Appendix

This document contains additional information about the methods and analyses reported in the manuscript '**An excess of positive results: Comparing the standard Psychology literature with Registered Reports**', as well as an additional robustness analysis (see section 4.1). This PDF was created in RMarkdown. The reproducible RMarkdown file, containing embedded code for the analysis in section 4.1, is available in our online repository at https://osf.io/dbhgr (`appendix/appendix.Rmd`).

1. Project timeline

Time	Event
Early September 2018	Project starts as the Bachelor End Project (BEP) of Mitchell Schijen (MS) under the supervision of Anne Scheel (AS) and Daniël Lakens (DL) at Eindhoven University of Technology
November 2018	Registered Reports population is determined by retrieving the Zotero database curated by the Center for Open Science
Early December 2018	Piloting
7th December 2018	Study is preregistered at https://osf.io/s8e97
January 2019	- Standard reports sample is drawn - MS and AS code all sampled standard reports and all Registered Reports (coding round 1) - MS submits BEP thesis
Spring/summer 2019	- Three unclear cases from the RRs sample are confirmed as RRs and coded (coding round 2) - Eight new standard reports are sampled and coded to replace the excluded ones (coding round 2); one of them has to be excluded and gets replaced by resampling (coding round 3) - AS and DL develop new coding criteria for replication status and re-code all SRs and RRs accordingly
Autumn 2019	- One RR and 2 SRs are found to have been been excluded erroneously in coding round 1 and should be included (i.e., sampling 8 replacement SRs had led to oversampling and a final sample of 152 SRs) - We find that one RR had accidentally been coded based only on the Stage-1 protocol (including pilot data but no final results) and replace it with the final Stage-2 report (coding round 4) - AS revises the hypothesis introduction phrases for all RRs (based on the existing hypothesis quotes)

2. Comparison of preregistration and eventual procedure

We preregistered our study on 7th December 2018. The preregistration document can be found at https://osf.io/sy927/. To facilitate the comparison between the preregistration and the procedure and analyses we eventually carried out, we annotated the preregistration document in Word, adding comments that for each 'decision unit' note whether it was carried out as preregistered and give a brief overview of relevant changes or additional details. The annotated preregistration document is provided in the file `appendix/preregistration_annotated.docx` (to view the annotations the file must be opened with a word processor like MS Word).

3. Method

3.1 Sample

3.1.1 Registered Reports

As described in the main manuscript, we selected Registered Reports by retrieving the Registered Reports database curated by the Center for Open Science[1] (COS) on 19th November 2018, determining the discipline each paper belonged to and excluding all non-Psychology papers, and then verifying if all Psychology papers on the list were indeed Registered Reports. The file `raw_data/source_data/RR_database_categorised_Nov2018.csv` contains the full list of papers that were included in the COS database when we accessed and copied it. The file also contains a variable indicating the discipline category we identified for each paper ('Discipline'), whether the papers was indeed a Registered Report ('RR', only checked for Psychology papers), and optional comments with details regarding this coding process ('comments').

Excluding Registered Replication Reports (RRRs): The decision to exclude all RRRs was made during the main data collection (coding round 1), i.e. after piloting and after the study had been preregistered. MS and AS both noted that RRRs were particularly difficult to code because no clear hypothesis and/or no clear conclusions could be identified. This led us to look up the editorial information for the RRR format, which clearly states that RRRs are not supposed to test hypotheses:
'The conclusion of a Registered Replication Report should avoid categorizing each result as a success or failure to replicate. Instead, it should focus on the cumulative estimate of the effect size' (retrieved from https://www.psychologicalscience.org/publications/replication). Therefore we decided to categorically exclude all RRRs, regardless of whether or not we were able to code them.

