Checking Robustness in 4 Steps

Dr. Michèle B. Nuijten

Sounds like Newton/Nowton

@MicheleNuijten

m.b.nuijten@uvt.nl

http://mbnuijten.com

Tilburg University

My background.

TILBURG UNIVERSITY

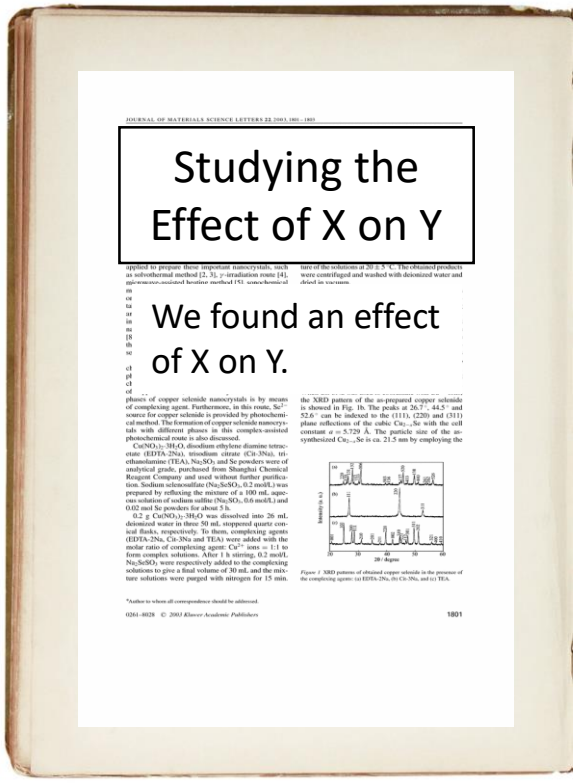
META-RESEARCH CENTER

Tilburg School of Social and Behavioral Sciences

Today.

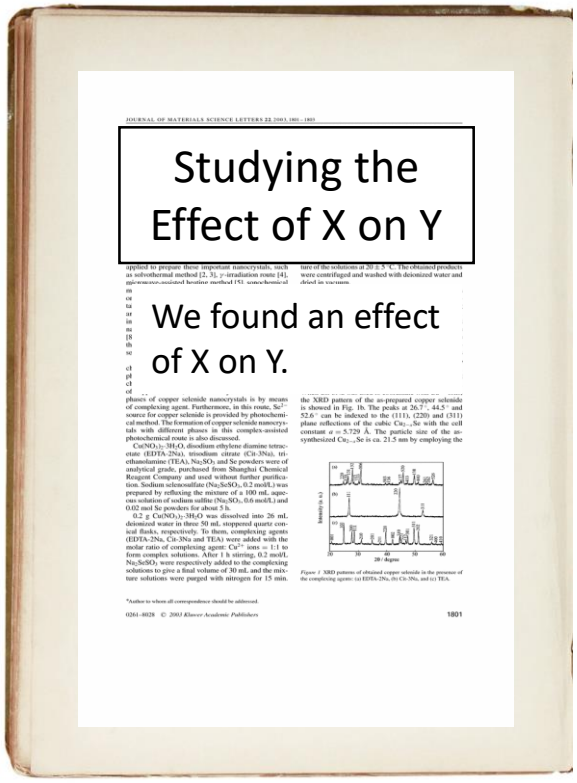
Assessing and **improving** robustness of psychological science in 4 steps (while using minimal resources).

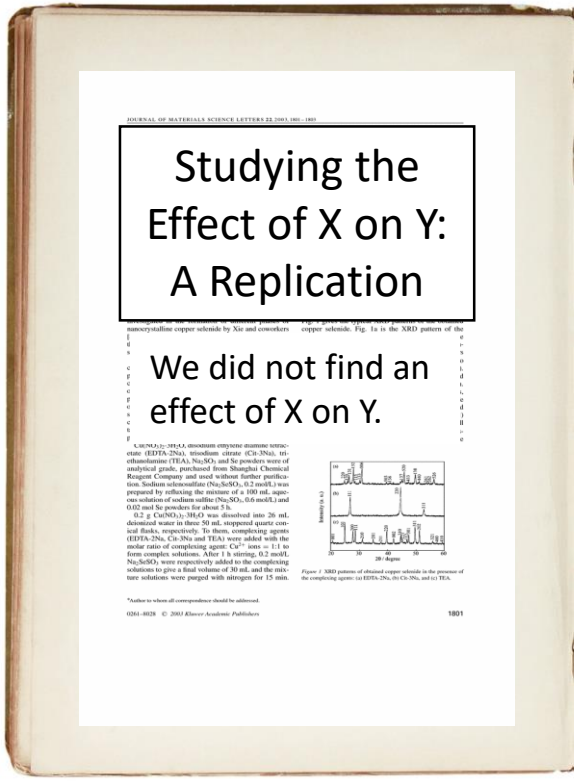
Robustness.



