Introduction to Recent “Open Science” Initiatives

Nullius in verba: “on the word of no one”
Is the published literature an unbiased representation of the available information?

| Journals: All Issues From January To December | Total Number of Research Reports (1) | Number of Research Reports Using Tests of Significance (2) | Number of Research Reports that Reject $H_0$ with $Pr(E|H_0) \leq .05$ (3) | Number of Research Reports that Fail to Reject $H_0$ (4) | Number of Research Reports That are Replication of Previously Published Experiments (5) |
|---------------------------------------------|-------------------------------------|--------------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-----------------------------------------------------------|
| Experimental Psychology (1955)              | 124                                 | 106                                                   | 105                                             | 1                                               | 0                                                          |
| Comparative and Physiological Psychology (1956) | 118                                 | 94                                                    | 91                                              | 3                                               | 0                                                          |
| Clinical Psychology (1955)                  | 81                                  | 62                                                    | 59                                              | 3                                               | 0                                                          |
| Social Psychology (1955)                    | 39                                  | 32                                                    | 31                                              | 1                                               | 0                                                          |
| Total                                       | 362                                 | 294                                                   | 286                                             | 8                                               | 0                                                          |
Statistical power in psychology has been estimated to be ~50% for detecting $d = 0.5$ (Cohen, 1962), 48% for detecting the median effect size of $r = 0.31$ (Sedlmeier & Gigerenzer, 1989), and as low as 38% for JPSP in 2011 (Schimmack, 2014).
This article presents evidence that published results of scientific investigations are not a representative sample of results of all scientific studies. Research studies from 11 major journals demonstrate the existence of biases that favor studies that observe effects that, on statistical evaluation, have a low probability of erroneously rejecting the so-called null hypothesis ($H_0$). This practice makes the probability of erroneously rejecting $H_0$ different for the reader than for the investigator. It introduces two biases in the interpretation of the scientific literature: one due to multiple repetition of studies with false hypothesis, and one due to failure to publish smaller and less significant outcomes of tests of true hypotheses. These practices distort the results of literature surveys and of meta-analyses. These results also indicate that practice leading to publication bias have not changed over a period of 30 years.

KEY WORDS: Bias; Null results; Publication bias; Tests of significance.
Open Science Collaboration (2015)
• Replicated 100 findings that were published in 2008
• Replication attempts had >90% power to detect original effects
• Only 36% of all studies replicated using statistical significance as criterion
• Only 25% of social psychology studies replicated using statistical significance as criterion
• Strength of evidence of original study predicted replication, whereas characteristics of original and replicating team did not

### Sample
<table>
<thead>
<tr>
<th>Sample</th>
<th>Standard meta-analysis models</th>
<th>Statistical power and $p$-values for the binomial tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FE: Overall effect</td>
<td>RE: Overall effect</td>
</tr>
<tr>
<td>Full</td>
<td>0.62 (0.58, 0.66)</td>
<td>0.68 (0.63, 0.74)</td>
</tr>
<tr>
<td>CI</td>
<td>0.74 (0.69, 0.77)</td>
<td>0.78 (0.76, 0.83)</td>
</tr>
<tr>
<td>CP</td>
<td>0.60 (0.52, 0.68)</td>
<td>NA</td>
</tr>
<tr>
<td>CV</td>
<td>0.24 (0.13, 0.35)</td>
<td>NA</td>
</tr>
<tr>
<td>SP</td>
<td>0.69 (0.60, 0.78)</td>
<td>NA</td>
</tr>
</tbody>
</table>

*Full = the full sample; CI = controlling impulses subsample; CP = cognitive processing subsample; CV = choice and volition subsample; SP = social processing subsample; FE = fixed-effect; RE = Random-effects. All $p$-values for overall effects and for the $Q$ statistics, with the exception of for the CP CV, and SP subsamples, were less than 0.001. Numbers given in parentheses are the lower and upper limits of the 95% confidence intervals. For each binomial test, average power was calculated from three possible sources: Pow$_{ind}$ = power based on the effect sizes reported within individual experiments; Pow$_{FE}$ = power based on the fixed-effect meta-analysis estimate of the overall effect size; Pow$_{RE}$ = power based on the random-effects meta-analysis estimate of the overall effect size.*
Registered Replication Reports: Ego Depletion

