Assessment of effect sizes and the prevalence of “positive” results in Registered Reports and preregistered studies in psychology

Abstract

Background: Preregistration, (i.e., publicly registering hypotheses and methods before collecting data), has been proposed as a means to reduce “questionable research practices” and biases which are thought to result in high rates of false positives and inflated effect sizes in the literature. Additionally, Registered Reports have been developed as a format that incorporates the benefits of preregistration and also reduces publication bias by basing a journal’s decision to accept a manuscript for publication on the quality of the preregistration before results are known. Both “simple” preregistration and Registered Reports are used by a growing number of psychological researchers, but it is currently unknown if, or to what extent, the two formats’ safeguards against bias affect the reported research outcomes.

Objective: The primary goal of this study is to descriptively assess the proportion of confirmed hypotheses and the effect sizes reported in preregistered studies and Registered Reports. In order to facilitate future efforts to disentangle the “pure” effects of these publication formats from potential confounding factors, a secondary goal is to explore any relevant characteristics of the types of research for which the two formats are currently used.

Method: At least two independent coders will identify the confirmatory hypotheses, analyses, and conclusions reported in each article. The proportion of confirmed hypotheses will be calculated by dividing the number of confirmed hypotheses (according to the respective authors) by the total number of hypotheses. Effect sizes will be coded for each hypothesis and analysed using a multi-level meta-analysis of dependent effect sizes.

Population: Eligible studies have to report primary research (i.e., excluding meta-analyses) in psychology. We will analyse the full population of Registered Reports published before 19th June 2018 and the full population of two types of preregistered studies published before 19th June 2018: (1) articles with a “Preregistered Badge”, and (2) articles judged to be eligible for the “Preregistration Challenge prize” by the Center for Open Science.

Background

The scientific research and publication process is affected by several biases that generally tend to favour “positive” results that confirm the stated hypotheses over “negative” results that falsify the hypotheses or are inconclusive. This can happen via selective reporting/publishing of samples, manipulations, measures, analyses, or whole studies with positive results; via other types of “*p*-hacking” (e.g. deciding to stop or continue data collection depending on the results; Simmons et al., 2011); via HARKing (hypothesising after results are known; Kerr, 1998), in which hypotheses are changed after the fact or in which exploratory research is presented as confirmatory; or via fraudulent behaviour like fabricating, deleting, or manipulating scientific records. These processes can be classified in two broad categories: At the level of individual scientists conducting and analysing their research, they are usually called “questionable research practices” (QRPs), or they are framed as exploitation of “researcher degrees of

freedom". The latter refers to the fact that most aspects of a study that are reported entail a host of lower-level decisions made by the researcher that are often not stated transparently, and can thus be exploited to nudge results or conclusions in a certain direction. At the level of scientific publication, they are usually summarised as "publication bias", a term which simply means that manuscripts reporting "positive" results have a better chance to get published, but which remains agnostic about the specific mechanisms leading to this outcome.

For readers of a scientific publication, it is usually difficult or impossible to determine to what extent QRPs and publication bias could have influenced the presented research. Scientific publications are typically retrospective summaries, and not presented in comparison to prior plans or intermediate reports of the performed research. Researchers themselves may not be aware of making biased decisions or forgetting some of their past goals and beliefs, and they are often motivated (intrinsically and extrinsically) to present a "clean" narrative of their research rather than a more accurate, but possibly less convincing one. Further, reviewers and editors may introduce bias, either by compelling researchers to change their reports (e.g. omit negative results or frame outcomes in more convincing, but less accurate ways) or simply by deciding to reject or accept manuscripts for publication depending on whether they present positive results.

Any literature which is affected by these processes to some degree, and in which no other processes exist that fully counteract their effects, will contain (a) more false positive results (i.e., results confirming a hypothesis that is actually false) than would be expected by random chance alone, and (b) inflated effect size estimates[1]. Such an unreliable literature hinders the accumulation of knowledge: some of the acquired "knowledge" will be wrong and fail when one tries to apply it, and resources will be wasted when trying to base new research on the flawed published findings. It is thus in the interest of researchers to prevent or reduce bias in scientific research and publications.

