplace in research, an intermediate solution could be for editors to explicitly encourage reviewers to compare the preregistration and the article. However, finding reviewers is already challenging as it is, and requiring them to do additional tasks would not make this any easier. An increase in workload may be prevented if reviewers need to verify only whether authors do what they promised in the preregistration, such as for registered reports (Chambers & Tzavella, 2022), next to checking the appropriateness of additional (so-called exploratory) analyses. At the very least, reviewers should be required

to check whether a preregistration exists and can be easily accessed if the authors mention one. A pilot of a so-called discrepancy review, in which reviewers are explicitly assigned to check for discrepancies between preregistration and article, found that this practice is effective and could feasibly be introduced without many obstacles (TARG Meta-Research Group and Collaborators, 2022). What may also help is if reviews would have a more prominent place in the reward structure of academia, for example, by making reviews public and assigning them DOIs. This would publicly show researchers' review, which could elevate reviews to be units of prestige besides regular peer-reviewed publications. Although there are some concerns (Rodríguez-Bravo et al., 2017), this development could even be beneficial to early-career researchers (van den Akker, 2019).

In all, the field needs efforts on multiple fronts to arrive at a situation with clearer hypotheses and less selective hypothesis reporting. On an individual level, researchers, editors, and reviewers can bundle forces to make comparisons between preregistrations and articles more feasible. On a more structural level, journals can implement the registered-reports format, and employers and funders can create more effective incentives for thorough reviews. This multifaceted approach could lead to clearer and more consistent hypotheses and with that, more certainty about the validity of results in the scientific literature.

Transparency

Action Editor: Katie Corker

Editor: David A. Sbarra

Author Contribution(s)

Olmo R. van den Akker: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Software; Visualization; Writing – original draft.

Marcel A. L. M. van Assen: Conceptualization; Investigation; Methodology; Resources; Supervision; Writing – review & editing.

Manon Enting: Investigation; Writing – review & editing.

Myrthe de Jonge: Investigation; Writing – review & editing.

How Hwee Ong: Investigation; Writing – review & editing.

Franziska Rüffer: Investigation; Writing – review & editing.

Martijn Schoenmakers: Investigation; Writing – review & editing.

Andrea H. Stoevenbelt: Investigation; Writing – review & editing.

Jelte M. Wicherts: Conceptualization; Funding acquisition; Methodology; Supervision; Writing – review & editing.

Marjan Bakker: Conceptualization; Investigation; Methodology; Resources; Software; Supervision; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by a Consolidator Grant (IMPROVE) from the European Research Council (Grant 726361).

Open Practices

This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iDs

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

Marcel A. L. M. van Assen  <https://orcid.org/0000-0002-7517-6081>

Acknowledgments

We thank Sophia Crüwell, Daniel Dunleavy, Mahmoud Elsherif, and Jackie Thompson for their contributions earlier in the project and Robert Thibault for comments on the preprint.

Notes

1. For procedures to extract outcomes from biomedical articles, see Chan et al. (2004) and Thibault et al. (2021).
2. Technically, the threshold value proposed by Jeffreys was $10^{1/2} \approx 3.16$, but it was later rounded to 3 to make statistical inference easier (see Jarosz & Wiley, 2014; Wetzels et al., 2011).
3. Note that we omitted article as Level 3, as preregistered, because the model including that level did not converge. Moreover, the preregistration incorrectly stated that an odds ratio below 1 indicates more selective hypothesis reporting instead of less (see Version 3 at <https://osf.io/z4awv>).
4. One way we deviated from our preregistration was by scoring hypotheses that changed from directional to null and from nondirectional to null as 0 instead of 1 for the variable “changed” because we would expect less significant results for such changes, not more.

References

- Allen, C., & Mehler, D. M. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, 17(5), Article e3000246. <https://doi.org/10.1371/journal.pbio.3000246>
- Anderson, M. S. (2000). Normative orientations of university faculty and doctoral students. *Science and Engineering Ethics*, 6(4), 443–461.
- Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, 73(1), 3–25. <https://doi.org/10.1037/amp0000191>