Robustness ≈ “Can I trust this result?”

Assessing robustness through replication?





Cons:

Focus on **reproducibility** first.

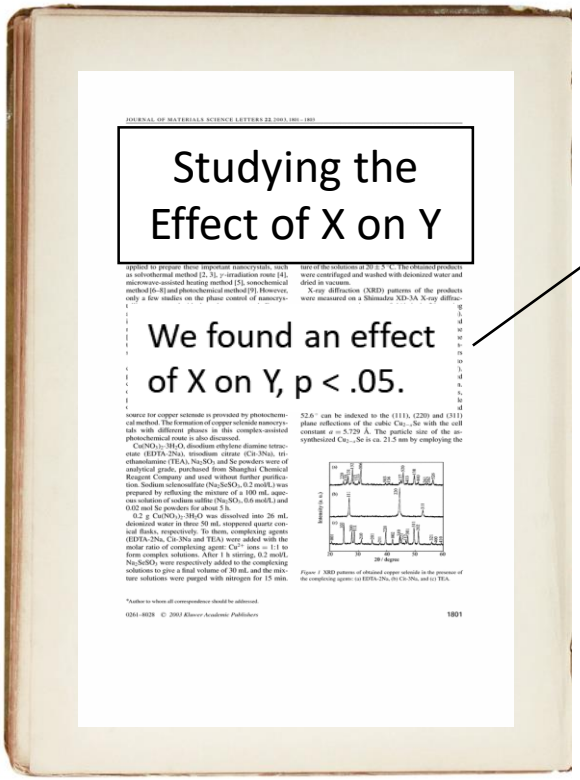
Replicability

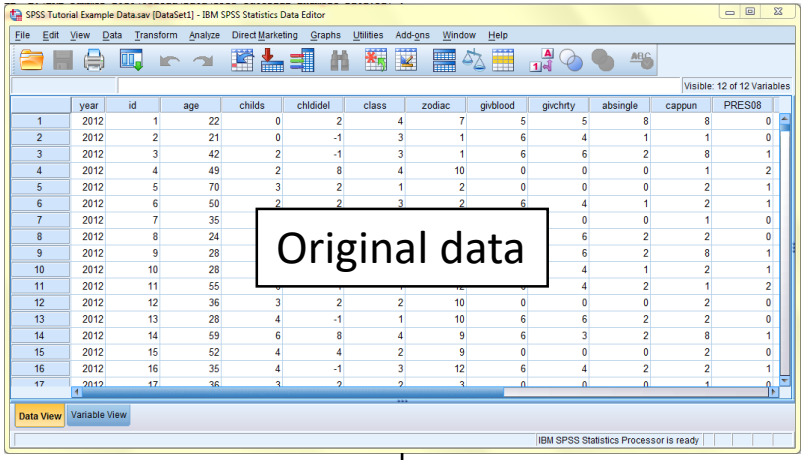
A study is successfully **replicated** if the same/a similar result is found in a new sample.

Reproducibility

A study is successfully **reproduced** if independent reanalysis of the original data, using the same analytic approach, leads to the same results.

Reproducibility is a prerequisite for replicability.

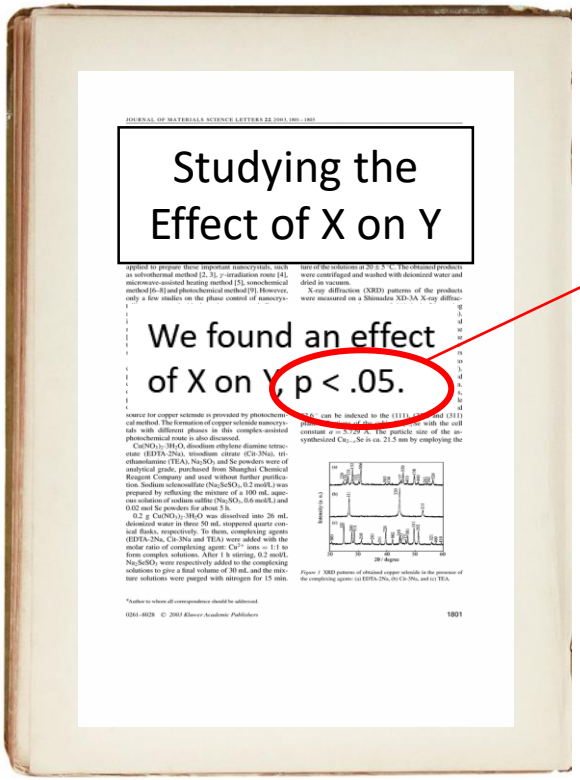




Reanalyze following reported procedures

p > .05??