3.1.2 Standard reports

Fanelli (2010) describes his sampling strategy as follows: 'The sentence "test* the hypothes*" was used to search all 10837 journals in the Essential Science Indicators database, which classifies journals univocally in 22 disciplines. When the number of papers retrieved from one discipline exceeded 150, papers were selected using a random number generator' (p. 8). The author informed us that the search was carried out in Web of Science using a Boolean search query which contained the key phrase and the full list of journal names for each scientific discipline (personal communication, 5th October 2018). We implemented this exact procedure with the following changes: We used a current version of the Essential Science Indicators (ESI) database[2] (retrieved on 4th December 2018), we only searched journals from the 'Psychiatry/Psychology' discipline, we used ISSNs instead of journal names in our search query (to prevent the risk that names of journals on the list could be sub-strings of names of journals not on the list, which would then be searched as well), and we restricted the search to papers published between 2013 and 2018 (to match them to the RR population).

Using a programming script written in Processing (The Processing Foundation, 2019) a single search query compatible with Web of Science was generated that contained the key phrase that had to occur in the paper ('test* the hypothes*'), as well as the desired publication years (2013 up to and including 2018), and the

[1] retrieved from https://www.zotero.org/groups/479248/osf/items/collectionKey/KEJP68G9
[2] retrieved from http://help.incites.clarivate.com/incitesLiveESI/ESIGroup/overviewESI/esiJournalsList.html

ISSNs of all 633 journals listed as Psychiatry/Psychology journals in the ESI database (the journal list is provided in the file `raw_data/source_data/ESI_journallist_Psychology_20181204.csv`). The resulting search query is available in the file `raw_data/source_data/SR_search_query_WoS.txt`.

The search yielded 1919 results which were exported and saved in the file `raw_data/source_data/SR_search_results_WoS_20190107.csv`. We randomly selected 150 papers from this list using the sample() function in R and the seed '20190120':

```
# Load all 1919 search results
allSRs <- read.csv("SR_search_results_WoS_20190107.csv")

# Draw a random sample of 150 papers and put them into a new dataframe
set.seed(20190120)
x <- sample(c(1:1919), 150)
SRsample <- allSRs[x, ]
```

The same strategy and random seed were used to sample 8 replacement papers and later another replacement paper after one of the initial 8 had to be excluded:

```
# Sample 8 additional papers and put them into a new dataframe
set.seed(20190120)
replacement <- sample(c(1:1919), 158)[151:158]
SRreplacement <- allSRs[replacement, ]

# Sample 1 additional paper and put it into a new dataframe
set.seed(20190120)
replacement2 <- sample(c(1:1919), 159)[159]
SRreplacement2 <- allSRs[replacement2, ]
```

We checked if there was any overlap between the SR sample and the RR population (naturally, the SR sample should not contain any RRs) but this was not the case at any point.

3.2 Coding procedure

3.2.1 Hypothesis support

In our email correspondence with Daniele Fanelli, we asked for potential additional details of the coding procedure, but were told that not 'much more' than the information given in Fanelli (2010) had been used for coding instructions. All criteria we used are described in the main manuscript.

3.2.2 Replication status

Since the difference between replication studies and original studies was not the main focus of our paper, and because definitions of (e.g.) direct versus conceptual replications can be difficult to apply, we originally aimed for a 'lightweight' coding criterion and coded two binary variables: 'contains replication' and 'contains original work'. Both of these pertained to a full paper and were not restricted to the first hypothesis, meaning that a paper could be coded as 'contains a replication' even if the coded hypothesis was clearly novel. A paper could be coded as containing only replication research (i.e., 'contains replication' = 1, 'contains original work' = 0), only original work (i.e., 'contains replication' = 0, 'contains original work' = 1), or both (i.e., 'contains replication' = 1, 'contains original work' = 1), but not neither. However, we did not define any precise criteria for what would vs would not constitute a replication, which led to coding disagreements between MS and AS for a relatively large number of papers (MS had used a relatively lenient, AS a quite strict definition of 'contains replication'). We therefore decided (after MS had completed his BEP thesis) to completely revise the way in which replication status was coded.

