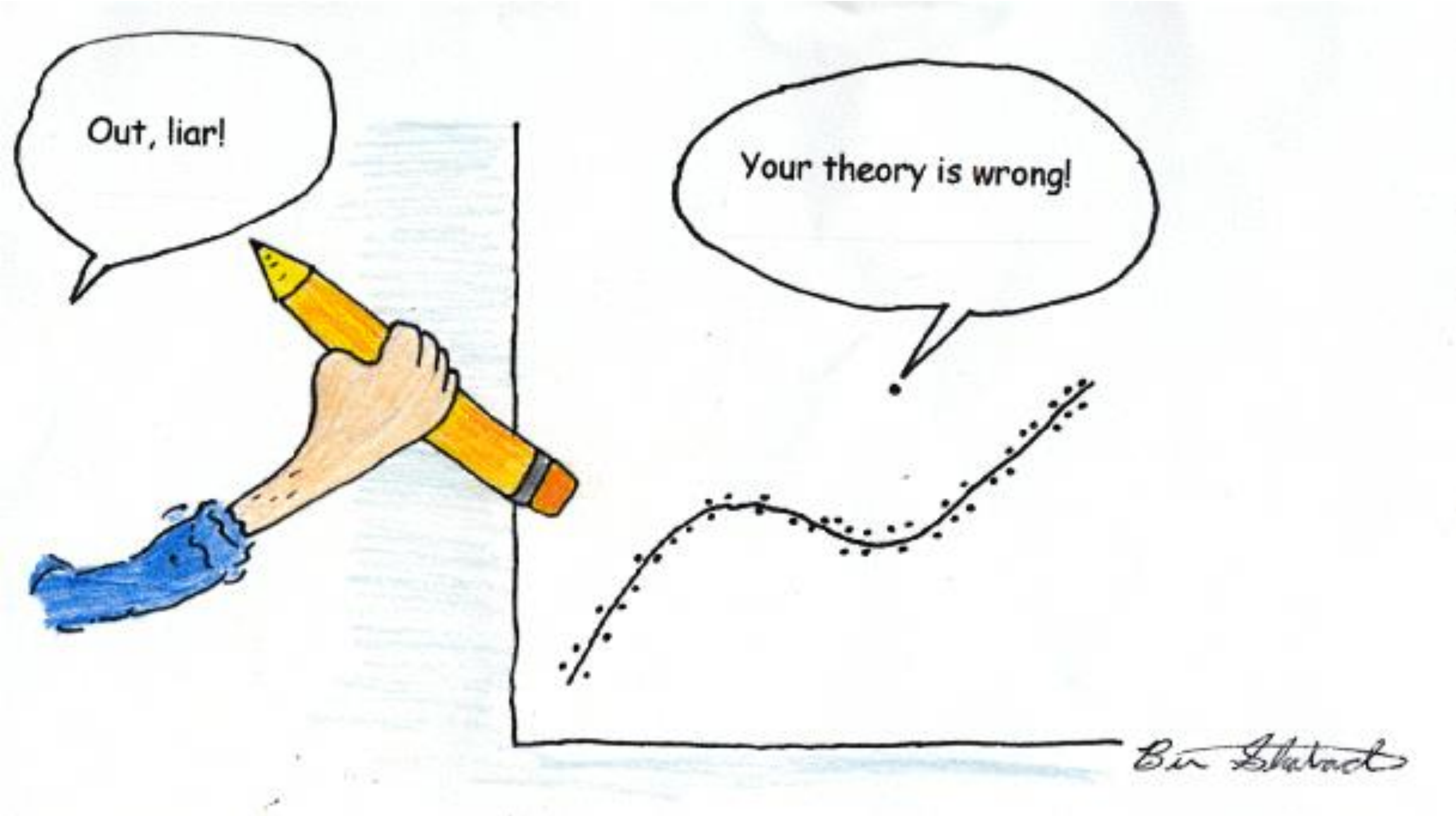
Research integrity in statistics: (mis)reporting and researcher degrees of freedom

Marjan Bakker; August 25, 2021; Amsterdam

TILBURG UNIVERSITY

Understanding Society

Choices



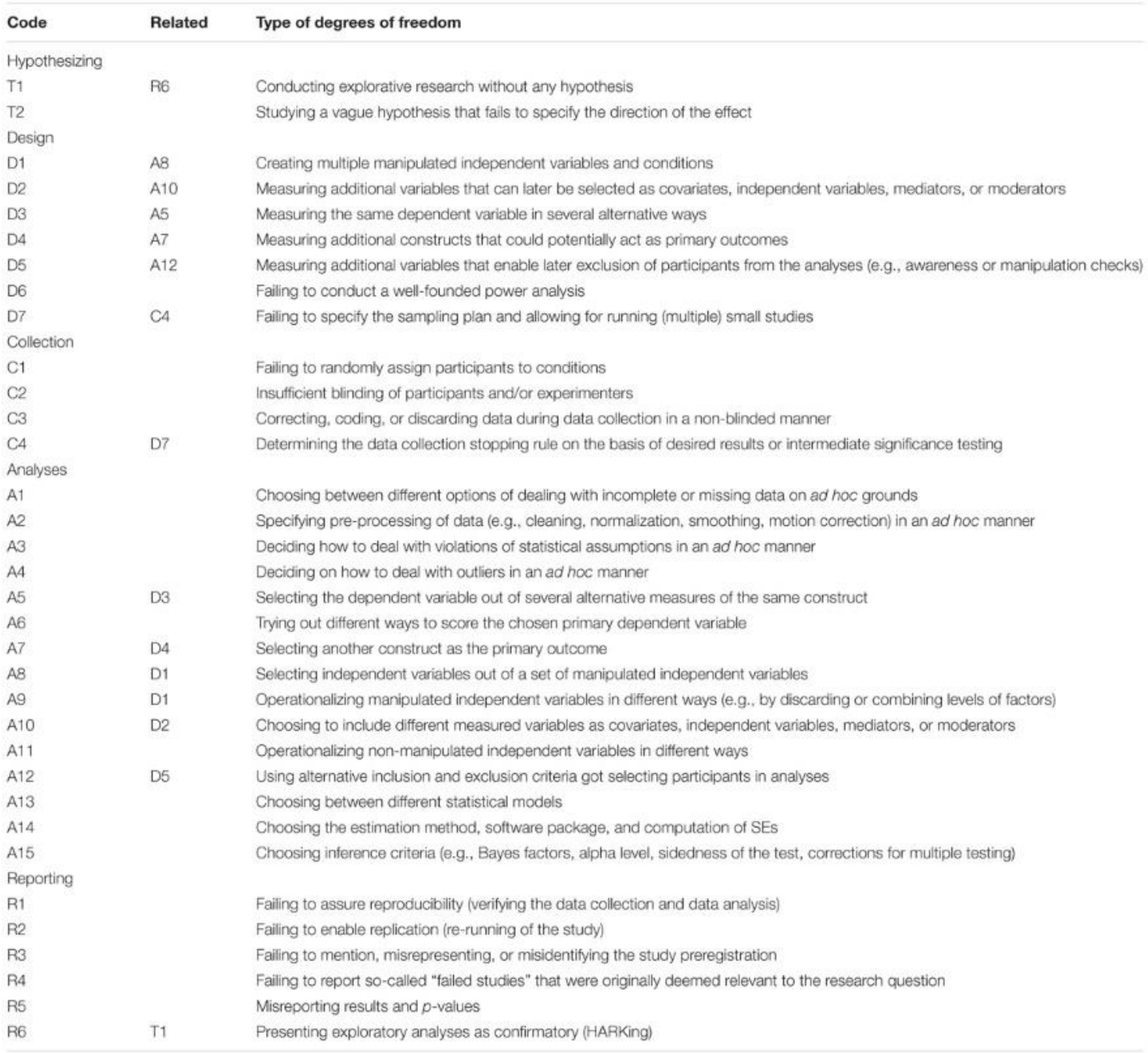
Choices

Remove?

Or don’t remove?