Reproducibility is a prerequisite for replication.



- If a result is not reproducible, it has no clear bearing on theory or practice
- An irreproducible number is effectively meaningless

You don't need replication to find out whether this finding is robust. It's not.

Today.

Assessing and **improving** robustness of psychological science in 4 steps (while using minimal resources).

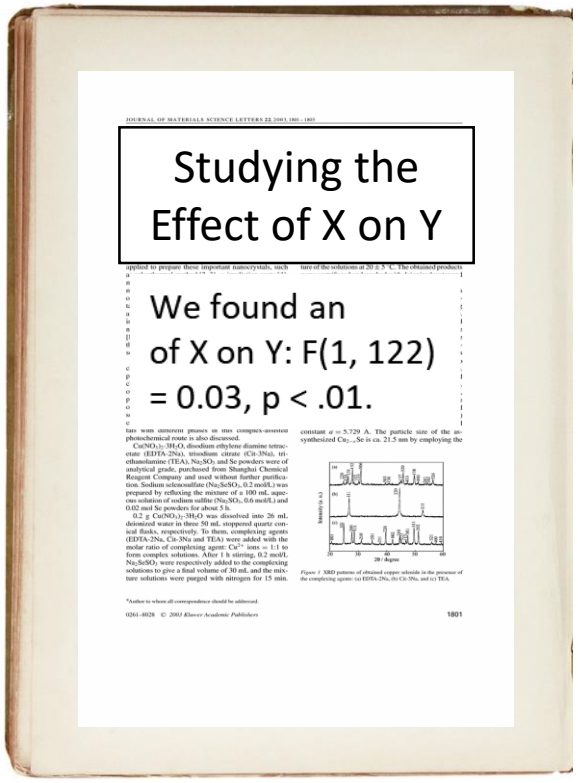
The 4-Step Robustness Check

1. Check the **internal consistency** of the statistical results
2. **Reanalyze** the data using the original analytical strategy
3. Check if the result is robust to **alternative analytical choices**
4. Perform a **replication** study in a new sample

Today.

Assessing and **improving** robustness of psychological science in 4 steps (while using minimal resources).

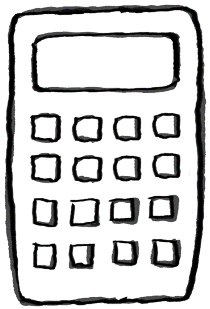
1. Check the internal consistency of the statistical results.



= Statistical sanity check

1. Check the internal consistency of the statistical results.

Also as expected, when priming condition was crossed with age group and time of memory prediction, interaction effects emerged for both the photo recall predictions, $F(1, 122) = 0.03, p < .01$, and the learned recall predictions, $F(1, 135) = 3.75, p < .06$.