The main reason why we coded replication status in the first place was the idea that direct replication studies may sometimes or often be motivated by the authors' scepticism of the original result, and that such

direct replication studies may be overrepresented in the Registered Reports population compared to the standard literature. Therefore, we chose to focus on whether the authors' goal appeared to be to test if a previously published result holds up by conducting a close replication of the respective original study. We would thus ignore replications that had the goal to test if a hypothesis or method holds in a new context or population (i.e., that did not have the goal to challenge or verify the results in the original context or population) and 'internal' replications (i.e., replications of studies published in the same paper).

To limit the amount of work that would be caused by re-analysing all coded hypotheses, we decided to first search the full texts of all papers for the string 'replic' and only re-analyse the hypotheses of papers that contained the string (assuming that it would be highly unlikely that authors conducting a close replication in the way described above would not use words such as 'replication' or 'replicate' anywhere in their paper).

3.2.3 Hypothesis introductions

Hypothesis introduction phrases were first coded by MS and AS as short phrases taken from the coded hypothesis quotes. The idea was to extract phrases roughly analogous to Fanelli's (2010) search phrase 'test* the hypothes*', but no clear coding criteria were defined. For the final analysis of hypothesis introductions in Registered Reports, AS revised the coding for all Registered Reports (but not standard reports) by starting with the coded hypothesis quotes and extracting minimal phrases that signalled that a hypothesis was being tested, stripped from all content-specific details in a consistent manner across all papers. Rather than focussing on the 'best' or just the first identifiable phrase, all identifiable hypothesis introduction phrases were coded, meaning that each paper could have one or several phrases. There was only one Registered Report for which no hypothesis introduction could be identified.

The revised hypothesis introduction phrases were captured in two new variables, 'RR_hyp_intro_abstract' and 'RR_hyp_intro_fulltext', depending on whether the phrase had been found in the abstract or the full text of the respective paper (this information was relevant since typically the contents of abstracts, but not full texts, are searchable via search engines). The revised hypothesis introduction phrases were coded by AS alone and not double-coded. However, they are very similar to the phrases coded by MS and AS in the initial coding round (the main difference being increased consistency across papers).

3.2.4 Additional measures

For the sake of completeness, we report all other recorded measures not described thus far in the following (*all* recorded measures are included and described in the dataset `raw_data/positive_results_in_registered_reports_data.xlsx`).

- We coded direct quotes of either 'finding', 'conclusion', or both, depending on how the hypothesis-relevant results were reported (we adopted these two variables from the data excerpt provided by Fanelli).
- To track the coding process and facilitate double-coding and disagreement resolution, we coded
  - time spent per paper
  - coding certainty (scale from 1 = not certain to 5 = very certain),
  - if the full text could be accessed
  - if the hypothesis had been coded from the abstract or full text
  - if the finding/conclusion had been coded from the abstract or full text
  - any individual coder's remarks
  - any remarks regarding the resolution of coding disagreements
- While revising our coding criteria for replication status, we also coded if the coded hypothesis introduction for a paper contained the string 'replic' and if a paper was part of a special issue focussed on replication, but in the end did not use these two variables (we did not run any analyses on them).

4. Analyses

4.1 Robustness analysis

Our preregistration does not specify whether or not excluded standard reports would be replaced (by resampling). In the initial analysis for his BEP thesis, MS did not replace any papers due to time constraints (his analysis included 142 SRs and 68 RRs), but we later decided that the SR sample should be filled up to achieve the planned sample size of 150 standard reports. As mentioned above and in the manuscript, we resampled SRs until 8 additional ones that met our inclusion criteria had been found, but later realised that 2 SRs excluded in the initial analysis should actually be included, meaning that we had oversampled and achieved a sample size of 152 rather than 150.