Bias-reducing practices

Several solutions have been proposed to prevent or attenuate QRPs and publication bias. First, research reports can be made more transparent if researchers share their materials, data, and analysis scripts, and explicitly assert that they do not omit relevant aspects or parts of their work in their report (Simmons et al., 2011; Eich, 2014). Second, HARKing can be prevented, and researcher degrees of freedom constrained, if researchers publicly register their hypotheses, methods, and sampling and analysis plan before collecting their data ("preregistration", De Groot, 1969; Wagenmakers et al., 2012). Third, publication bias can be eliminated if the decision of whether or not an article gets accepted for publication is made before the data collection and/or the analyses have been carried out ("Registered Report"; Chambers, 2013)[2]. Of all these practices, Registered Reports (RRs) provide the strongest safeguard against bias: they usually require authors to share their data on a public repository,

[1] If a small, positive effect exists at the population level, different studies examining the effect with samples drawn from the population will find effect size estimates that are distributed around the true effect size (sampling error). This means that some studies will overestimate the effect, and others will underestimate the effect - to the degree that some studies may estimate the effect to be zero or even negative. In this situation, QRPs and publication bias will lead to an overrepresentation of overestimations compared to underestimations of the effect in the literature.

[2] Or before the results of the analyses are known, which can also be achieved by results-blind reviewing, see e.g. "Allied Initiatives" at https://cos.io/rr.

they offer the benefits of preregistration (preventing HARKing and constraining researcher degrees of freedom) -- combined with peer review of the preregistration, which can help weed out errors and eliminate remaining researcher degrees of freedom --, and they guarantee publication independent of the main results, removing the risk of negative results getting "file-drawered" or rejected. Additionally, RRs make it harder for reviewers and editors to coerce authors to report their results in inaccurate ways.

All of these bias-reducing practices are becoming increasingly popular in the field of psychology (and beyond). Many journals have introduced changes to their editorial policies which encourage or require more transparent reporting (Nosek et al., 2015, McNutt, 2016)[3], acknowledge open practices such as data sharing and preregistration (e.g. Eich, 2014)[4], or allow them to submit their work as a Registered Report (i.e., prior to data collection or analysis; Jonas & Cesario, 2016; Nosek & Lakens, 2014)[5]. Additionally, the Center for Open Science (COS) announced a "preregistration challenge" in 2016, promising to reward researchers with $1,000 for a published study if it had previously been preregistered on the OSF, adhering to a set of eligibility criteria[6]. At the time of writing (8th June 2018), an estimated number of around 80-90 preregistered studies which adhere to the minimal standards to be awarded a preregistration badge[7], the Preregistration Challenge prize[8], or both[9], and around 100-120 RRs [10] have been published in psychological research areas.

Given these numbers, it is now possible to empirically assess the results reported in articles that make use of these bias-reducing practices. Comparing research outcomes in preregistered studies, RRs, and non-registered studies could potentially allow us to quantify the effect of bias on research outcomes in the literature.

Potential confounds

Neither preregistering a study nor using the RR format is currently compulsory in psychological subfields. Since researchers make an active decision to use these formats, it is plausible to expect selection effects, or confounds, that make it difficult to compare research outcomes of preregistered studies and RRs with those of non-registered studies (and even with each other). For example, authors might use the RR format selectively for hypotheses with low prior odds (i.e., "riskier" hypotheses) because it guarantees publication even in the event of a negative result. Selection effects of this kind could lead to a systematic difference in the rate of true hypotheses or the true effect sizes between the formats, which in turn could cause a difference in research outcomes even if there was no difference in bias. Another plausible assumption is that researchers who choose to preregister, researchers who conduct RRs, and researchers who use neither format might differ from each other (e.g. in terms of knowledge or personality traits) in ways that could affect the bias in research outcomes even if preregistration and RRs were completely ineffective in reducing bias: For instance, researchers who are more