… or correct, or winsorize, or use a different statistical technique, or …

Researcher Degrees of Freedom



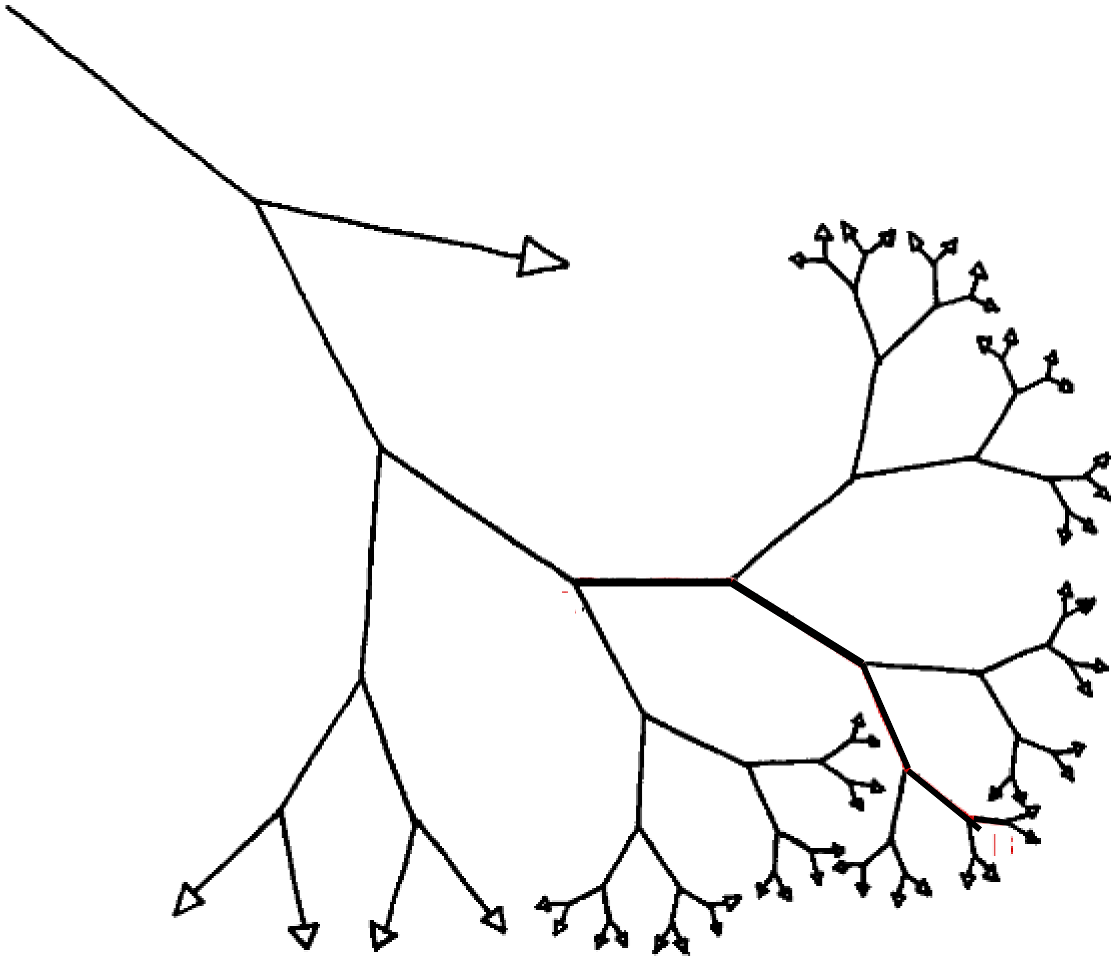
Code	Related	Type of degrees of freedom
Hypothesizing		
T1	R6	Conducting explorative research without any hypothesis
T2		Studying a vague hypothesis that fails to specify the direction of the effect
Design		
D1	A8	Creating multiple manipulated independent variables and conditions
D2	A10	Measuring additional variables that can later be selected as covariates, independent variables, mediators, or moderators
D3	A5	Measuring the same dependent variable in several alternative ways
D4	A7	Measuring additional constructs that could potentially act as primary outcomes
D5	A12	Measuring additional variables that enable later exclusion of participants from the analyses (e.g., awareness or manipulation checks)
D6		Failing to conduct a well-founded power analysis
D7	C4	Failing to specify the sampling plan and allowing for running (multiple) small studies
Collection		
C1		Failing to randomly assign participants to conditions
C2		Insufficient blinding of participants and/or experimenters
C3		Correcting, coding, or discarding data during data collection in a non-blinded manner
C4	D7	Determining the data collection stopping rule on the basis of desired results or intermediate significance testing
Analyses		
A1		Choosing between different options of dealing with incomplete or missing data on *ad hoc* grounds
A2		Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, motion correction) in an *ad hoc* manner
A3		Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner
A4		Deciding on how to deal with outliers in an *ad hoc* manner
A5	D3	Selecting the dependent variable out of several alternative measures of the same construct
A6		Trying out different ways to score the chosen primary dependent variable
A7	D4	Selecting another construct as the primary outcome
A8	D1	Selecting independent variables out of a set of manipulated independent variables
A9	D1	Operationalizing manipulated independent variables in different ways (e.g., by discarding or combining levels of factors)
A10	D2	Choosing to include different measured variables as covariates, independent variables, mediators, or moderators
A11		Operationalizing non-manipulated independent variables in different ways
A12	D5	Using alternative inclusion and exclusion criteria got selecting participants in analyses
A13		Choosing between different statistical models
A14		Choosing the estimation method, software package, and computation of SEs
A15		Choosing inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing)
Reporting		
R1		Failing to assure reproducibility (verifying the data collection and data analysis)
R2		Failing to enable replication (re-running of the study)
R3		Failing to mention, misrepresenting, or misidentifying the study preregistration
R4		Failing to report so-called "failed studies" that were originally deemed relevant to the research question
R5		Misreporting results and *p*-values
R6	T1	Presenting exploratory analyses as confirmatory (HARKing)

Wicherts et al. (2016)

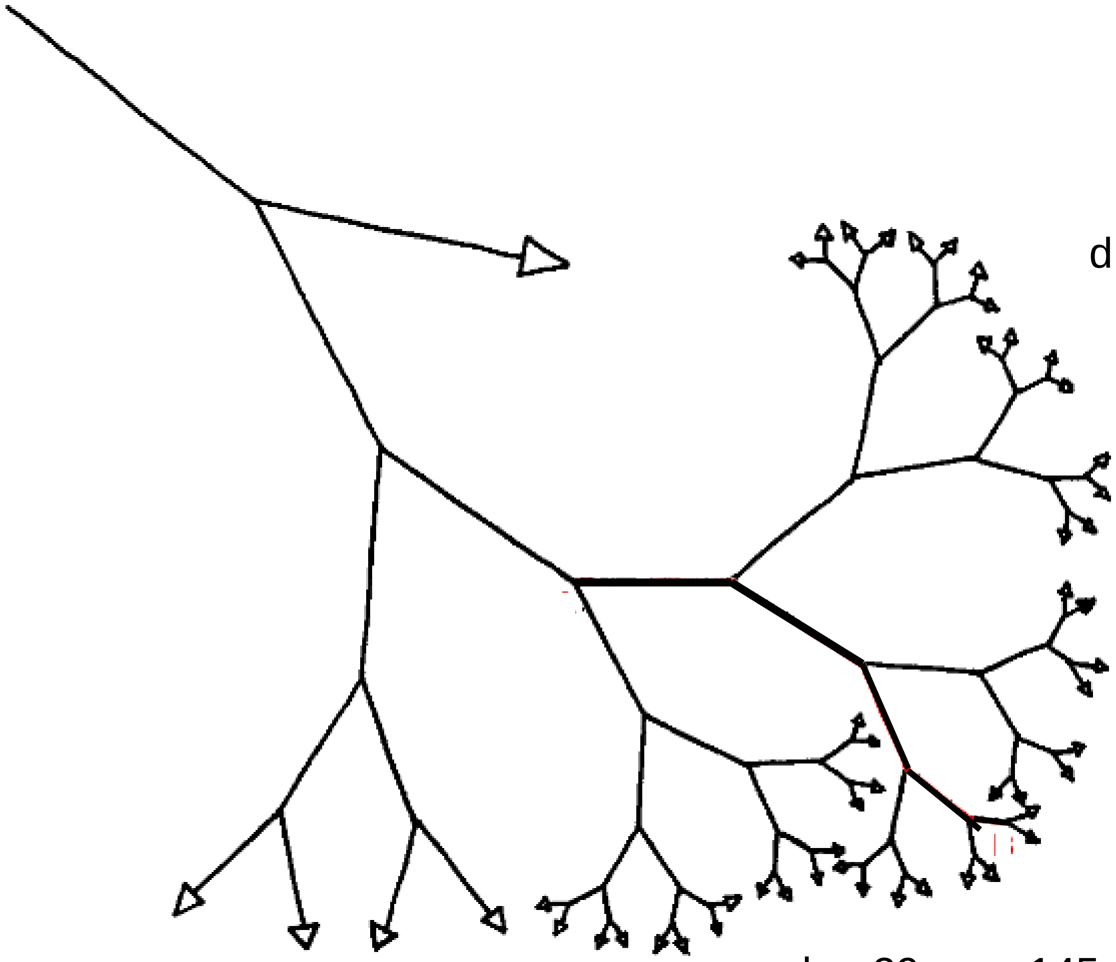
Researcher Degrees of Freedom

- 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
- 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)