p = .86

1. Check the internal consistency of the statistical results.

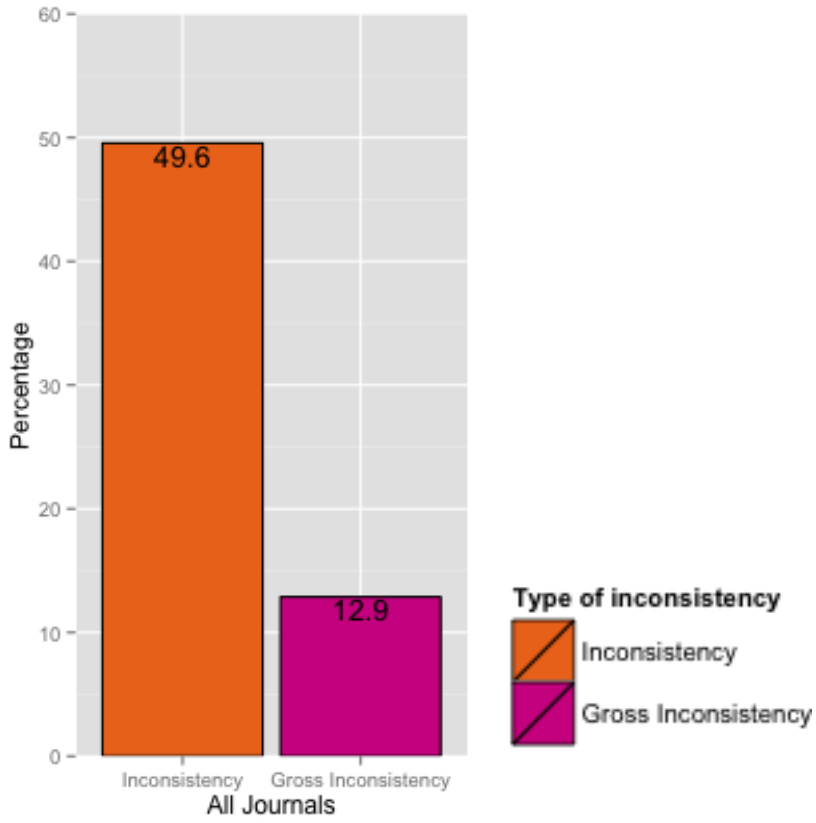
statcheck

R package CRAN 1.3.0

Epskamp & Nuijten, 2014

16,000+ Psychology papers

Nuijten et al. (2016)



1. Check the internal consistency of the statistical results.

statcheck

R package CRAN 1.3.0

Epskamp & Nuijten, 2014

When $ab \neq c - c'$: Published errors in the reports of single-mediator models

John V. Petrocelli · Josh
Melanie B. Whitmi

Algorithmic ide discrepancies between published r reported confidence interv ues FREE

Co athan D Wren

ume 34, Issue 10, 15 May 2018, Pages 1758–1766,

The GRIM Test: echnique Detects Numerous Anomalies in the Reporting of Results in Psych

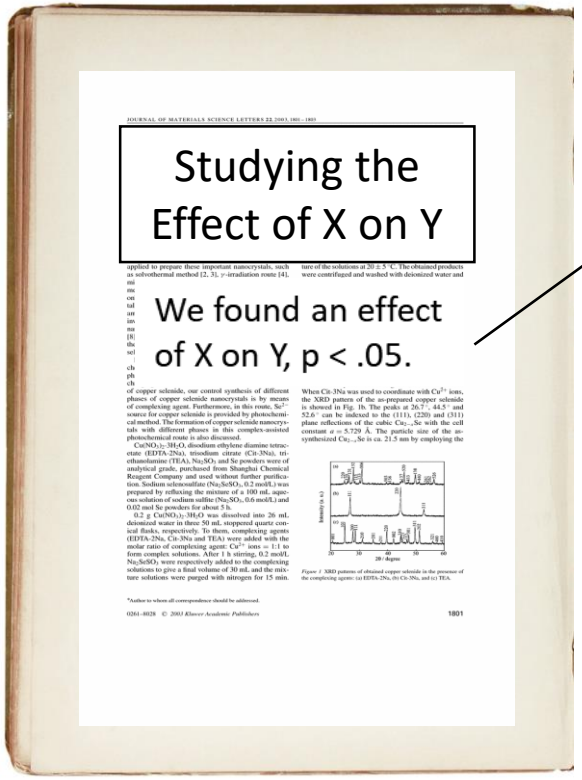
Nicholas J. L. Brown, James A. J. Heathers

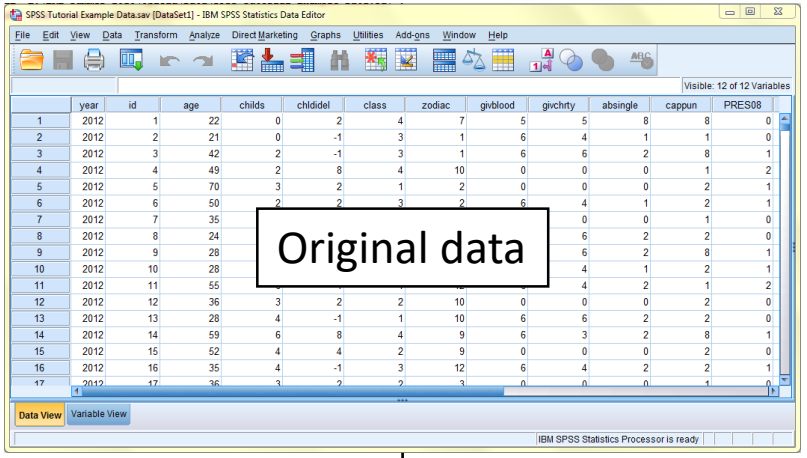
First Published October 18, 2016 | Research Article | Check for updates