[3] https://cos.io/our-services/top-guidelines/
[4] https://osf.io/tvyxz/wiki/5.%20Adoptions%20and%20Endorsements/; https://cos.io/prereg/
[5] https://cos.io/rr/
[6] https://cos.io/our-services/prereg-more-information/
[7] https://www.zotero.org/groups/2155377/rr_coding_project/items/collectionKey/RXZWF8J9
[8] https://www.zotero.org/groups/2155377/rr_coding_project/items/collectionKey/BP5UWD28
[9] https://www.zotero.org/groups/2155377/rr_coding_project/items/collectionKey/PSCVVGVT
[10] https://www.zotero.org/groups/479248/osf/items/collectionKey/KEJP68G9

conscientious, and a priori less likely to use QRPs, might find preregistration and RRs more appealing than less conscientious researchers.

As long as only a small minority of articles is preregistered or published as RRs, and authors choose the format voluntarily, the outcomes of these articles cannot be taken as representative of all research that is being conducted, nor can strong inferences about the effects of bias-reducing practices be drawn from them. Estimating the "pure" influence of QRPs and publication bias on research outcomes, and the effectiveness of preregistration and RRs to reduce them, requires an extensive research programme. Such a programme would include observational studies to assess characteristics of the research questions, practices, and authors of published studies in different formats, as well as experimental studies (e.g. RCTs) to test how different formats or requirements affect research practices and outcomes. The study proposed here can be seen as a first step of this programme: an assessment of research outcomes -- specifically, effect size estimates and proportion of confirmed hypotheses -- in published preregistered studies and published RRs in psychology.

We assume that reduced bias will lead to lower proportions of confirmed hypotheses and lower effect size estimates in the published literature. Since RRs provide more, and stronger, safeguards against bias than preregistration alone, we expect RRs to have the lowest values on both of these outcome variables, the non-registered, "conventional" literature to have the highest values, and preregistered studies to fall somewhere in between. However, the aforementioned unknown confounds and selection effects make it difficult to interpret any differences we may observe. Not only that: they make it near impossible to define a meaningful non-registered comparison group in the first place. Because the number of RRs and preregistered studies that have been published to date is relatively small, it is possible to assess their respective full populations, but the same is not true for the non-registered literature, making it necessary to draw a sample. However, it is unclear what the sampling criteria should be: Should RRs and preregistered studies be compared to a non-registered sample matched by subfield, by journal, by authors, or some other criteria? The sampling procedure should be based on what is known about how the potential matching variables affect the outcome variables. Because our knowledge about any such effects is extremely limited at the moment, we postpone drawing a sample of non-registered studies to a point in the future when additional research has shed enough light on this issue to make informed decisions about sampling criteria.

The proposed study is therefore explicitly descriptive and exploratory in focus. The data we will obtain are meant to provide a first overview of research outcomes in RRs and preregistered studies, enabling more informed choices in the design of subsequent studies in this research programme. In order to help put the data into context, we will descriptively compare them to the results of previous meta-analytical, field-wide assessments of hypothesis tests and effect sizes, such as Richard et al. (2003), Fanelli (2010), Wetzel et al. (2011), Szucs & Ioannidis (2016), and Motyl et al. (2017), but refrain from drawing inferences about the effects of bias-reducing practices and safeguards. In order to increase the value of our results for future studies, we will also collect data on "peripheral" variables such as subfield and keywords.

Replication studies vs. original work

Since their inception, preregistration and RRs have been particularly popular for replicating previously published studies (see e.g. Simons, Holcombe, & Spellman, 2014). The hypotheses that are tested in preregistered replication studies, or replication RRs, have previously been subjected to the QRPs and publication bias of the non-registered literature -- they "survived" the conventional publication process. As another peculiarity, replication efforts may often be motivated by a distrust in the methods or evidence of a published study. This means that the prior odds of hypotheses tested in preregistered replication studies or replication RRs may be inherently different from those of hypotheses tested in original preregistered studies or RRs (and maybe even from the average prior odds in the non-registered literature), which would lead us to expect differences in the reported research outcomes. For this reason, we will code whether the hypotheses tested in a given preregistered study or RR constitute a replication effort or original work, and treat these as separate categories. We will classify hypotheses as replication efforts if they meet the criteria for "exact", "very close", or "close" replications proposed by LeBel, McCarthy, Earp, Elson, & Vanpaemel (in press, see their Figure 1).