Garden of forking paths



Garden of forking paths



d = .12, p = .702

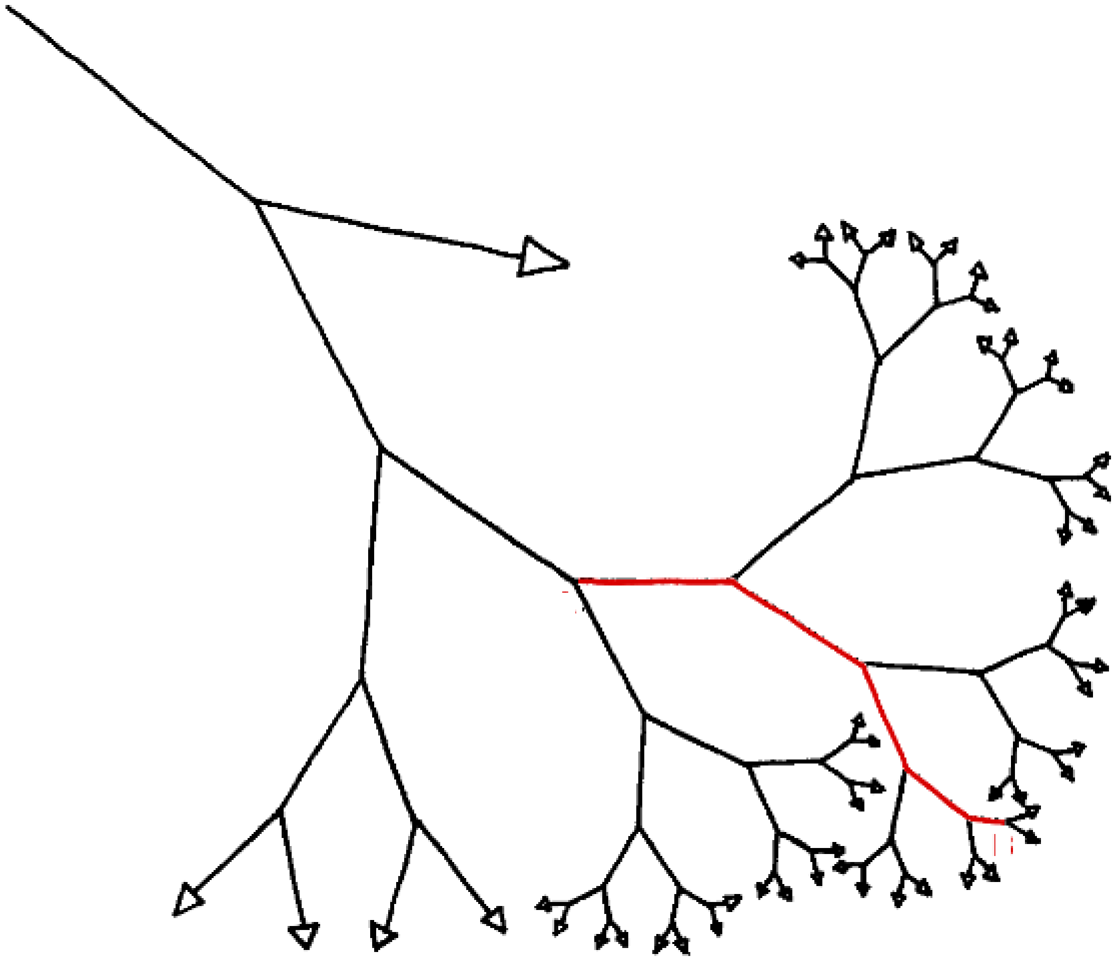
d = .39, p = .053

d = .62, p = .002

d = .29, p = .145

d = .13, p = .516

Garden of forking paths



d = .62, p = .002

Questionable Research Practices

John et al. (2012)

I have at least once….	(self admittance rate)
Failing to report all of a study’s dependent measures	(63.4%)
Deciding whether to collect more data after looking to see whether the results were significant	(55.9%)
Failing to report all of a study’s conditions	(27.7%)
Stopping collecting data if the result is already significant	(15.6%)
‘Rounding off’ a p value (e.g. p = .054, report p < .05)	(22.0%)
Selectively reporting studies that ‘worked’	(45.8%)
Deciding whether to exclude data after looking at the impact of doing so	(38.2%)
Reporting an unexpected finding as having been predicted from the start	(27.0%)

Simmons, Nelson, and Simonsohn (2011)

Listening to The Beatles makes you younger!

Increase in Type I error rate

Type I error: incorrect rejection of a true null hypothesis.

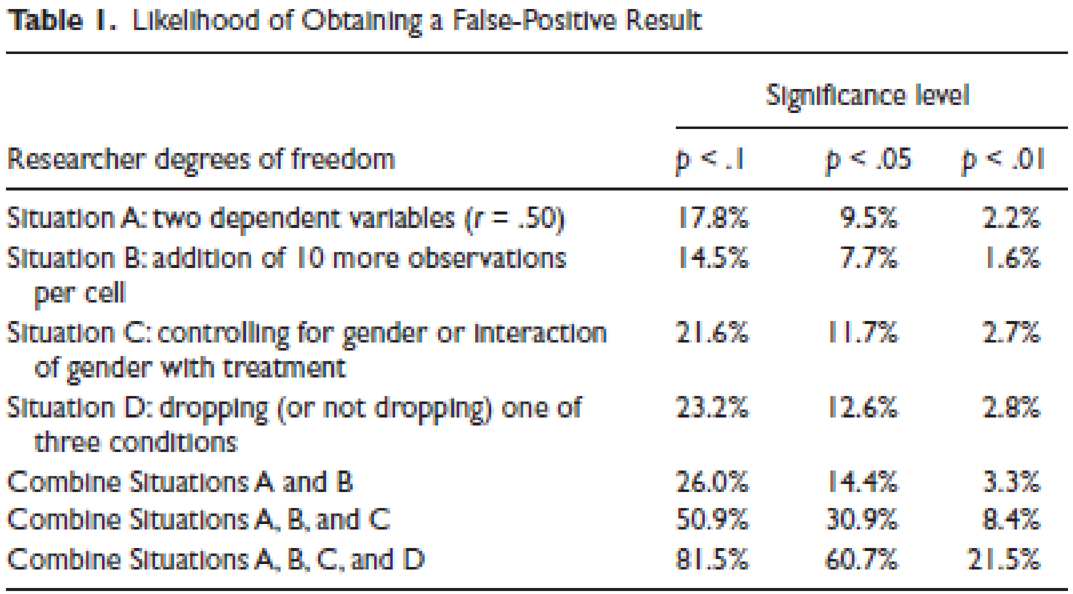
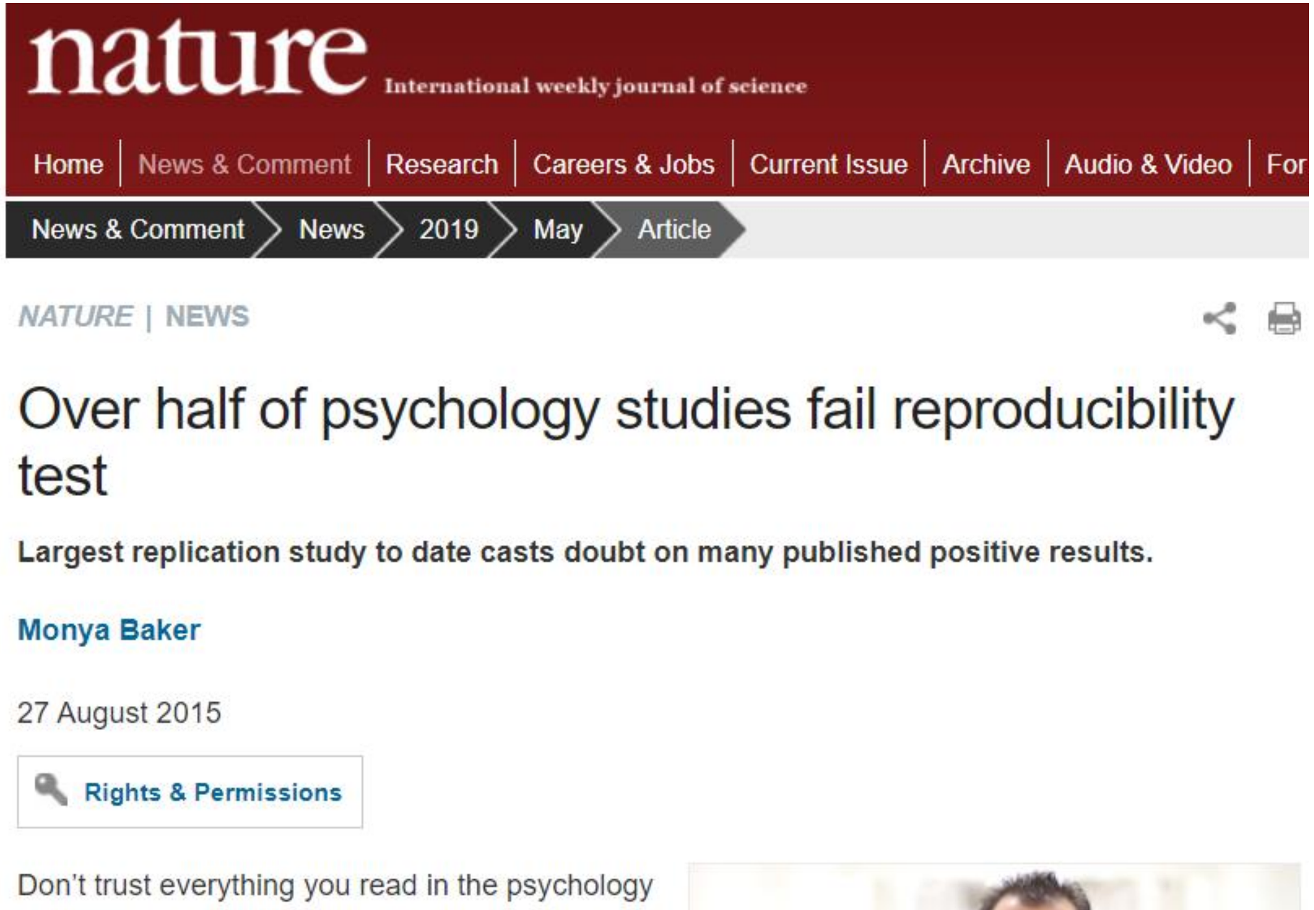