https://doi.org/10.1177/1948550616673876

No raw data needed

2. Reanalyze the data using the original analytical strategy.

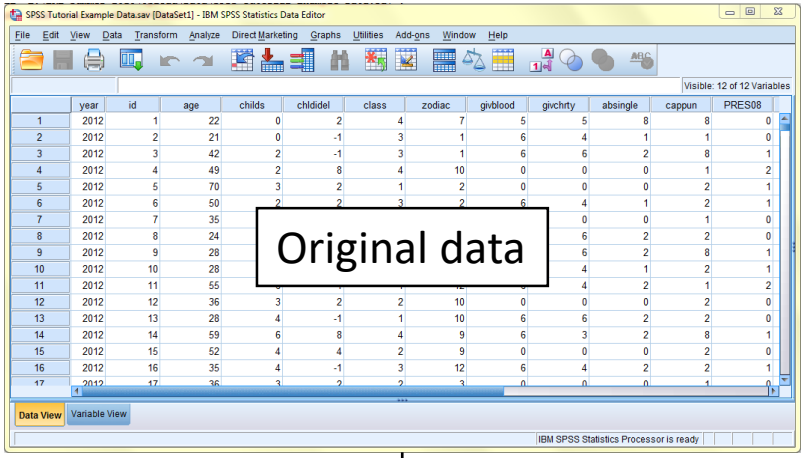




Reanalyze following reported procedures

p = ?

2. Reanalyze the data using the original analytical strategy.



Original data

Reanalyze following reported procedures

p > .05??

Data in psychology often not available

Alsheikh-Ali et al. (2011); VanPaemel et al. (2015); Nuijten et al. (2017); Hardwicke et al. (2019)

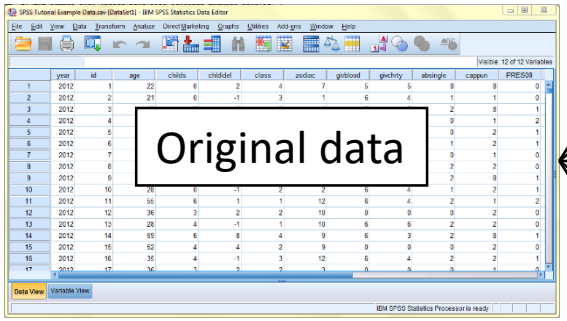
Unusable data or analytical procedure unclear

Kidwell et al. (2016); Hardwicke et al. (2019)

Results not reproducible

Ebrahim et al. (2014); Hardwicke et al. (2018); Maassen et al. (forthcoming)