Goals of the present study

The primary goal of this study is to assess (1) the proportion of confirmed hypotheses and (2) the size of reported effects in Registered Reports (RRs) and preregistered but non-RR studies (PRs) in psychology. Since the planned research is purely observational, any differences between these formats and previous observations from the non-registered literature could be due to factors other than reduced QRPs and publication bias in PRs and RRs. However, the results will provide a first estimate of the distribution of research outcomes, which can serve as a basis for later investigations of factors that might influence this distribution in different contexts.

A secondary goal is to build a database of psychological effects sourced from RRs alone, following the rationale that this population will be least contaminated by the sources of bias mentioned above. Such a database can help researchers plan new studies and base their sampling and analysis plans on more realistic assumptions than previously possible. Specifically, it allows to estimate (1) the rate of true hypotheses in the field of psychology with fewer additional assumptions than previously possible (see e.g. Johnson et al., 2017) and (2) "typical" effect sizes, and the distribution of effect sizes, in psychological research. The population size is very small at the time of writing, which limits the possible inferences. The goal is to continuously fill the database with new RRs as they get published after the initial assessment taking place in this project.

Preregistration

Outline of the planned research

We will code all confirmatory hypothesis tests in the full population of published Registered Reports (RRs), and the full population of two types of published preregistered studies (PRs): (1) articles with a "preregistration badge" and (2) articles which have been determined to be eligible for the "Preregistration Challenge" prize offered by the Center for Open Science. Our investigation will be limited to empirical research in the field of psychology. For each hypothesis test, we will code the reported test specifications and test statistics, including the effect size (or the information necessary to calculate the effect size), and the authors' explicit conclusion (confirmed, disconfirmed, inconclusive). We will calculate two main outcome variables: The proportion of confirmed hypotheses per article, and effect sizes transformed to r. We will present these outcome variables separately for RRs versus PRs, and separately for hypothesis tests of replication studies versus original studies.

Abbreviations:
PR: a preregistered study which is not a Registered Report (operationalised here as either carrying the Preregistered Badge, or judged to be eligible for the COS Preregistration Challenge prize, or both)
RR: a Registered Report

Piloting:
The coding procedure will be piloted on 8 articles from the final sample by two coders who are fully aware of the purpose and goals of the study. The following preregistration plan will then be adjusted with regard to any emerging problems or shortcomings, and subsequently re-registered (highlighting any changes), before the full sample will be coded.

A Hypotheses

We predict that all else being equal, the less a set of published articles is affected by publication bias and QRPs, the smaller the reported effect sizes and the proportions of confirmed hypotheses will be. Since RRs seek to provide more and stricter safeguards against publication bias and QRPs than PRs, we would expect smaller effect sizes and proportions of confirmed hypotheses in RRs compared to PRs, and smaller effect sizes and proportions of confirmed hypotheses in PRs compared to the non-registered literature (as reported in existing field-wide meta-analyses, e.g. Szucs & Ioannidis, 2016). However, as we have explained above, all else is not equal: A number of possible confounds currently makes it impossible to draw strong inferences from any observed differences in these outcome variables alone. In addition to that, the statistical power to detect any true differences between the formats is limited by the relatively small size of the population at this point (roughly 200 RR and PR articles in total as of 8th June, 2018), and by the large within-group variance that can be expected when studying hypothesis tests from a whole field.

For these reasons we will refrain from performing confirmatory hypothesis tests on the data we plan to collect. The two main outcome variables (effect size and proportion of confirmed

hypotheses) will be plotted per publication format (RR and PR), separately for replication studies and original studies. We will further compare them to existing data on the non-registered literature, in particular to Fanelli (2010; only proportion of confirmed hypotheses), OSC (2015), and Szucs and Ioannidis (2016; only effect sizes), but again refrain from performing confirmatory hypothesis tests.