Table 1. Likelihood of Obtaining a False-Positive Result

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

Many published null results

- Too many positive findings
- Failure to replicate


nature
International weekly journal of science
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For
News & Comment > News > 2019 > May > Article
Rights & Permissions

NATURE | NEWS

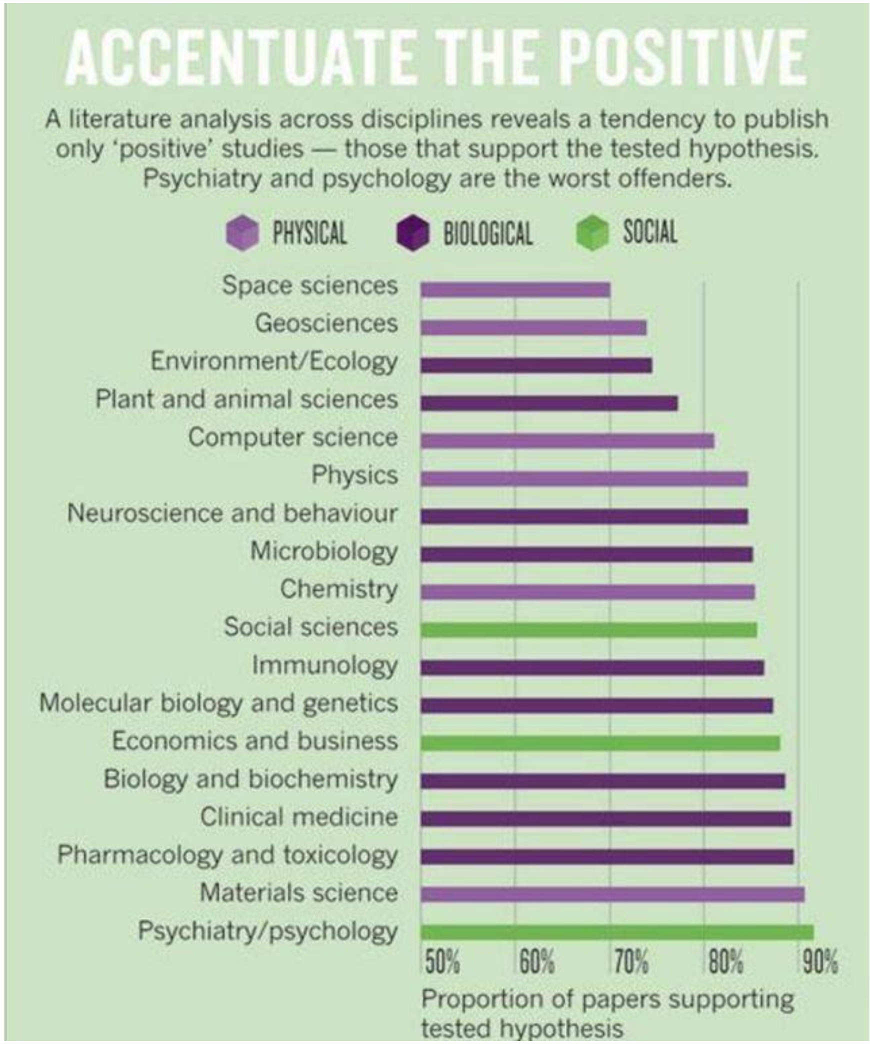
Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

Don't trust everything you read in the psychology



ACCENTUATE THE POSITIVE

A literature analysis across disciplines reveals a tendency to publish only 'positive' studies — those that support the tested hypothesis. Psychiatry and psychology are the worst offenders.

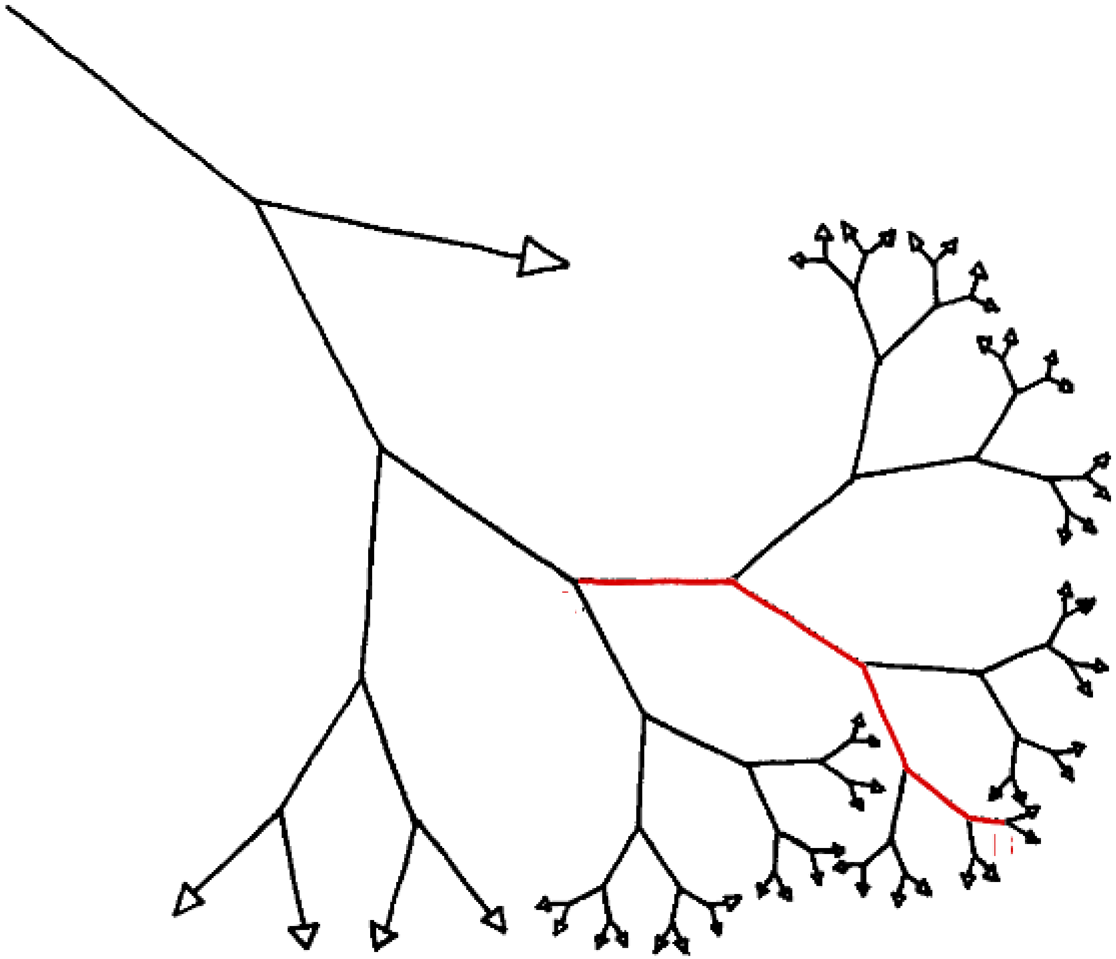
Fanelli,2010,

Solutions

- Preregistration: specifying your research plan in advance of your study and submitting it to a registry