<table>
<thead>
<tr>
<th>Sample</th>
<th>PET</th>
<th>PEES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b_0$</td>
<td>$b_1$</td>
</tr>
<tr>
<td>Full</td>
<td>$-0.10$</td>
<td>$2.72^{***}$</td>
</tr>
<tr>
<td>CI</td>
<td>$0.21^{*}$</td>
<td>$0.22^{***}$</td>
</tr>
<tr>
<td>CP</td>
<td>$0.02$</td>
<td>$0.39$</td>
</tr>
<tr>
<td>CV</td>
<td>$0.06$</td>
<td>$0.27$</td>
</tr>
<tr>
<td>SP</td>
<td>$0.18$</td>
<td>$0.47$</td>
</tr>
</tbody>
</table>

$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05; ^{0.1}p < 0.10$. Full = the full sample; CI = controlling impulses subsample; CP = cognitive processing subsample; CV = choice and volition subsample; SP = social processing subsample. For PET and PEES, $b_0$ = the intercept (i.e., the corrected estimate of the overall effect), $b_1$ = the coefficient for standard error or variance (i.e., the test for funnel plot asymmetry). Numbers given in parentheses are the lower and upper limits of the 95% confidence intervals.
How “Open Science” is tackling these issues?

Curate Science: All Replication Attempts of Ego Depletion Studies
By developing tools and incentives to improve research!
3 ways “Open Science” can be incorporated into your research

1. Use collaborative tools that archive your research workflow
2. Preregistering your research
3. Communicate your adherence to good research practices

Using collaborative tools that archive your research workflow
Goals of archiving your research workflow

- Maintain version control of all research materials
- Have non-local storage
- Can be shared privately among research team and, if you want, publically among research community
Other tips to archive your data

• Annotate your research workflow
  • Keep a lab journal or a project wiki in the OSF
  • Annotate your analysis code (I use Google’s Style Guide for R)
  • Create your materials as if they will be shared. The most likely person to access your materials is future you. Future you will thank you for the effort.

• Minimize the extent to which the information resides in one person’s head (Bus Rule!)
  • Parallel coding, parallel analyses, etc.

• Save your data in plain text files (such as .csv files)
  • Is readable by most statistical software programs
  • Is relatively robust to software updates
  • SPSS codes expire
Wichert (2006) found that 73% of authors defied APA guidelines of not sharing their data for reanalysis.

Wichert, Bakker, & Molenaar (2011)

- Reluctance to share data associated with weaker evidence against the null and more reporting errors.
- Particularly when evidence had bearing on statistical significance (using $\alpha = .05$).
What if you want to share your materials?

- It’s as easy as clicking “make public” on OSF
- If you want credit for stimuli/data, you can use figshare.org to create doi for individual components
Pre-registering your studies
(i.e., how to fool yourself less)

Same Data, Different Conclusions
Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.
Simmons et al. demonstrated how easy it is to inflate Type 1 error:
• With as few as 4 data analysis decisions, Type 1 error rate can go from $\alpha = 0.05$ to $\alpha = 0.607$. 

---

**Table 1. Likelihood of Obtaining a False-Positive Result**

<table>
<thead>
<tr>
<th>Researcher degrees of freedom</th>
<th>$p &lt; .1$</th>
<th>$p &lt; .05$</th>
<th>$p &lt; .01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situation A: two dependent variables ($r = .50$)</td>
<td>17.8%</td>
<td>9.5%</td>
<td>2.2%</td>
</tr>
<tr>
<td>Situation B: addition of 10 more observations per cell</td>
<td>14.5%</td>
<td>7.7%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Situation C: controlling for gender or interaction of gender with treatment</td>
<td>21.6%</td>
<td>11.7%</td>
<td>2.7%</td>
</tr>
<tr>
<td>Situation D: dropping (or not dropping) one of three conditions</td>
<td>23.2%</td>
<td>12.6%</td>
<td>2.8%</td>
</tr>
<tr>
<td>Combine Situations A and B</td>
<td>26.0%</td>
<td>14.4%</td>
<td>3.3%</td>
</tr>
<tr>
<td>Combine Situations A, B, and C</td>
<td>50.9%</td>
<td>30.9%</td>
<td>8.4%</td>
</tr>
<tr>
<td>Combine Situations A, B, C, and D</td>
<td>81.5%</td>
<td>60.7%</td>
<td>21.5%</td>
</tr>
</tbody>
</table>