3. Check if the result is robust to alternative analytical choices.



Test two-tailed instead of one-tailed → p > .05

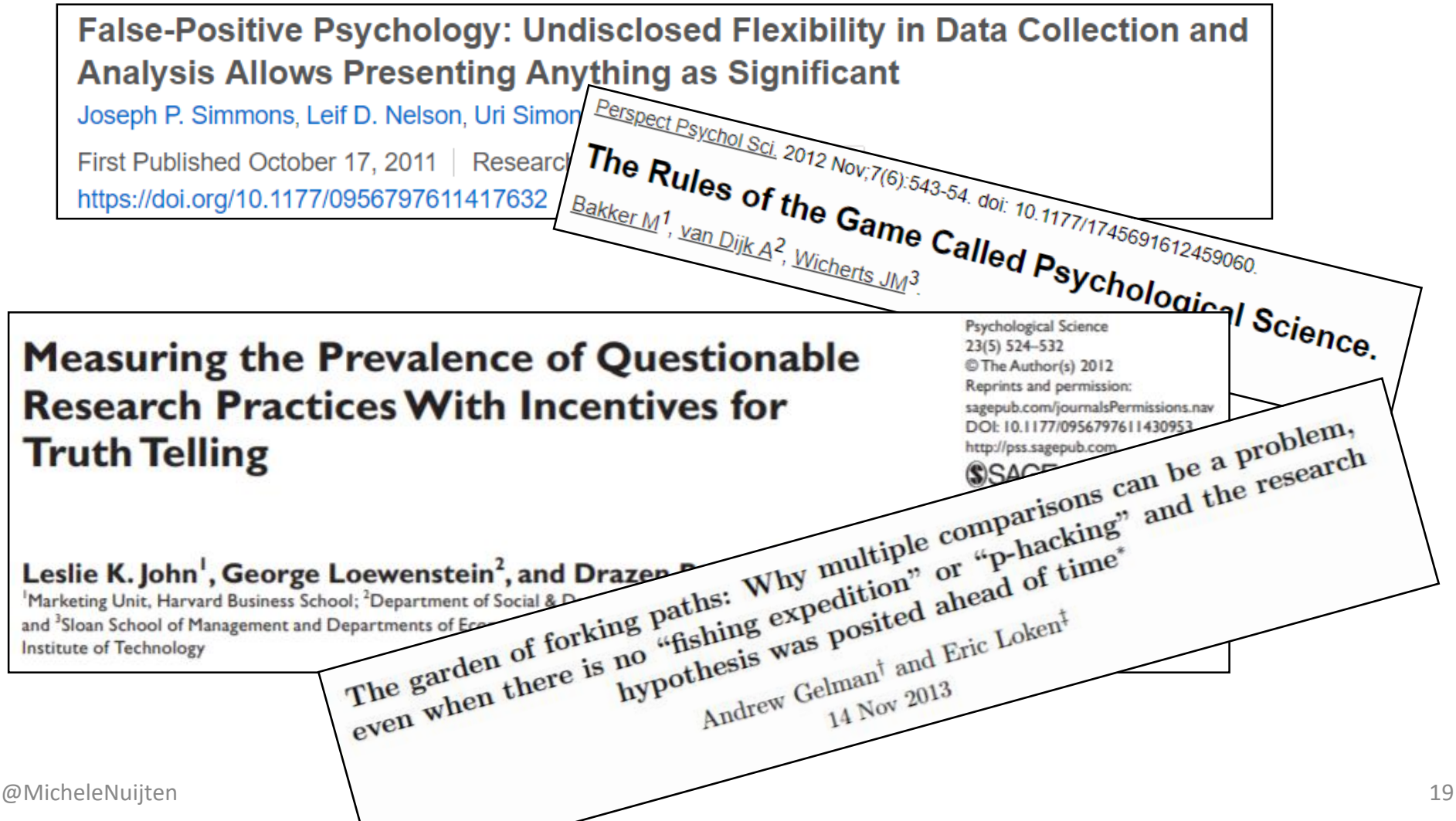
Remove one seemingly arbitrary covariate → p > .05

Reanalyze following reported procedures → p < .05

Include the outlier that was removed → p > .05

Exclude the last observation → p > .05

3. Check if the result is robust to alternative analytical choices.



4. Perform a replication study in a new sample.

1. Check the **internal consistency** of the statistical results

2. **Reanalyze** the data using the original analytical strategy

3. Check if the result is robust to **alternative analytical choices**

4. Perform a **replication** study in a new sample

Failed replication more likely to have bearing on the effect

Today.

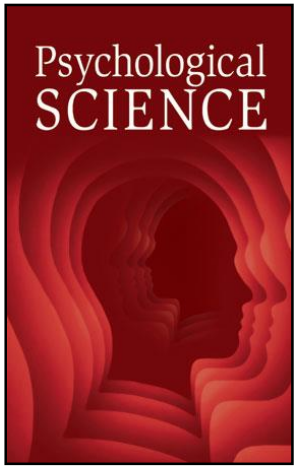
Assessing and **improving** robustness of psychological science in 4 steps (while using minimal resources).

Improving robustness.

1. Check the **internal consistency** of your own statistical results

- Use statcheck and related tools for self-checks / in the peer review process

http://statcheck.io



Improving robustness.

2. Facilitate **reanalyis** of the data

- Share data
- Share well-documented data
- Share analysis scripts
- “In-house” code review (co-authors = co-pilots)
- Code review during peer review
- Fully reproducible dynamic manuscripts (R Markdown, Code Ocean, Docker, etc.)

Improving robustness.

3. Report whether your result is robust to **alternative analytical choices**

- 21-word solution

  Simmons et al. (2011)

These 21 words in a Methods section can *say it* succinctly:

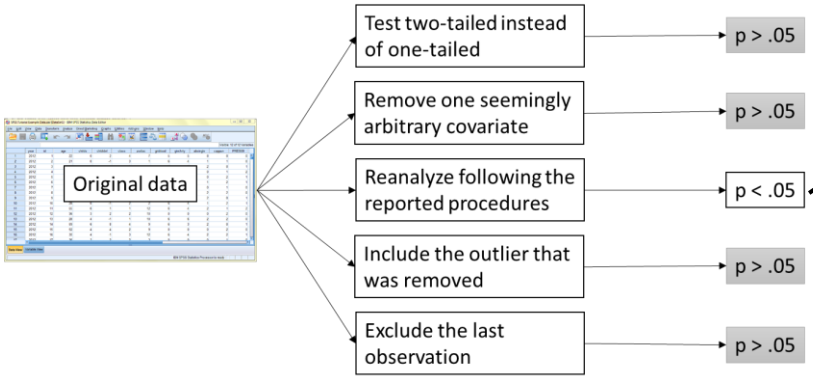
> "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study."

