Open science: Replicability, transparency, p-hacking

David Saldaña
@DavidSaldana_es
@DavidSaldana_en
Trondheim, 28th August 2019
What should we be aiming for?

Valid and useful knowledge that is available to those who may need/want to use it
19th century scientist

I must find the explanation for this phenomenon in order to truly understand Nature...

21st century scientist

I must get the result that fits my narrative so I can get my paper into Nature..

https://pandelisperakakis.info/wp-content/uploads/2017/05/scientist_vs_academic.png
Threats to “truthfulness” of our results
Valid...means as “true” as possible
Some concepts

REPRODUCIBILITY:

Refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator.

• E.g., a researcher uses the same raw data, builds same analysis files, and same statistical procedures to make sure that same results obtained as in published study.

• Differences could be due to:
  o Processing (e.g., treatment missing data) of data
  o Application of statistical method (e.g., different defaults)
  o Accidental errors in original analysis (or follow-up analysis)
  o Reproducibility is a minimum necessary condition for a finding to be believable and informative.

Report of the Subcommittee on Replicability in Science of the SBE Advisory Committee to the National Science Foundation
13 May 2015 Presentation at SBE AC Spring Meeting by K. Bollen
Some concepts

REPRODUCIBILITY:

Refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator

• E.g., a researcher uses the same raw data, builds same analysis files, and same statistical procedures to make sure that same results obtained as in published study

• Differences could be due to:
  o Processing (e.g., treatment missing data) of data
  o Application of statistical method (e.g., different defaults)
  o Accidental errors in original analysis (or follow-up analysis)
  o Reproducibility is a minimum necessary condition for a finding to be believable and informative.
Some concepts

REPLICABILITY

Refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected

- a failure to replicate occurs when one study documents relations and a subsequent attempt with new data fails to yield the same relations

- null results or nonzero results could be replications
  - E.g., failure to find intervention to work in two different data sets is a replication as would be the finding of positive effect

- Same researcher performing second study more likely to replicate
  - Fully aware of procedures

- Second researcher in another location less likely because:
  - Did not directly observe the first study
  - Relies on text description of first study
  - Critical details not fully understood or described
  - Failure to replicate might be due to different procedures
Some concepts

GENERALIZABILITY

refers to whether the results of a study apply in other contexts or populations that differ from the original one

- degree to which found relations apply in different situations
- E.g., do findings based on college students apply to adult population of the United States?
- E.g., does an experiment that uses one type of persuasive message work when researcher tries other types of persuasive messages?
- Failure to generalize directs attention to operation of limiting conditions on relationship

- Chance to advance theory as these limiting conditions are uncovered
Replication study of the Open Science Collaboration

• 2008 articles of: Psychological Science (PSCI), Journal of Personality and Social Psychology (JPSP), and Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP: LMC)

• 100 replications by 270 contributing authors

Does replication show a $p < .05$ in the same direction?

Gilbert et al. (2016)

35 replication studies showed significant effects versus 98 in the original studies
Is the original effect size within 95\% CI of the replication?

It is in 48\% of the studies

Gilbert et al. (2016)
Looks like we have a false positive problem! Why?

We are only human:

• Apophenia (the tendency to see patterns in random data)
• Confirmation bias (the tendency to focus on evidence that is in line with our expectations or favoured explanation)
• Hindsight bias (the tendency to see an event as having been predictable only after it has occurred)

Why? Some flaws – e.g. difficulty in understanding how p-values work: base rate fallacy

Testing 100 drugs
Top 10 truly effective
Power of 80% -> I detect 8 of the 20 truly effective

$p = .05 \Rightarrow 5$ of the $90$ non-effective are false-positives
I "see" $5 + 8$ as "effective" = $13$
But I am getting $5/13 = 38\%$ of false discovery! (not $5\%$)

Reinhart, 2015
Difficulty in understanding how p-values work: e.g. multiple testing

Reinhart, 2015

Difficulty in understanding how p-values work: e.g. multiple testing

Reinhart, 2015
Difficulty in understanding how p-values work: e.g. multiple testing

Reinhart, 2015
Difficulty in understanding how p-values work: e.g. multiple testing

Reinhart, 2015
Difficulty in understanding how p-values work: e.g. multiple testing

Reinhart, 2015
Difficulty in understanding how p-values work: e.g. multiple testing

Reinhart, 2015
### 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 and another after collecting an additional 10 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 × 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).
Optional stopping

Fig. 1. Likelihood of obtaining a false-positive result when data collection ends upon obtaining significance (p ≤ .05, highlighted by the dotted line). The figure depicts likelihoods for two minimum sample sizes, as a function of the frequency with which significance tests are performed.
Difficulty in understanding how p-values work

https://www.statisticsdonewrong.com
Why many results in Psychology may be false

• p-hacking (fishing for significant results) or just misunderstanding
  • Using only p-values
  • Underpowered studies
  • Optional stopping
  • Selective removal of outliers
  • Selective reporting of results

• Publication bias of negative or complex findings

• HARKing: hypothesizing after results are known.