Note: The table reports the percentage of 15,000 simulated samples in which at least one of a set of analyses was significant. Observations were drawn independently from a normal distribution. Baseline is a two-condition design with 20 observations per cell. Results for Situation A were obtained by conducting three t-tests, one on each of two dependent variables and a third on the average of these two variables. Results for Situation B were obtained by conducting one t-test after collecting 20 observations per cell. Results for Situation C were obtained by conducting a t-test, an analysis of covariance with a gender main effect, and an analysis of covariance with a gender interaction (each observation was assigned a 50% probability of being female). We report a significant effect if the effect of condition was significant in any of these analyses or if the Gender x Condition interaction was significant. Results for Situation D were obtained by conducting t-tests for each of the three possible pairings of conditions and an ordinary least squares regression for the linear trend of all three conditions (coding: low = −1, medium = 0, high = 1).
O’Boyle et al. studied the Chrysalis Effect
- Found 142 management dissertations that were subsequently published
- Compared dissertations to the resulting publications

Prior to ClinicalTrials.gov, 57% of large RCTs funded by National Heart Lung, and Blood Institute (NHLBI) were significant
- After ClinicalTrials.gov, 8% of these trials were significant
- Not due to different features of methods (e.g., active controls) or due to industry-sponsored research
Why pre-register?

- Accurately records your hypotheses and intentions before you have to communicate what your hypotheses and intentions were
  - After you see results, people overestimate how likely they would be to have predicted those results a priori
  - Garden of forking paths
  - p-values change as researchers intentions change, even if the data do not change
  - You need ammo for fighting against editors and reviewers

Also, preregistration increases evidentiary value of results

NON-PRE-REGISTERED STUDY

THEORY > (IV IS VALID) > (DV IS VALID) > (CONDITIONS WERE REALIZED) > (SAMPLING PLAN WAS NOT DATA DEPENDENT) > (HYPOTHESES WERE NOT DATA DEPENDENT) > (ETC.) > DATA

PRE-REGISTERED STUDY

THEORY > (IV IS VALID) > (DV IS VALID) > (CONDITIONS WERE REALIZED) > (SAMPLING PLAN WAS NOT DATA DEPENDENT) > (HYPOTHESES WERE NOT DATA DEPENDENT) > (ETC.) > DATA
How to pre-register a study

• I have used 3 methods:
  • A simple word document with your research strategy
  • Use OSF templates
  • AsPredicted.org

If you have completed an IRB, then you are nearly done with your preregistration

<table>
<thead>
<tr>
<th>AsPredicted.org</th>
<th>NIU IRB Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What's the main question being asked of hypothesis being tested in this study?</td>
<td>Describe the purpose of your study and the reason(s) this study is needed. Include any necessary background information and a description of your hypotheses and your research question.</td>
</tr>
<tr>
<td>2. Describe the key dependent variable(s) specifying how they will be measured.</td>
<td>Describe the data collection procedures including what data will be collected, how it will be collected (include a description of any interventions to be used), the duration of participants in the study session(s), and how session(s) will end.</td>
</tr>
<tr>
<td>3. How many conditions and which conditions will participants be assigned to?</td>
<td></td>
</tr>
<tr>
<td>4. Specify exactly which analyses you will conduct to examine the main question/hypothesis.</td>
<td></td>
</tr>
<tr>
<td>5. Any secondary analyses?</td>
<td></td>
</tr>
<tr>
<td>6. How many observations will be collected or what will determine the sample size?</td>
<td>Target number of participants for the entire study (including controls) from start to finish.</td>
</tr>
<tr>
<td>7. Anything else you would like to pre-register?</td>
<td></td>
</tr>
<tr>
<td>8. Have data been collected for this study already?</td>
<td>Typically “no” for new IRB applications</td>
</tr>
</tbody>
</table>
Preregistration does not limit the analyses you do.