Because the decision to add 8 SRs could potentially affect our results, we re-ran our main analysis without the resampled papers as a robustness check. This analysis thus includes the 144 standard reports from the first sample we drew that met our inclusion criteria and the same 71 Registered Reports as the main analysis reported in the manuscript. Note that these are 3 Registered Reports more than included in MS' initial analysis: One had been part of the initial sample but was excluded for reasons we later found were erroneous, and two were cases about whose Registered Report status we were not sufficiently sure at the time of MS' initial analysis, but which we later learned met the Registered Reports criteria and could be included.

Of the 144 standard reports sampled in Round 1, 138 had positive results, a positive result rate of 95.83% (95% CI [91.15, 98.46]). The difference between this group of standard reports and all Registered Reports (52.17%) is thus nearly identical to the one found in the main analysis (52.39%). Recalculating the one-sided proportions test and the equivalence test (both with an alpha level of 5%) shows that this difference is still statistically significant, $\chi^2 = 73.89$, $p < .001$, and still not statistically equivalent to a range between -6% and 6%, $z = 7.55$, $p > .999$, meaning that we cannot reject differences more extreme than 6%. We thus conclude that the decision to replace excluded standard reports by resampling does not affect our results or conclusions in a meaningful way.

4.2 Calculating combinations of the proportion of true hypotheses and statistical power compatible with the observed results (Figure 3)

The following serves to further explain the calculation and reasoning underlying Figure 3 in the manuscript.

Psychologists most commonly test hypotheses using frequentist statistics, specifically null-hypothesis significance testing (NHST). Typically this means testing a null hypothesis which predicts that the tested effect or relationship is exactly zero. If a significant p-value is observed ($p < \alpha$, typically with $\alpha = .05$), the null hypothesis is rejected and its complement, the alternative hypothesis (which predicts that the tested effect or relationship is *not* zero), is accepted. Usually it is the alternative hypothesis that researchers are interested in, or that most closely corresponds to the 'research hypothesis' that is being studied.

For example, a research hypothesis could be 'watching an episode of Sesame Street increases positive affect as measured with the PANAS'. This hypothesis would typically be studied by testing the null hypothesis that the effect of watching Sesame Street on positive affect is exactly zero. A significant result would then be interpreted as *support* for the research hypothesis. This is why we refer to significant results as 'positive' results: They are typically used to argue in favour of the research hypothesis presented by the researcher, not to argue *against* the null hypothesis (which mainly serves as a statistical tool and is not itself of focal interest).

Statistically, the probability of observing a significant result depends on three factors: the probability that the tested null hypothesis is true, the probability of obtaining a significant result when the null hypothesis is true (α), and the probability of obtaining a significant result when the null hypothesis is false (power, i.e. $1 - \beta$). However, if research hypotheses are commonly the alternative hypothesis and thus complementary to the null hypothesis (meaning that if one is true, the other must be false, and vice versa), and significant results are 'positive' results in favour of the research hypothesis, we can flip the language in the following way:

> The probability of observing a positive result depends on three factors: the probability that the

research hypothesis is true (i.e., that the null hypothesis is false), the probability of obtaining a positive result when the tested research hypothesis is false (α), and the probability of obtaining a positive result when the research hypothesis is true (power, i.e. $1 - \beta$).

This can be expressed in the equation presented in the note of Figure 3 in the manuscript:

$$PRR = (1 - t) * \alpha + t * (1 - \beta)$$

PRR is the positive result rate (the proportion of positive results, i.e. the number of positive results divided by the number of all results), t is the proportion of true (research) hypotheses (i.e., the number of true hypotheses divided by the number of all hypotheses), α is the probability of obtaining a positive result when testing a false hypothesis, and $1 - \beta$ is the probability of obtaining a positive result when testing a true hypothesis (power). By solving this equation for t and for $1 - \beta$, respectively, we can calculate the combinations of these two factors that would produce a given positive result rate for a given level of α. Figure 3 in the manuscript was produced by solving for power, $1 - \beta = (PRR - \alpha + \alpha * t)/t$, and calculating power for all levels of t from 0 to 1, $\alpha = .05$, and the observed positive result rates of original standard reports, original Registered Reports, and all Registered Reports.