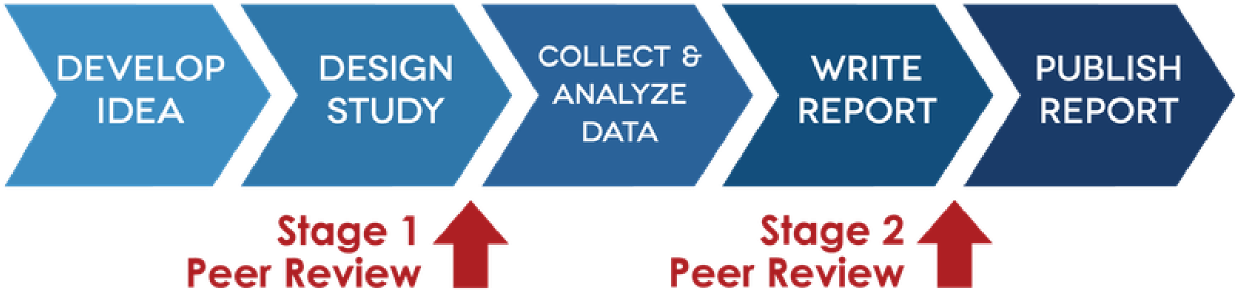
- Clear distinction between two modes of research:
  - Confirmatory testing (data is collected to test predictions)
    - Prediction
  - Exploratory analysis (data is used to generate predictions that could be tested in the future)
    - Postdiction

Register your choices



Registered Reports

- Registered reports
  - Submit pre-registration to journal for review: introduction and method section
  - Receive ‘in principle acceptance’
  - Submit paper: results and discussion reviewed for correspondence with original introduction and method
  - Benefits:
    - No incentive for significant results
    - Reviewers can contribute to improving methods



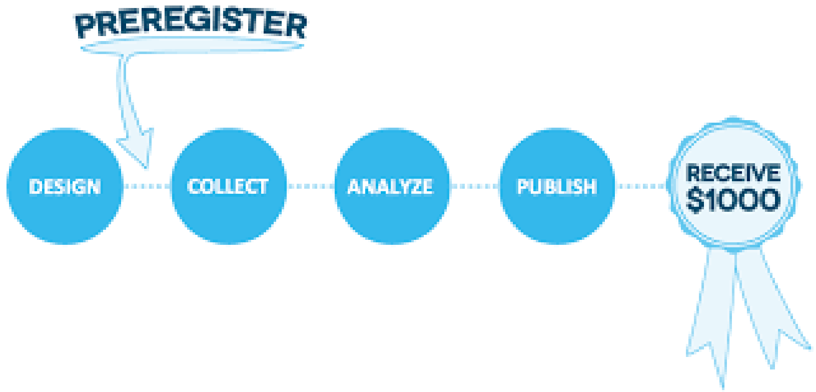
Different formats

- Overview on: https://osf.io/zab38/wiki/home/

• OSF prereg	Most extensive template
• As predicted	Only 8 questions
• Open ended	Snapshot of current project with time stamp
• Replication recipe	For replication studies
• Qualitative research	Haven & Van Grootel, …
• Secondary Data	Van den Akker et al. (2019)
• Cognitive Modeling	Cruwell & Evans (2019)
• fMRI	Flannery (2018)

From theory to practice

- Preregistration
  - The number of preregistrations at OSF has approximately doubled yearly with 38 in 2012 to 36,675 by the end of 2019
  - Preregistration Challenge
  - Preregistration badges
    - 75 journals award badges
- Registered reports
  - Over 300 journals offer this format



PREREGISTRATION CHALLENGE

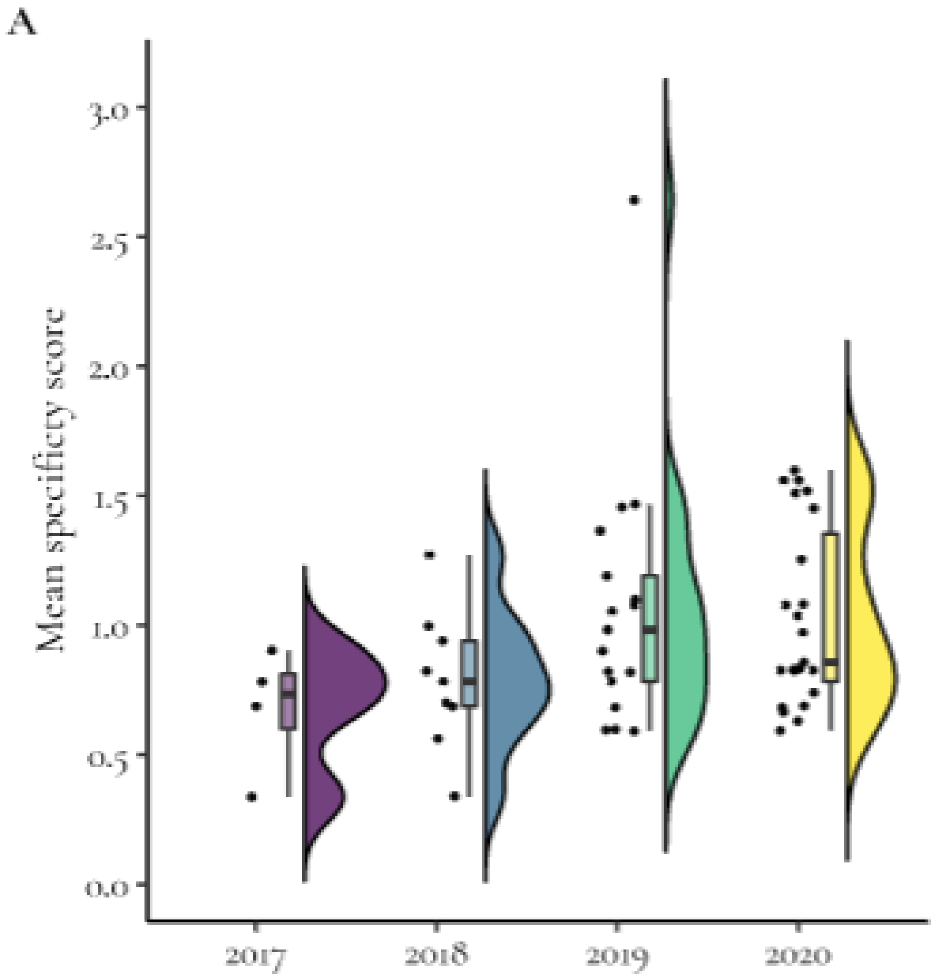
… and to Research

- Do preregistered studies prevent the opportunistic use of researcher degrees of freedom?
  - Comparison of Prereg Challenge Registrations (extensive guidelines) with Standard Pre-Data Collection Registrations (almost no guidelines)
  - Are they specific, precise, and exhaustive

- Results:
  - Prereg Challenge Registrations prevent more opportunistic use of researcher degrees of freedom.
  - However, still room for the opportunistic use of researcher degrees of freedom.
  - For example: often number of hypotheses was not clear.