Preregistration allows you to communicate which analyses were a priori and which were not.

Q: Am I unethical if I don’t preregister my studies?

A: No. It just means that you are not maximizing the potential evidentiary value of your data.
Q: Won’t preregistration add a lot of work?

A: No. You need to write your hypotheses and analysis plan anyways; this just changes when you do it.

Communicate your adherence to good research practices (and reap the rewards)
Why?

- Sets social norms for good research practices.
- You are being compared to others. Get credit where credit is due.
- Jobs are looking for this stuff now.
- T&P committees are looking for this stuff now.
- Publish like it is 2016 (i.e., publish as if the internet actually exists).

21-word statement

“We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.”

Submit to journals that value your good habits

Polish your application materials to brag up your good habits

• **SOCIAL PSYCHOLOGY**: The successful applicants will be joining the new *Social, Social Development, and Personality Psychology Cluster*, which includes research in social psychology, social development, and personality and individual differences. The positions are in social psychology but candidates whose research interests cross social psychology, social development, and personality are also encouraged to apply. For the positions in Social Psychology, applicants must demonstrate evidence of an active program of research that utilizes innovative approaches to social behaviour that will attract external funding. Applicants are also invited to describe their approach to open science, interdisciplinarity, new methods and analytic procedures in social research, and teaching interests.
Use OSF to Include links to your materials

Use OSF to Include links to supplementary analyses
Registered Replication Report: Strack, Martin, & Stepper (1988)

E.-J. Wagenmakers, Titia Beek, Laura Dijkhoff, and Quentin F. Gronau
Department of Psychological Methods, University of Amsterdam, The Netherlands


Protocol vetted by: Ursula Hess
Protocol edited by: Daniel J. Simons


Data and registered protocols: https://osf.io/plkd65/
Journals have special issues, publication options, or exclusively accept registered reports

- E.g., *JESP* special issue of registered reports
- E.g., *CRISP* only accepts registered reports
- Etc.
An example of an “Open Science” project

- Analyses preregistered
- Data/materials archived in online repository
- Possibly make data/stimuli freely available
- Use 21-word statement in article
- Publish in journal that rewards your practices
- Include links to prereg, stimuli, materials, supplementary analyses, etc.
- Brag about it to your future employers/T & P committee/funding agencies

If you are unsure about what to do, use the Taco Bell heuristic!!!
As a peer reviewer, Encourage authors to be transparent

• "I request that the authors add a statement to the paper confirming whether, for all experiments, they have reported all measures, conditions, data exclusions, and how they determined their sample sizes. The authors should, of course, add any additional text to ensure the statement is accurate. This is the standard reviewer disclosure request endorsed by the Center for Open Science (see http://osf.io/project/hadz3). I include it in every review."

Making tools and incentives for sharing data and materials
Why most of psychology is statistically unfalsifiable

Richard D. Morey
Cardiff University
Daniel Lakens
Eindhoven University of Technology

Abstract

Low power is a recurring problem in the psychological literature, and the evidence showing in the current manuscript is consistent with it. One cannot do strict replication analyses or use the secondary science that would be possible with replication results published in the literature, because the sample sizes were small. The consequence is that the literature is not a friend to theoretical progress, but rather its mortal enemy.

Combined with publication bias (Sterling, 1959; see also Lakens, 2015) and flexible analysis methods (Gelman and Loken, 2014), it is clear that the fallacies explain psychology's current problems with power. Most studies have low power, and hence statistical tests cannot differentiate even large effects from noise. Publication bias leads to large published effects by necessity, because at the small sample sizes common in the literature these are the only ones that are statistically significant. Researchers power to these large effect sizes, perpetuating the problem and creating a noise-laden literature. The consequence is that the literature is not a friend to theoretical progress, but rather its mortal enemy.
Figure 3. P-curves for *Journal of Personality and Social Psychology (JPSP)* studies suspected to have been *p*-hacked (A) and not *p*-hacked (B). Graphs depict p-curves observed in two separate sets of 20 studies. The first set (A) consists of 20 *JPSP* studies that only report statistical results from an experiment with random assignment, controlling for a covariate; we suspected this indicated *p*-hacking. The second set (B) consists of 20 *JPSP* studies reported in articles whose full text does not include keywords that we suspected could indicate *p*-hacking (e.g., *exclude*, *covariate*).
Figure 3. *P*-curves for Journal of Personality and Social Psychology (JPSP) studies suspected to have been *p*-hacked (A) and not *p*-hacked (B). Graphs depict *p*-curves observed in two separate sets of 20 studies. The first set (A) consists of 20 JPSP studies that only report statistical results from an experiment with random assignment, controlling for a covariate; we suspected this indicated *p*-hacking. The second set (B) consists of 20 JPSP studies reported in articles whose full text does not include keywords that we suspected could indicate *p*-hacking (e.g., exclude, covariate).
My position on “Power Poses”