B Method

B1 Sample

B1.1 Data source

RRs:
Articles will be identified using the Zotero list of published Registered Reports[11] maintained by the Center for Open Science. Additionally, the editors of journals which are offering the RR format as of 19th June 2018 (determined via https://cos.io/rr/), and which publish psychological research, will be contacted and asked to provide a list of all RRs that have been published in their journals. If no answer can be obtained within two weeks, all issues of the respective journal starting from the date of RR adoption will be searched by hand to identify published RRs not contained in the COS Zotero list.

PRs: Preregistration Challenge prize
Articles will be identified using the Zotero list of published Preregistration-Prize-eligible articles[12] maintained by the Center for Open Science. Additionally, the members of the Center for Open Science in charge of maintaining the list will be contacted and asked if the list is complete and up to date.

PRs: Preregistered Badge
Articles will be identified using the Zotero list of articles with open science badges[13] maintained by Fiona Fidler, Felix Singleton Thorn, and Steven Kambouris (https://osf.io/jsva7/). Additionally, the the editors of journals which are offering the Open Practice Badges as of 19th June 2018[14], and which publish psychological research, will be contacted and asked to provide a list of all articles with a Preregistered Badge that have been published in their journals. If no answer can be obtained within six weeks, all issues of the respective journal starting from the date of badge adoption will be searched by hand to identify published studies which have been awarded the Preregistered Badge.

All articles will be downloaded from the respective journal website in PDF format.

[11] https://www.zotero.org/groups/479248/osf/items/collectionKey/KEJP68G9/order/title/sort/desc
[12] https://www.zotero.org/groups/479248/osf/items/collectionKey/D77RMN4N/order/dateModified/sort/desc
[13] https://www.zotero.org/groups/2146879/open_science_badges
[14] https://osf.io/tvyxz/wiki/5.%20Adoptions%20and%20Endorsements/

B1.2 Inclusion criteria

RRs and PRs:

- The article must have been published in a peer-reviewed journal before 19th June 2018 and not have been retracted.
- The article must be written in English.
- The article must describe empirical research in the field of psychology.
  Subject area will be determined with the help of the Scopus subject area retrieved from the article's meta-data, using the R package *rscopus* and an Elsevier API key. Articles whose meta-data contain the subject area PSYC (psychology) will be included. Articles whose meta-data do not contain the subject area PSYC will be reviewed by two independent coders. If, based on reading the title and abstract, both coders agree that the article reports psychological research, it will be included. Disagreements will be resolved by discussion.
- The article must describe an empirical investigation other than a meta-analysis or systematic review (i.e., primary research)

RRs:
The article must have been published as a "Registered Report" as specified at https://cos.io/rr/.

PRs:
Articles must belong to one or both of the following two groups:

1. Determined to be eligible for the "Preregistration Challenge" prize by the Center for Open Science (eligibility criteria: https://osf.io/4uxbj/)

2. Published with a "Preregistered Badge" as specified at https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/

Articles which meet the criteria for both groups (i.e., which could be classified as RRs *and* as PRs) will be classified as RRs.

B1.3 Sample size

RRs: The full population of RRs, according to the criteria specified in B1.1 and B1.2, will be examined. As of 8th June, 2018, around 100-120 RRs have been published which meet the criteria specified in B1.1 and B1.2 (this is a rough estimate, the articles have not yet been systematically analysed at the time of writing).
Link to Zotero library

PRs: The full population of PRs, according to the criteria specified in B1.1 and B1.2, will be examined. As of 8th June, 2018, around 80-90 PRs have been published which meet the criteria specified in B1.1 and B1.2 (this is a rough estimate, the articles have not yet been systematically analysed at the time of writing).
Link to Zotero library

B1.4 Data collection termination rule

If the dataset could not be fully coded by at least two independent coders by 31st October 2018, we will reconsider if it is still feasible to finish the study.