Bakker et al. 2020

… and to Research



Heirene et al. (2021)

TILBURG UNIVERSITY

Research: adherence to preregistered plans

Claesen, A., Gomes, S. L. B. T., Tuerlinckx, F., & vanpaemel, w. (2019, May 9). Preregistration: Comparing Dream to Reality. https://doi.org/10.31234/osf.io/d8wex

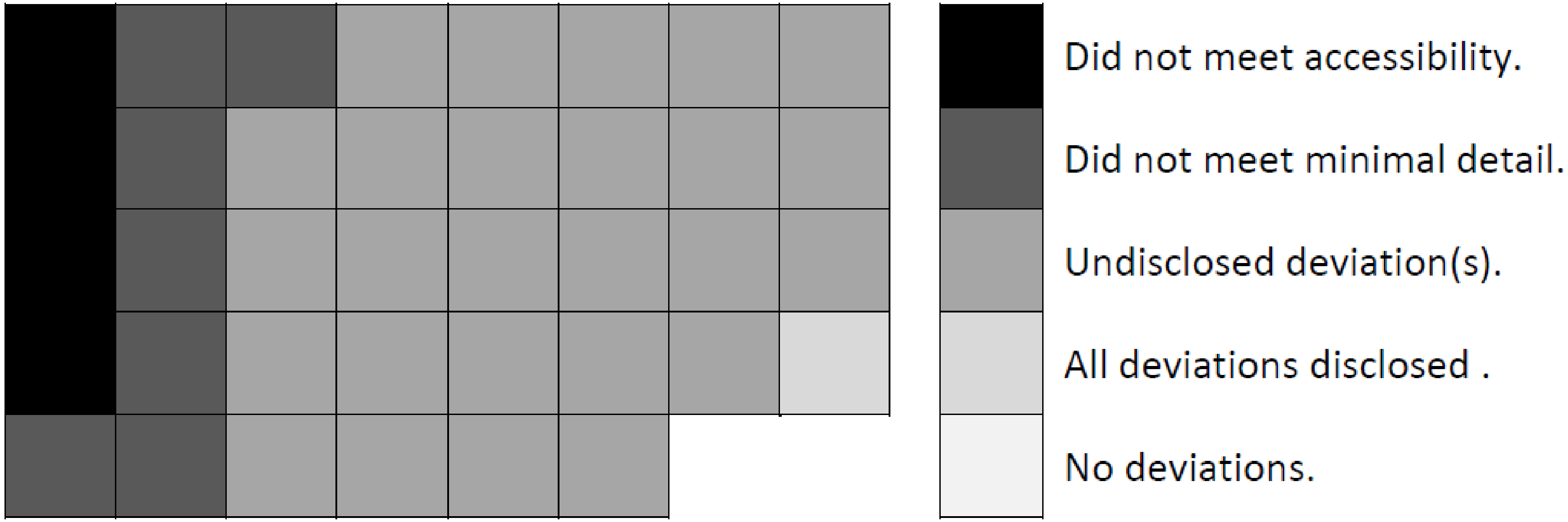
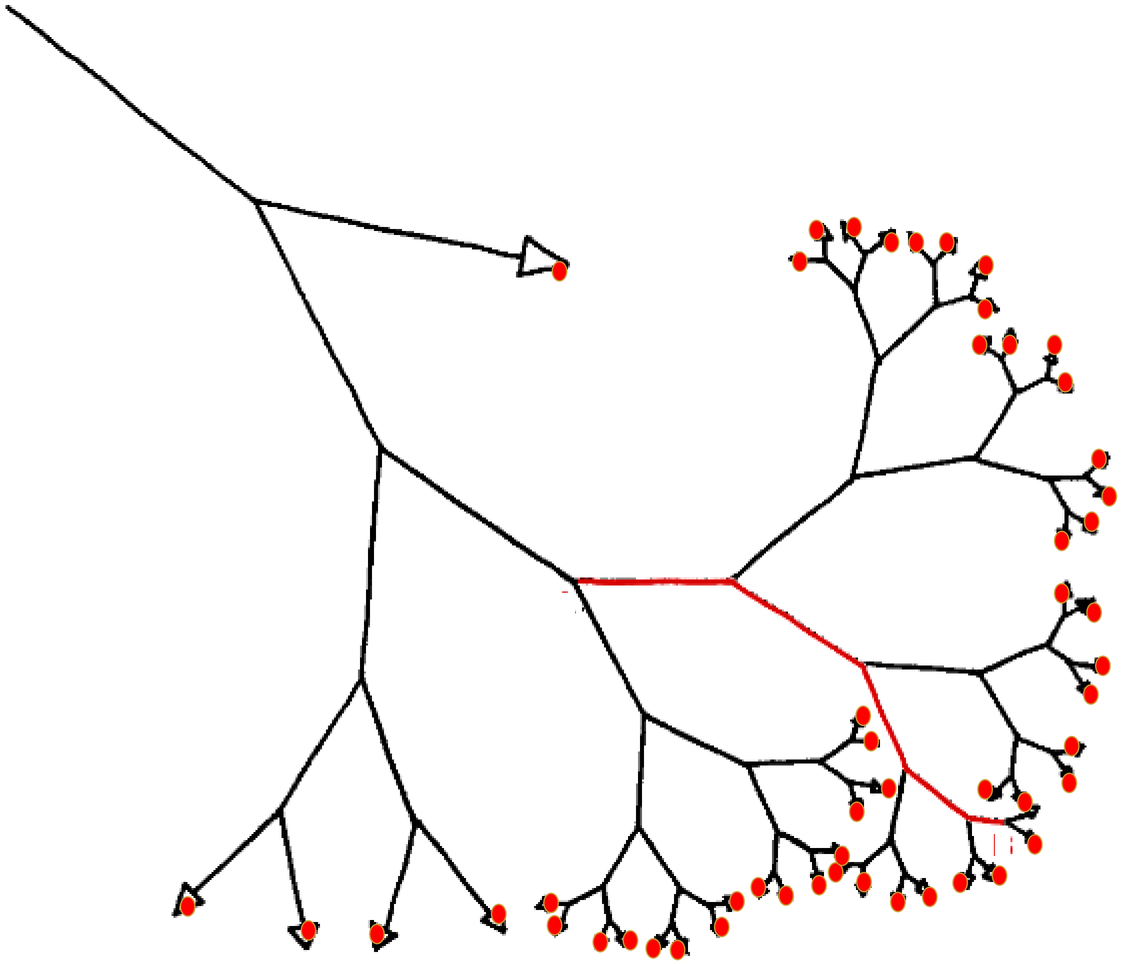


Figure 1. Assessment on preregistration level. Each cell represents one preregistration plan. None of the plans was adhered to without deviations.

Solutions

- Preregistration: specifying your research plan in advance of your study and submitting it to a registry
- **Multiverse analysis: check all paths**

… check all paths!



Multiverse analysis

- Sensitivity analysis
  - Only a few choices are tested independently
  - E.g., with and without outlier removal
- Specification Curve (Simonsohn, Simmons, & Nelson, 2019)
  - Focus on graphical display of results
- Multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016)

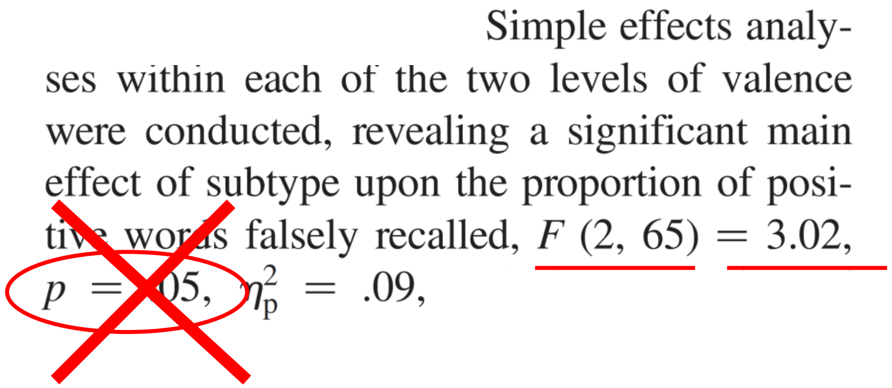
Solutions

- Preregistration: specifying your research plan in advance of your study and submitting it to a registry
- Multiverse analysis: check all paths
- **Be transparent about all the paths you went on**
    - Open lab notebooks

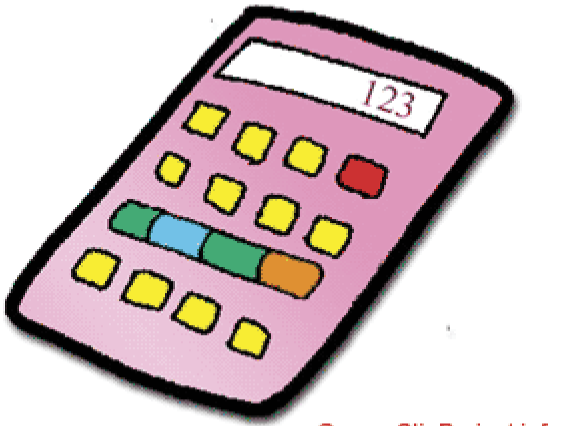
Errors

- Humans make errors

Reporting of Statistical Results



Simple effects analyses within each of the two levels of valence were conducted, revealing a significant main effect of subtype upon the proportion of positive words falsely recalled, F (2, 65) = 3.02, ~~p = .05~~, η^2_p = .09,



© www.ClipProject.info

***p* = .06**

Occurrence of errors

- Half of the papers showed an error
- 1 in 8 showed a gross error (an error that affected the statistical conclusion

(Bakker & Wicherts, 2011)

Reporting Errors in Other Fields

- Garcia-Berthou & Alcaraz (2004)
  - Nature and Britisch Medical Journal
  - 38% and 25% of the articles contained at least one error.