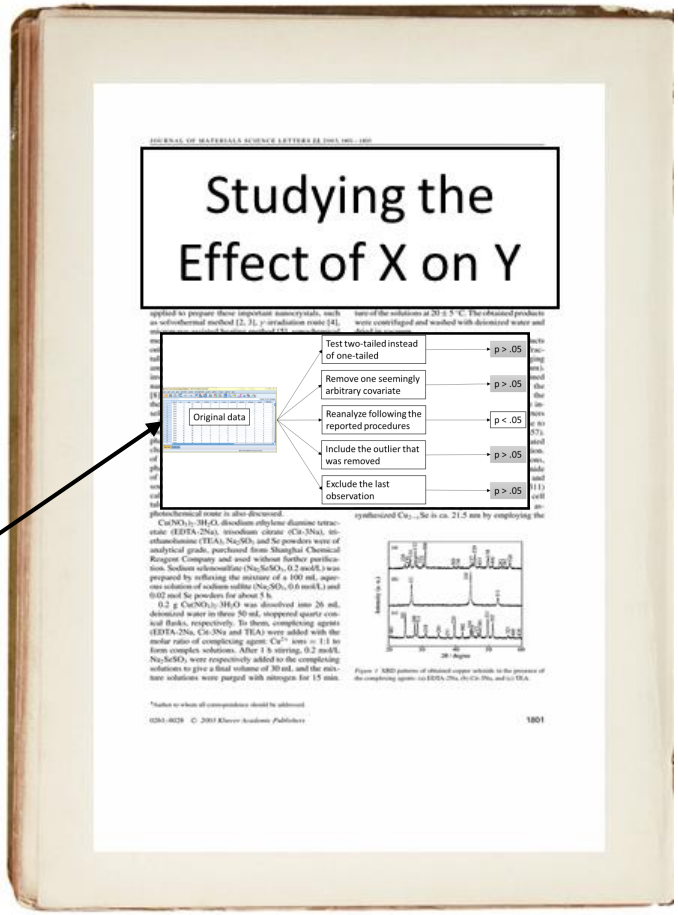
Improving robustness.

3. Check and report whether your result is robust to **alternative analytical choices**

- Journals could require sensitivity analyses
- Multiverse analysis

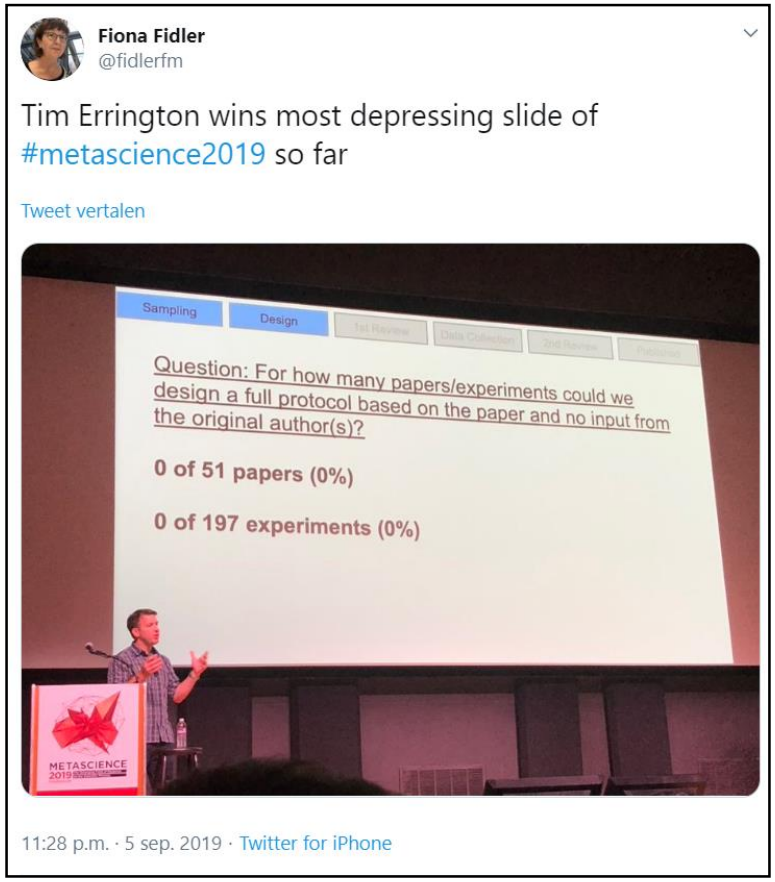
  Steegen et al. (2016)





Improving robustness.