B1.5 Exclusion criteria

There are currently no obvious reasons why an article should be excluded if it had been judged to meet the inclusion criteria. If any such reasons should emerge during the data collection, they will be made fully transparent, and the final results will be presented with and without the excluded studies (depending on which of the variables could still be coded for these excluded studies).

After the data collection has been finished, we will check all included articles for retractions that may have occurred after the start of our data collection, and exclude any such cases (given that they would no longer meet our inclusion criteria).

B2 Procedure

B2.1 Independent variables for primary analyses

(1) Article type: RR vs PR (two levels, between-"subject")

(2) Replication vs. original study (two levels, between-"subject")

B2.2 Dependent variables for primary analyses

(1) Proportion of confirmed hypotheses

(2) Effect sizes

B2.3 Coding scheme

Table 1 shows a summarised overview of the coded variables. The detailed coding scheme is available here.

As a general note: we will only code content from the final published manuscript of each article. This means that we will not verify the content of any preregistration plan that is not part of the manuscript itself, and rely on the information provided in the manuscript regarding which aspects of the research were preregistered.

Table 1. Overview of coded variables

Variable	Retrieved from/coded by
Article information:	
article type: RR vs PR (for PR articles: badge vs prize)	Anne Scheel
DOI	Anne Scheel
article title	retrieved from Scopus
author names	retrieved from Scopus
publication date	retrieved from Scopus
journal	retrieved from Scopus
subfields	retrieved from Scopus
keywords	retrieved from Scopus
handling editor	Anne Scheel
Hypotheses:	
total number of confirmatory hypotheses *Only preregistered hypotheses will be coded. If an article contains unregistered studies/hypotheses, they will be counted, but their content, results, and conclusions will not be coded.*	two independent coders
number of explicitly exploratory tests or analyses	two independent coders
for each confirmatory hypothesis:	
replication vs. original *Hypotheses will be classified as replication attempts if they meet the criteria for "exact", "very close", or "close" replications in Figure 1 of LeBel et al. (in press)*	two independent coders
content (quote), directedness, independence from other hypotheses	two independent coders
coding difficulty	two independent coders
Results:	
number of statistical tests for each hypothesis	two independent coders
for each statistical test:	
inferential statistics (e.g. type of test, test statistic, degrees of freedom, *p*-value; Bayes factor and priors)	two independent coders
decision criterion (e.g. alpha, BF-cutoff)	two independent coders
reported or recalculated effect size	two independent coders
coding difficulty	two independent coders
Conclusions:	
Authors' conclusion for each hypothesis (confirmed, disconfirmed, inconclusive)	two independent coders
coding difficulty	two independent coders
Main outcome variables	
Proportion of confirmed hypotheses	calculated as $\frac{number\ of\ confirmed\ confirmatory\ hypotheses}{number\ of\ all\ confirmatory\ hypotheses}$
Mean effect size in *r*	calculated by transforming the reported test statistic or effect size to *r* and performing a three-level meta-analysis of dependent effect sizes (Noortgate et al., 2013)

B2.5 Coding procedure

Piloting

The coding procedure will be piloted on 8 articles (4 RRs and 4 PRs, randomly chosen) from the final sample by two coders (AS and EH) who are fully aware of the purpose and goals of the study. If it becomes apparent during piloting that the preregistration plan (this document) has to be changed, it will be updated and re-registered with marked changes.

The primary purpose of the piloting phase is to test and adjust the coding scheme. A preliminary version of the coding scheme is available here and will be preregistered along with this document. After the piloting phase, it will be updated and adjusted with regard to any emerging problems or shortcomings, and subsequently re-registered, before the remaining sample will be coded.