- Berle and Starcevic (2007)
  - Two psychiatry journals
  - 36% of the articles contained at least one error

Reporting Errors

statcheck

(Epskamp & Nuijten, 2014)

Inconsistency
Gross Inconsistency
Inconsistency
Inconsistency
Inconsistency

- Half of the papers in psychology contain at least one inconsistent *p*-value
- In 1 in 8 papers, this may have affected the conclusion

Reported p < .05 and computed p > .05, or vice versa

(Nuijten et al., 2016)

Questionable Research Practices

John et al. (2012)

I have at least once….	(self admittance rate)
• Failing to report all of a study's dependent measures	(63.4%)
• Deciding whether to collect more data after looking to see whether the results were significant	(55.9%)
• Failing to report all of a study's conditions	(27.7%)
• Stopping collecting data if the result is already significant	(15.6%)
• **'Rounding off' a p value (e.g. p = .054, report p < .05)**	**(22.0%)**
• Selectively reporting studies that 'worked'	(45.8%)
• Deciding whether to exclude data after looking at the impact of doing so	(38.2%)
• Reporting an unexpected finding as having been predicted from the start	(27.0%)

Tilburg University

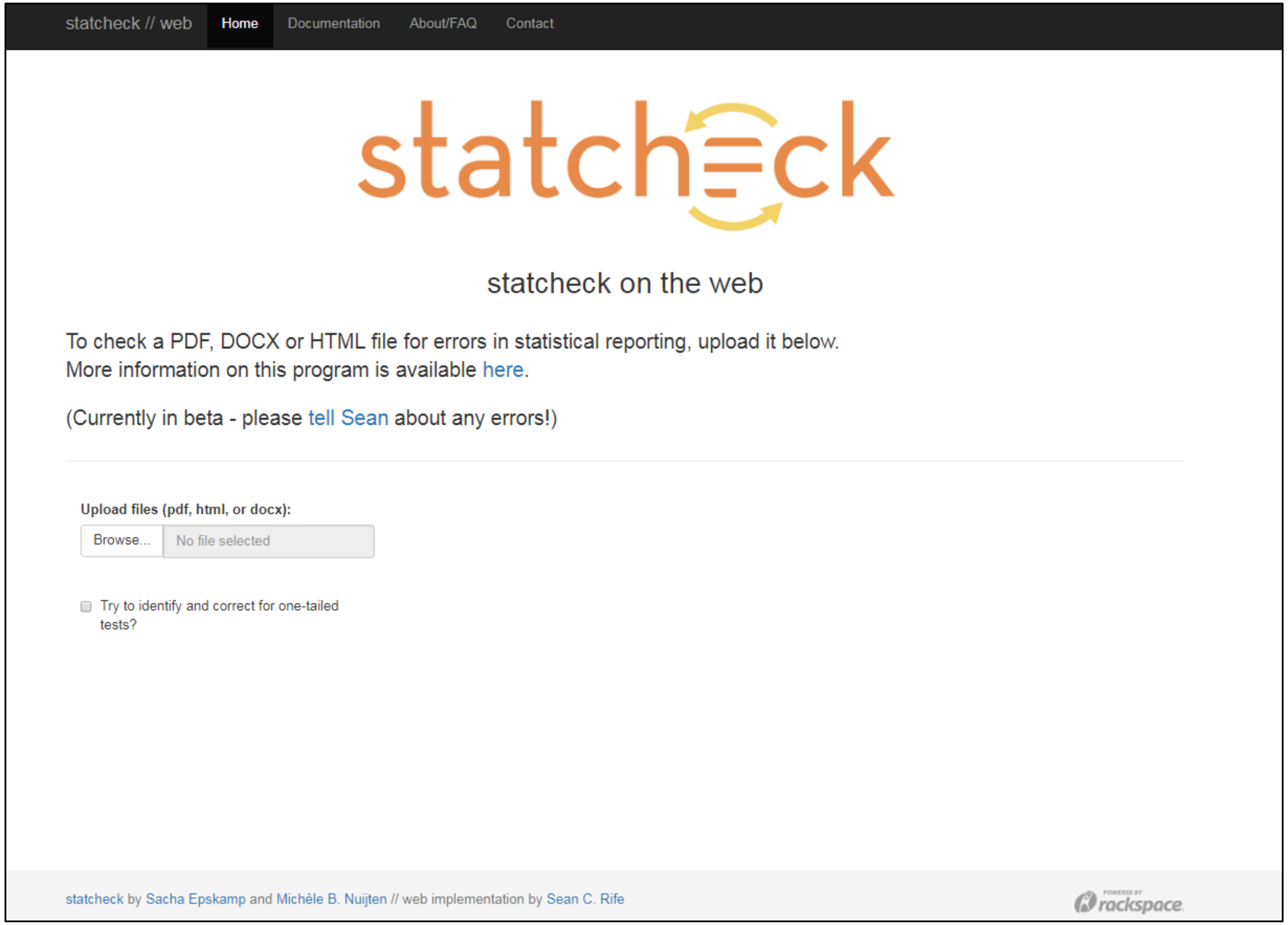
Preventing reporting errors

http://statcheck.io

A "spellchecker" for statistics

(Epskamp & Nuijten, 2014)

- > 28,800 visits since its launch in Sept. 2016
- Used in the peer review process of PS & JESP



statcheck on the web

To check a PDF, DOCX or HTML file for errors in statistical reporting, upload it below.
More information on this program is available here.

(Currently in beta - please tell Sean about any errors!)

Upload files (pdf, html, or docx):

Browse... No file selected

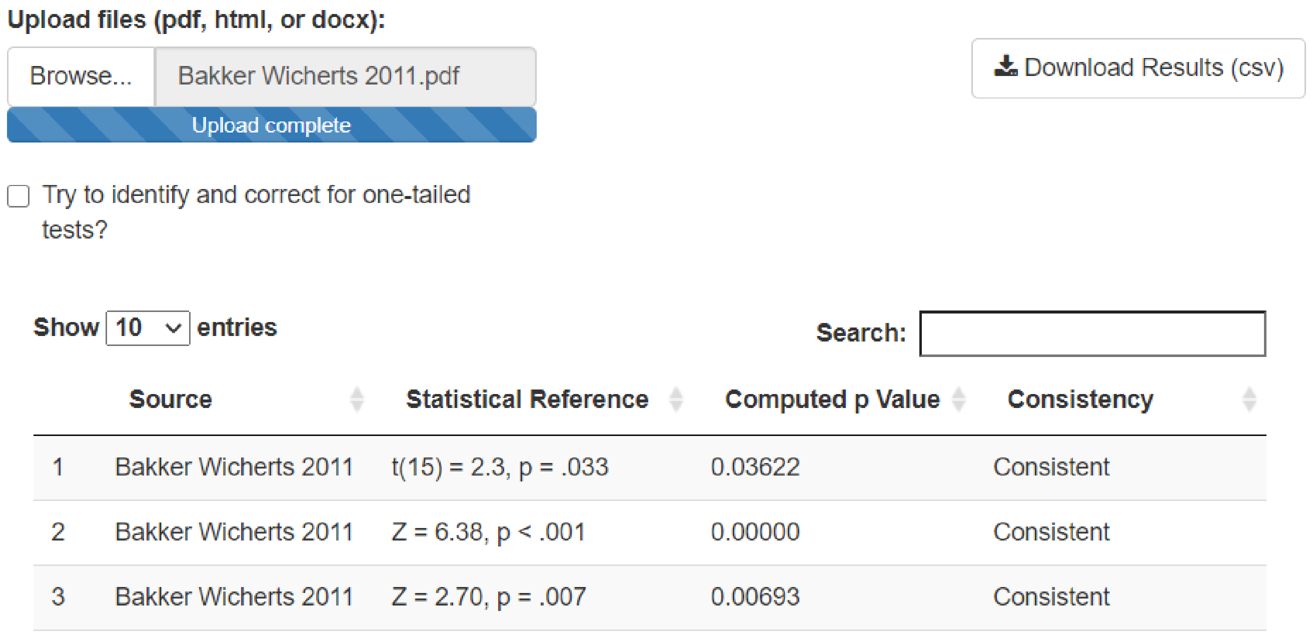
□ Try to identify and correct for one-tailed tests?