Regarding: Carney, Cuddy & Yap (2010).

Reasonable people, whom I respect, may disagree. However, since early 2015 the evidence has been mounting suggesting there is unlikely any embodied effect of nonverbal expansiveness (vs. contractiveness)—i.e., “power poses” —on internal or psychological outcomes.

As evidence has come in over these past 2+ years, my views have updated to reflect the evidence. As such, I do not believe that “power pose” effects are real.

Any work done in my lab on the embodied effects of power poses was conducted long ago (while still at Columbia University from 2008-2011) well before my views updated. And so while it may seem I continue to study the phenomenon, those papers (emerging in 2014 and 2015) were already published or were on the cusp of publication as the evidence against power poses began to convince me that power poses weren’t real. My lab is conducting no research on the embodied effects of power poses.

The “review and summary paper” published in 2015 (in response to Ranehill, Dreber, Johannesson, Leiberg, Sul, & Weber (2015) seemed reasonable, at the time, since there were a number of effects showing positive evidence and only 1 published that I was aware of showing no evidence. What I regret about writing that “summary” paper is that it
Heat Priming-Aggressive Cognitions

Statistical Inference Results
Studies contain evidential value (right-skewed)
$\chi^2(4)=0.56, p=0.96782$
Studies lack evidential value (flatter than 33% power)
$\chi^2(4)=10.96, p=0.0273$
Studies lack evidential value and were intensely p-hacked (left-skewed)
$\chi^2(4)=8.16, p=0.0857$

$p$-curve results

<table>
<thead>
<tr>
<th>Percent of p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>0% 0% 0% 7% 100%</td>
</tr>
<tr>
<td>p-value</td>
</tr>
<tr>
<td>0.01 0.02 0.03 0.04 0.05</td>
</tr>
</tbody>
</table>

$p$-curve Elderly Priming

- uniform
- 33% power
- observed
An Ethical Approach to Peeking at Data

Brad J. Sagarin, James K. Ambler, and Ellen M. Lee
Northern Illinois University

Abstract
When data analyses produce unanticipated results, researchers may be tempted to peek at the data, which may transform a study from an a priori design into an exploratory analysis. This may affect the Type I error rate. To address this, Frick, Pocock, and colleagues (2017) proposed a set of guidelines for data peering. However, this approach may not be adequate to address the potential for data peering. In such situations, many researchers (at least 55.9%) according to the results of John et al. (2012) might make a post-hoc decision to augment the dataset—thus engaging in a "questionable research practice" according to John et al. We believe, however, that it is not the dataset augmentation but the magnitude of the resulting Type I error rate that should be considered.

Table 2. Techniques Available at Each Stage of Dataset Augmentation

<table>
<thead>
<tr>
<th>Stage</th>
<th>Dataset augmentation</th>
<th>Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1: Before data analysis begins</td>
<td>a priori</td>
<td>Botella, Ximénez, Revuelta, and Suero (2006)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Frick (2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pocock (1977)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>adjusted $P_{aug}$ (see also Baguley, 2012; Lakens &amp; Evers, 2014)</td>
</tr>
<tr>
<td>Stage 2: After discovering non-significant results</td>
<td>contemplating post hoc</td>
<td>conditional power analysis (Lakens &amp; Evers, 2014)</td>
</tr>
<tr>
<td>Stage 3: After unplanned dataset augmentation</td>
<td>post hoc</td>
<td>$P_{aug}$</td>
</tr>
</tbody>
</table>