Pilot sample

On 8th June, 2018, all RRs and PRs available through the three Zotero lists described in B1.1 were listed in two tables[15] (both will be preregistered along with this document) and numbered (ordered by last name of the first author and year). The RR list contained 126 entries, the PR list contained 81 entries. On 19th June 2018, 5 articles from each list will be randomly chosen by executing the following commands in R:

```
set.seed(x)
sample(126, size = 4, replace=FALSE)
sample(81, size = 4, replace=FALSE)
```

where "x" will be replaced by a number obtained by retrieving the measured temperature in Oslo, Norway, at 8:00 am on 19th June, 2018, from https://www.yr.no/place/Norway/Oslo/Oslo/Oslo/almanakk.html?dato=2018-06-19 (in °C with one decimal, e.g. 16.2 on 11th June 2018) and multiplying it by 10. If an article chosen with this procedure does not fulfill the inclusion criteria (see B1.2), it will be replaced by an article chosen by re-running the same code (this time with "size = 1").

Coders

Coders should should hold an undergraduate or postgraduate degree in psychology or a closely related subject area, be knowledgeable about psychological research and the psychological research literature, and have sufficient expertise in statistics to perform the coding.

Coders may only code articles which they have not co-authored.

For training purposes each coder will first code the 8 articles in the pilot sample (see above). After the piloting phase (and before further coding), we will define a cutoff criterion coders have to meet (with regard to inter-rater agreement in the pilot sample) in order to be allowed to go on to code articles from the full sample, and update this preregistration accordingly.

One coder (AS) will act as the "first coder" and attempt to code all articles in the full sample. Each article has to be coded by two independent coders. It is our goal to hold

---

15 https://docs.google.com/spreadsheets/d/1IpaWsazTaRnzmxLQsBAJNgTWEID8h_8sHeoNPxgjbsE/edit?usp=sharing

the number of "second coders" small by requiring coders to code at least 5-10 (but preferably more) articles in addition to the piloting sample.

Blinding

Each article will be downloaded in PDF format from the journal website. We will not redact information from the PDFs for the coding procedure since it is highly likely that group identity (RR vs PR) will be obvious to any reader. Additionally, PR articles may report non-registered studies alongside preregistered ones, which coders will have to record. Data coding will thus not be blind.

Procedure

Each article will be coded by two independent coders according to the above coding scheme, using an online form (created in *formr*, Arslan & Tata, 2017). Coders will mark the text for each coded hypothesis, result, and conclusion in each PDF in order to facilitate external evaluation. These marked-up PDFs will be saved and made available openly or upon request, depending on the copyright of the article in question. Our goal is to make this dataset as open as possible.
Disagreement between the coders of one article will be resolved by discussion or by a third independent coder.

C Analysis plan

C1 Calculation of the main outcome variables

(1) proportion of confirmed hypotheses (1 value per article)

→ calculate proportion of confirmed hypotheses per article

→ calculate proportion of confirmatory (vs. exploratory) tests per article

(2) effect sizes

Following OSC (2015), we will calculate effect sizes as "correlation coefficient per df", by converting test statistics (*t*, *F*, *z*, χ^2) or effect sizes other than *r* into *r*. We will use the same formulas and code for this as the OSC (see their appendix A3, https://osf.io/z7aux/):

> Whenever possible, we calculated the 'correlation coefficient per *df*' as effect size measure based on the reported test statistics. This was possible for the *z*, *χ2*, *t*, and *F* statistic. The code for the calculation is:

```
esComp <- function(
x,
df1,
df2,
N,
esType){
esComp <- ifelse(esType=="t",
sqrt((x^2*(1 / df2)) / (((x^2*1) / df2) + 1)),
ifelse(
esType=="F",
```

```
sqrt((x*(df1 / df2)) / (((x*df1) / df2) + 1))*sqrt(1/df1),
ifelse(
esType=="r",
x,
ifelse(
esType=="Chi2",
sqrt(x/N),
ifelse(
esType == "z",
tanh(x * sqrt(1/(N-3))),
NA
)
)
)
))
return(esComp)}
```

The *z* statistic is transformed into a correlation using sample size *N* with $z = r_f \sqrt{(N-3)}$, with r_f the Fisher-transformed correlation. The *χ2* is transformed into the or correlation coefficient with $\phi = \sqrt{\chi^2/N}$. The *t* and *F* statistic are transformed into a 'correlation per *df*' using $r = \sqrt{\frac{F\frac{df_1}{df_2}}{F\frac{df_1}{df_2}+1}\sqrt{\frac{1}{df_1}}}$, where $F = t^2$. The expression in the first square-root equals the proportion of variance explained by the df_1 predictors of the variance not yet explained by these same predictors. To take into account that more predictors can explain more variance, we divided this number by df_1 to obtain the "explained variance by predictor". Taking the square root gives the correlation, or more precisely, it gives the correlation of each predictor assuming that all df_1 predictors contribute equally to the explained variance of the dependent variable.