statcheck by Sacha Epskamp and Michèle B. Nuijten // web implementation by Sean C. Rife

TILBURG UNIVERSITY

Using statcheck

- To check your own papers before submitting
- To help peer review
- To do meta-research
- As a first robustness check

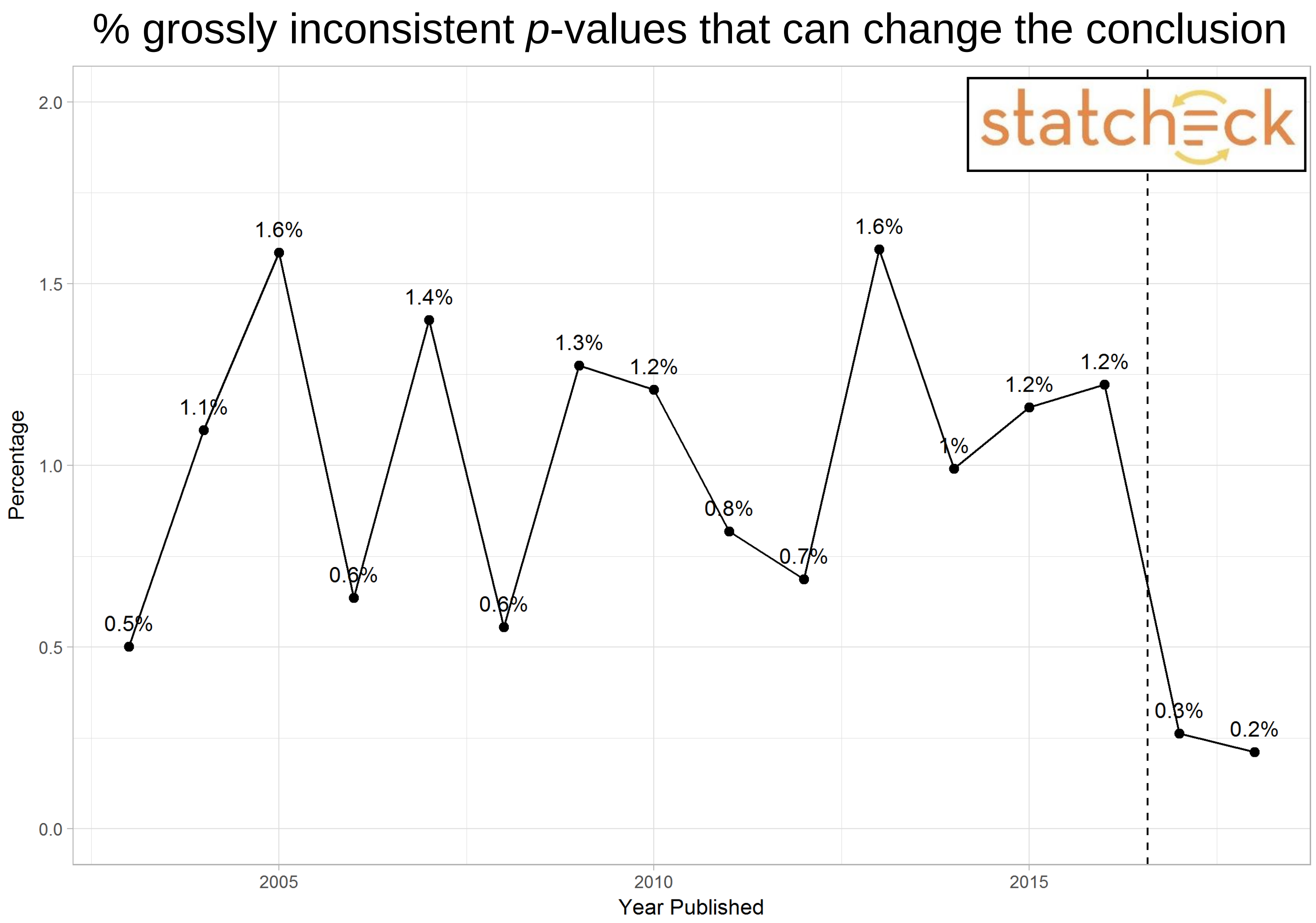


Upload files (pdf, html, or docx):

□ Try to identify and correct for one-tailed tests?

	Source	Statistical Reference	Computed p Value	Consistency
1	Bakker Wicherts 2011	t(15) = 2.3, p = .033	0.03622	Consistent
2	Bakker Wicherts 2011	Z = 6.38, p < .001	0.00000	Consistent
3	Bakker Wicherts 2011	Z = 2.70, p = .007	0.00693	Consistent

Preventing reporting errors



TILBURG UNIVERSITY

To conclude

- Many researcher degrees of freedom exist
  - Preregister your study
  - Do a multiverse analysis
  - Be extremely transparent about all the research decisions that you made on the way

- It is easy to make errors
  - Use statcheck!

References and further reading

- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524-532.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-aac4716.
- 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
- Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*(3), 666-678. doi:10.3758/s13428-011-0089-5
- Berle, D., & Starcevic, V. (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research, 16*(4), 202-207. doi:10.1002/mpr.225
- Garcia-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and P values in medical papers. *BMC Medical Research Methodology, 4:13*. doi:10.1186/1471-2288-4-13
- Nuijten, M. B., Hartgerink, C. H. J., Van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*. Retrieved from https://osf.io/e9qbp/. doi:10.3758/s13428-015-0664-2
- Epskamp, S., & Nuijten, M. B. (2018). statcheck: Extract statistics from articles and recompute p-values (1.3.1) [R package]. Retrieved from https://cran.r-project.org/web/packages/statcheck/index.html.
- Heirene, R., LaPlante, D., Louderback, E. R., Keen, B., Bakker, M., Serafimovska, A., & Gainsbury, S. M. (2021, July 16). Preregistration specificity & adherence: A review of preregistered gambling studies & cross-disciplinary comparison. https://doi.org/10.31234/osf.io/nj4es

References and further reading

- Crüwell S, Evans NJ. Preregistration in Complex Contexts: A Preregistration Template for the Application of Cognitive Models. 2019, December 7. https://doi.org/10.31234/osf.io/2hykx
- Flannery J. fMRI Preregistration Template. 2018. Retrieved from https://osf.io/dvb2e/
- Haven TL, Van Grootel DL. Preregistering qualitative research. Accountability in Research. 2019;26(3):229-244.
- Claesen A, Gomes SLBT, Tuerlinckx F, Vanpaemel W. Preregistration: Comparing Dream to Reality. https://doi.org/10.31234/osf.io/d8wex. 2019, May 9.
- Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., . . . Wicherts, J. M. (2020; in press). Ensuring the quality and specificity of preregistrations. Plos Biology. doi:10.31234/osf.io/cdgyh
- Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*(3), 609-610. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010945212003735. doi:10.1016/j.cortex.2012.12.016
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600-2606. doi:10.1073/pnas.1708274114
- Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & Van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies. A checklist to avoid p-hacking. *Frontiers in psychology, 7*, 1832. doi:10.3389/fpsyg.2016.01832
- Van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., … Bakker, M. (2019, November 20). Preregistration of secondary data analysis: A template and tutorial. https://doi.org/10.31234/osf.io/hvfmr
- Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637
- Simonsohn, Uri and Simmons, Joseph P. and Nelson, Leif D., Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications (October 29, 2019). Available at SSRN: https://ssrn.com/abstract=2694998 or http://dx.doi.org/10.2139/ssrn.2694998