- Bakker, M., Veldkamp, C. L., van Assen, M. A., Cromptvoets, E. A., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, 18(12), Article e3000937. <https://doi.org/10.1371/journal.pbio.3000937>
- Cairo, A. H., Green, J. D., Forsyth, D. R., Behler, A. M. C., & Raldiris, T. L. (2020). Gray (literature) matters: Evidence of selective hypothesis reporting in social psychological research. *Personality and Social Psychology Bulletin*, 46(9), 1344–1362. <https://doi.org/10.1177/0146167220903896>
- Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, 6(1), 29–42.
- Chan, A. W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association*, 291(20), 2457–2465.
- 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>
- DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A. J., Schroeder, T. V., Sox, H. C., & Van Der Weyden, M. B., & International Committee of Medical Journal Editors. (2005). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *Archives of Dermatology*, 141(1), 76–77. <https://doi.org/10.1503/cmaj>
- DeVito, N. J., Bacon, S., & Goldacre, B. (2020). Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: A cohort study. *The Lancet*, 395(10221), 361–369.
- Dwan, K., Altman, D. G., Clarke, M., Gamble, C., Higgins, J. P., Sterne, J. A., Williamson, P. R., & Kirkham, J. J. (2014). Evidence for the selective reporting of analyses and discrepancies in clinical trials: A systematic review of cohort studies of clinical trials. *PLOS Medicine*, 11(6), Article e1001666. <https://doi.org/10.1371/journal.pmed.1001666>
- Dwan, K., Gamble, C., Williamson, P. R., & Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias: An updated review. *PLOS ONE*, 8(7), e66844. <https://doi.org/10.1371/journal.pone.0066844>
- European Commission. (2012). *Commission guideline: Guidance on posting and publication of result-related information on clinical trials in relation to the implementation of Article 57(2) of Regulation (EC) No 726/2004 and Article 41(2) of Regulation (EC) No 1901/2006*. <https://web.archive.org/web/20230305174330/https://op.europa.eu/en/publication-detail/-/publication/9a64920e-1134-11e2-8e28-01aa75ed71a1/language-en>
- Food and Drug Administration Amendments Act of 2007. (2018). <https://www.govinfo.gov/content/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf>
- Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*, 7(1), 8–12.
- Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, 7(6), 562–571.
- Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, 17(1), 239–251.
- 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>
- Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, 7(1), Article 2. <https://doi.org/10.7771/1932-6246.1167>
- Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.
- Kambouris, S., Singleton Thorn, F., van den Akker, O. R., De Jonge, M., Rüffer, F., Head, A., & Fidler, F. (2020). *Database of articles with Open Science Badges: 2020-02-21 snapshot*. OSF Registries. <https://doi.org/10.17605/osf.io/q46r5>
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196–217.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzaska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443–490. <https://doi.org/10.1177/2515245918810225>
- Lakens, D., & DeBruine, L. M. (2021). Improving transparency, falsifiability, and rigor by making hypothesis tests machine-readable. *Advances in Methods and Practices in Psychological Science*, 4(2). <https://doi.org/10.1177/2515245920970949>
- Logg, J. M., & Dorison, C. A. (2021). Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes*, 167, 18–27.
- Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, 7(6), 537–542.
- Mazzola, J. J., & Deuling, J. K. (2013). Forgetting what we learned as graduate students: HARKing and selective outcome reporting in I–O journal articles. *Industrial and Organizational Psychology*, 6(3), 279–284.
- Merton, R. K. (1973). The normative structure of science. In R. K. Merton (Ed.), *The sociology of science: Theoretical and empirical investigations* (pp. 267–280). University of Chicago Press. (Original work published 1942).
- 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.

- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, USA*, 115(11), 2600–2606.
- Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, 45(3), 137–141.
- Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer*, 31(3). <https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science>
- O'Boyle, E. H., Jr., Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, 43(2), 376–399.
- Ofosu, G. K., & Posner, D. N. (2023). Pre-analysis plans: An early stocktaking. *Perspectives on Politics*, 21(1), 174–190. <https://doi.org/10.1017/S1537592721000931>
- Open Science Badges enhance openness, a core value of scientific practice. (n.d.). COS. <https://web.archive.org/web/20230418120332/https://www.cos.io/initiatives/badges>
- Pennington, C. R. (2023). *A student's guide to open science: How the replication crisis can reform psychology*. Open University Press.
- Pham, M. T., & Oh, T. T. (2021). On not confusing the tree of trustworthy statistics with the greater forest of good science: A comment on Simmons et al.'s perspective on preregistration. *Journal of Consumer Psychology*, 31(1), 181–185.
- Preregistration Challenge. (n.d.). Center for Open Science. <https://web.archive.org/web/20230305173237/https://www.cos.io/initiatives/prereg-more-information>
- Rodríguez-Bravo, B., Nicholas, D., Herman, E., Boukacem-Zeghmouri, C., Watkinson, A., Xu, J., Abrizah, A., & Świgoń, M. (2017). Peer review: The experience and views of early career researchers. *Learned Publishing*, 30(4), 269–277. <https://doi.org/10.1002/leap.1111>
- Scheel, A. M., Schriren, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science*, 4(2). <https://doi.org/10.1177/25152459211007467>
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Journal of Pharmacology and Pharmacotherapeutics*, 1(2), 100–107.
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association*, 54, 30–34.
- Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, 49(1), 108–112.
- 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., & Munafo, M. (2021). *Estimating the prevalence of discrepancies between study registrations and publications: A systematic review and meta-analyses*. medRxiv. <https://doi.org/10.1101/2021.07.07.21259868>
- van den Akker, O. R. (2019, October 10). *Why I think open peer review benefits PhD students*. Behavioural and Social Sciences Community. https://socialsciences.nature.com/posts/54659-why-i-think-open-peer-review-benefits-phd-students?channel_id=2140-is-it-publish-or-perish
- van den Akker, O. R., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., Hall, A., Kosie, J., Kruse, E., Olsen, J., Ritchie, S. J., Valentine, K. D., van 't Veer, A., & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, 5. <https://doi.org/10.15626/MP.2020.2625>
- Van der Steen, J. T., Van den Bogert, C. A., Van Soest-Poortvliet, M. C., Fazeli Farsani, S., Otten, R. H., Ter Riet, G., & Bouter, L. M. (2018). Determinants of selective reporting: A taxonomy based on content analysis of a random selection of the literature. *PLOS ONE*, 13(2), Article e0188247. <https://doi.org/10.1371/journal.pone.0188247>
- Vinkers, C. H., Lamberink, H. J., Tijdkink, J. K., Heus, P., Bouter, L., Glasziou, P., Moher, D., Damen, J. A., Hooft, L., & Otte, W. M. (2021). The methodological quality of 176,620 randomized controlled trials published between 1966 and 2018 reveals a positive trend but also an urgent need for improvement. *PLOS Biology*, 19(4), Article e3001162. <https://doi.org/10.1371/journal.pbio.3001162>
- 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.
- Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*, 2(3), 214–227.
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, 6(3), 291–298.