Zwaan, 2015
The researcher degrees of freedom: design

- Conducting explorative research without any hypothesis
- Studying a vague hypothesis that fails to specify the direction of the effect
- Creating multiple manipulated independent variables and conditions
- Measuring additional variables that can later be selected as covariates, independent variables, mediators, or moderators
- Measuring the same dependent variable in several alternative ways
- Measuring additional constructs that could potentially act as primary outcomes
- Measuring additional variables that enable later exclusion of participants from the analyses (e.g., awareness or manipulation checks)
- Failing to conduct a well-founded power analysis
- Failing to specify the sampling plan and allowing for running (multiple) small studies

Wicherts et al. (2016)
The researcher degrees of freedom: data collection

- Failing to randomly assign participants to conditions
- Insufficient blinding of participants and/or experimenters
- Correcting, coding, or discarding data during data collection in a non-blinded manner
- Determining the data collection stopping rule on the basis of desired results or intermediate significance testing

Wicherts et al. (2016)
The researcher degrees of freedom: analysis

• Choosing between different options of dealing with incomplete or missing data on ad hoc grounds

• Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, motion correction) in an ad hoc manner

• Deciding how to deal with violations of statistical assumptions in an ad hoc manner

• Deciding on how to deal with outliers in an ad hoc manner

• Selecting the dependent variable out of several alternative measures of the same construct

• Trying out different ways to score the chosen primary dependent variable

• Selecting another construct as the primary outcome

Wicherts et al. (2016)
The researcher degrees of freedom: analysis

- Selecting independent variables out of a set of manipulated independent variables
- Operationalizing manipulated independent variables in different ways (e.g., by discarding or combining levels of factors)
- Choosing to include different measured variables as covariates, independent variables, mediators, or moderators
- Operationalizing non-manipulated independent variables in different ways
- Using alternative inclusion and exclusion criteria got selecting participants in analyses
- Choosing between different statistical models
- Choosing the estimation method, software package, and computation of SEs
- Choosing inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing)

Wicherts et al. (2016)
The researcher degrees of freedom: reporting

• Failing to assure reproducibility (verifying the data collection and data analysis)
• Failing to enable replication (re-running of the study)
• Failing to mention, misrepresenting, or misidentifying the study preregistration
• Failing to report so-called “failed studies” that were originally deemed relevant to the research question
• Misreporting results and p-values
• Presenting exploratory analyses as confirmatory (HARKing)

Wicherts et al. (2016)
The researcher degrees of freedom issue

Exploratory research

Confirmatory research
“Small” mistakes that undermine our science

Munafò et al. (2017)
Simple Solution to the Problem of False-Positive Publications: Authors

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.
2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.
3. Authors must list all variables collected in a study.
4. Authors must report all experimental conditions, including failed manipulations.
5. If observations are eliminated, authors must also report what the statistical results are if those observations are included.
6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

Simmons, 2011
Be transparent

Table 3. Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

Using the same method as in Study 1, we asked 20 34 University of Pennsylvania undergraduates to listen only to either “When I’m Sixty-Four” by The Beatles or “Kalimba” or “Hot Potato” by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. Then, in an ostensibly unrelated task, they indicated only their birth date (mm/dd/yyyy) and how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with “computers are complicated machines;” their father’s age, their mother’s age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as “the good old days,” and their gender. We used father’s age to control for variation in baseline age across participants.

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to “When I’m Sixty-Four” (adjusted M = 20.1 years) rather than to “Kalimba” (adjusted M = 21.5 years), F(1, 17) = 4.92, p = .040. Without controlling for father’s age, the age difference was smaller and did not reach significance (Ms = 20.3 and 21.2, respectively), F(1, 18) = 1.01, p = .33.
More Solutions: data sharing
Some rules for data sharing

• Anticipate how your data will be used
• Keep raw data raw
• Store data in open formats
• Data structured for analysis (tidy data)
• Data should be uniquely identified
• Link relevant metadata
• Adopt proper privacy protocols
• Systematic backup scheme (2 onsite and 1 offsite)
• Analyse your capacity needs

Pre-registration
ClinicalTrials.gov is a database of privately and publicly funded clinical studies conducted around the world.

Explore 314,644 research studies in all 50 states and in 209 countries.

ClinicalTrials.gov is a resource provided by the U.S. National Library of Medicine.

IMPORTANT: Listing a study does not mean it has been evaluated by the U.S. Federal Government. Read our disclaimer for details.

Before participating in a study, talk to your health care provider and learn about the risks and potential benefits.
Welcome to PROSPERO
International prospective register of systematic reviews

Register a review
Registering a review is quick and easy. Just follow these simple steps to register your review in PROSPERO.

Search PROSPERO
Search for PROSPERO registrations by entering words in the record or the registration number below.

Register your review now

Accessing and completing the registration form
Registered reports

204 journals  https://cos.io/rr/
Availability

**Figure 3. Five Open Science schools of thought (Fecher and Friesike, 2014)**
*Figure 1. Promoting openness at different stages of the research process (Open Science and Research Initiative, 2014)*

https://www.fosteropenscience.eu
Open Access publishing

- **Green Route**
  - Often manuscript
  - Institutional Repository
  - Disciplinary Repository
  - All OA, No Fees

- **Gold Route**
  - Journal article
  - Hybrid (OA with fees and non OA)
  - All OA, Fees Mandatory

https://canterbury.libguides.com/scholarly/OA
Open Access publishing: the Green Route

https://canterbury.libguides.com/scholarly/OA
A potential workflow

Not yet accepted

• Pre-print for discussion on repository

Accepted

• Pre-print on repository
• Post-print on repository (potential embargo)
• Post-print on personal page
• Disseminate via social media (ResearchGate and similar)
Thanks!