4. Facilitate **replication** in a new sample



Write detailed methods sections/appendices and share materials & protocols!

Discussion.

Assessing and **improving** robustness of psychological science in 4 steps (while using minimal resources).

- If you're interested in the robustness of a **specific study**
- **Context** matters: an inconsistency in the 3rd decimal doesn't automatically mean you shouldn't replicate
- Regardless of the logic of the 4-step robustness check:

All published research should always be reproducible!

Thank you!

A 4-step robustness check to **assess** and **improve** psychological science.

1. Check the **internal consistency** of the statistical results
2. **Reanalyze** the data using the original analytical strategy
3. Check if the result is robust to **alternative analytical choices**
4. Perform a **replication** study in a new sample

META-RESEARCH CENTER
Tilburg School of Social and Behavioral Sciences

NWO
Netherlands Organisation for Scientific Research

Tilburg University


@MicheleNuijten


m.b.nuijten@uvt.nl


http://mbnuijten.com

References.

Alsheikh-Ali, A. A., et al. (2011). "Public availability of published research data in high-impact journals." PLoS One 6(9): e24357.

Bakker, M., et al. (2012). "The rules of the game called psychological science." Perspectives on Psychological Science 7(6): 543-554.

Brown, N. J. L. and J. A. J. Heathers (2016). "The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology." Social Psychological and Personality Science 8(4): 363-369.

Ebrahim, S., Sohani, Z. N., Montoya, L., Agarwal, A., Thorlund, K., Mills, E. J., & Ioannidis, J. P. A. (2014). Reanalyses of Randomized Clinical Trial Data. Jama-Journal of the American Medical Association, 312(10), 1024-1032. doi:10.1001/jama.2014.9646

Epskamp, S. and M. B. Nuijten (2014). statcheck: Extract statistics from articles and recompute p values. R package version 1.0.0. Available from http://CRAN.R-project.org/package=statcheck.

Gelman, A. and E. Loken (2014). "The statistical crisis in science data-dependent analysis - a "garden of forking paths" - explains why many statistically significant comparisons don't hold up." American Scientist 102(6): 460.

Georgescu, C. and J. D. Wren (2017). "Algorithmic identification of discrepancies between published ratios and their reported confidence intervals and p-values." Bioinformatics 34(10): 1758-1766.

Hardwicke, T. E., et al. (2018). "Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition." Royal Society open science 5(8).

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2019). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014-2017). Preprint retrieved from https://osf.io/preprints/metaarxiv/6uhg5/.

John, L. K., et al. (2012). "Measuring the prevalence of questionable research practices with incentives for truth-telling." Psychological science 23: 524-532.

Kidwell, M. C., et al. (2016). "Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency." PLoS biology 14(5): e1002456.

Maassen, E., Van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A. & Wicherts, J. M. (in preparation). Investigating the Reproducibility of Meta-Analyses in Psychology.

Nuijten, M. B. (2018). Research on research: a meta-scientific study of problems and solutions in psychological science. Doctoral dissertation. Available from https://psyarxiv.com/qtk7e.

Nuijten, M. B., et al. (2016). "The prevalence of statistical reporting errors in psychology (1985-2013)." Behavior Research Methods 48(4): 1205-1226.

Nuijten, M. B., et al. (2017). "Journal data sharing policies and statistical reporting inconsistencies in psychology." Collabra: Psychology 3(1): 1-22.

Petrocelli, J., et al. (2012). "When ab ≠ c – c': Published errors in the reports of single-mediator models: Published errors in the reports of single-mediator models." Behavior Research Methods: 1-7.

Simmons, J. P., et al. (2011). "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." Psychological science 22: 1359 –1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. Available at SSRN 2160588. Chicago

Steegen, S., et al. (2016). "Increasing transparency through a multiverse analysis." Perspectives on Psychological Science 11(5): 702-712.

Vanpaemel, W., et al. (2015). "Are we wasting a good crisis? The availability of psychological research data after the storm." Collabra 1(1): 1-5.