(OSC 2015, appendix A3, p. 7)

If results are presented as Bayes factors only (i.e., none of the test statistics above, no effect size, and no information allowing to re-calculate the effect size are given), and a scaled Cauchy prior is used, the following code by Richard Morey (retrieved from https://gist.github.com/richarddmorey/1b408eaa608943379e01204fc4333bd5) will be used to calculate *t*-values from the BF_{10} (if Bayes factors are reported as BF_{01} only, the BF_{10} will be calculated as $1/BF_{01}$):

```
## Given a (scaled Cauchy) Bayes factor for the null against the
## alternative (Rouder et al 2009), yields the t statistic
## that would yield it. The ... arguments are passed to
## the ttest.tstat function.
bf.inv = Vectorize(function(b10, ...){
  fn = Vectorize(function(t,...){
    BayesFactor::ttest.tstat(t,...)[["bf"]]
  }, "t")
  t0 = optimize(function(t0, ...){
    t = sqrt(t0 / (1 - t0))
    b = fn(t, ...)
    (b10 - exp(b))^2
  }, c(0,1), ...)$minimum
```

```
  sqrt(t0 / (1 - t0))
},"b10")
```

(Morey, 2017, https://gist.github.com/richarddmorey/1b408eaa608943379e01204fc4333bd5)

We will present all obtained effect sizes individually in a table following PRISMA guidelines.We expect that every hypothesis will be examined with typically one, but possibly more than one result, and that multiple hypotheses will be tested in the same article. We will therefore report descriptives after performing a three-level meta-analysis of dependent effect sizes, following Noortgate, López-López, Marín-Martínez, and Sánchez-Meca (2013).

C2 Planned data analysis

No confirmatory hypothesis tests are planned.
We will compare the central tendency and distribution of the two main outcome variables (proportion of confirmed hypotheses and average effect size) between RRs and PRs using summary statistics and by visualising the data, but we will not perform a significance test.

References

Arslan, R.C., & Tata, C.S. (2017). formr.org survey software (Version v0.17.12).

De Groot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences* (J. A. A. Spiekerman, Trans.). In J. T. Barendregt et al. (Eds.), Psychological Studies (Vol. 6). The Hague, Netherlands: Mouton & Co.

Eich, E. (2014). Business Not as Usual. *Psychological Science*, *25*(1), 3–6. https://doi.org/10.1177/0956797613512465

Fanelli, D. (2010). "Positive" Results Increase Down the Hierarchy of the Sciences. *PLoS ONE*, *5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the Reproducibility of Psychological Science. *Journal of the American Statistical Association*, *112*(517), 1–10. https://doi.org/10.1080/01621459.2016.1240079

Jonas, K. J., & Cesario, J. (2016). How can preregistration contribute to research in our field? *Comprehensive Results in Social Psychology*, *1*(1–3), 1–7. https://doi.org/10.1080/23743603.2015.1070611

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

McNutt, M. (2016). Taking up TOP. *Science*, *352*(6290), 1147–1147. https://doi.org/10.1126/science.aag2359

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., … Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology*, *113*(1), 34–58. https://doi.org/10.1037/pspa0000084

Noortgate, W. V. den, López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. Behavior Research Methods, 45(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, Brian A., & Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

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

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An Introduction to Registered Replication Reports at *Perspectives on Psychological Science*. *Perspectives on Psychological Science*, *9*(5), 552–555. https://doi.org/10.1177/1745691614543974

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

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. https://doi.org/10.1177/1745691611406923