# Psychology: From crisis to change

Marjan Bakker; Zurich; 03-09-2018

# Problems in Psychology

- Fraud
  - 2011: Diederik Stapel

# Problems in Psychology

- Publication of “impossible results”
  - Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of personality and social psychology*, *100*(3), 407.

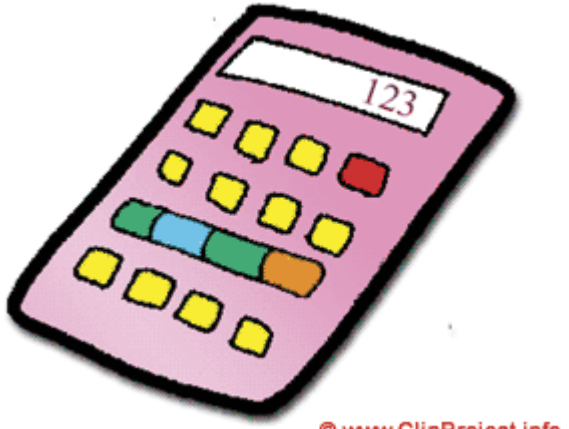
# Problems in Psychology

- Replication problems
  - Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind?. *PloS one*, *7*(1), e29081.
  - Reproducibility project (2015); Many labs projects.

# Problems in Psychology

- Reporting errors

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, $\eta_p^2$ = .09,



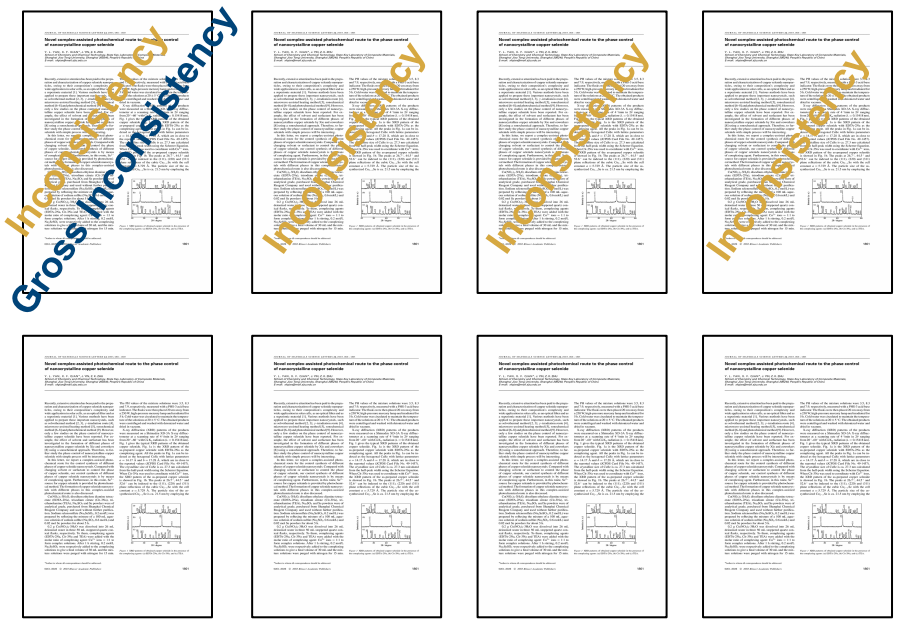
© www.ClipProject.info

***p* = .06**

# Reporting Errors

statcheck

(Epskamp & Nuijten, 2014)




- 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*

(Bakker & Wicherts, 2011)
(Nuijten et al., 2016)

Tilburg University

# Problems in Psychology


- Questionable Research Practices (QRPs), also called p-hacking, or opportunistic use of Researcher Degrees of Freedom (RDFs), are common
  - 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.

# Researcher Degrees of Freedom

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%) |

# Researcher Degrees of Freedom

- Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832.

# 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) |

# 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)

# Scientist are human

- Prone to seeing patterns that don’t exist
- Prone to cognitive biases (confirmation bias, hindsight bias)
- Prone to forgetfulness
- Prone to human error
- Prone to the incentives for obtaining significant results

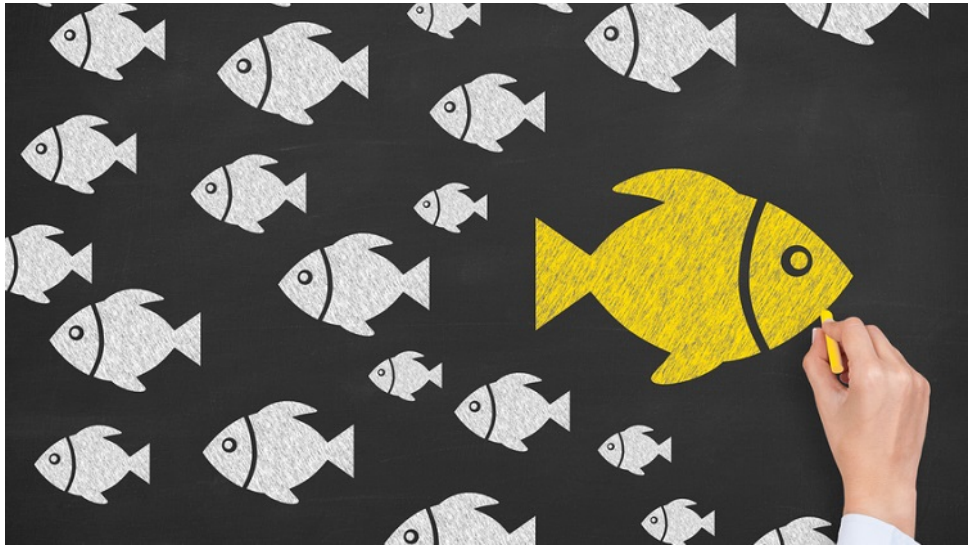
# Results?

- False positive findings
- Distorted meta-analytical results
- Overestimation of effect sizes

- … distrust in science

See also Simmons, Nelson, & Simonsohn (2011), or Bakker, Van Dijk, & Wicherts (2012)

# But also change!



# Preventing reporting errors

## http://statcheck.io

A “spellchecker” for statistics

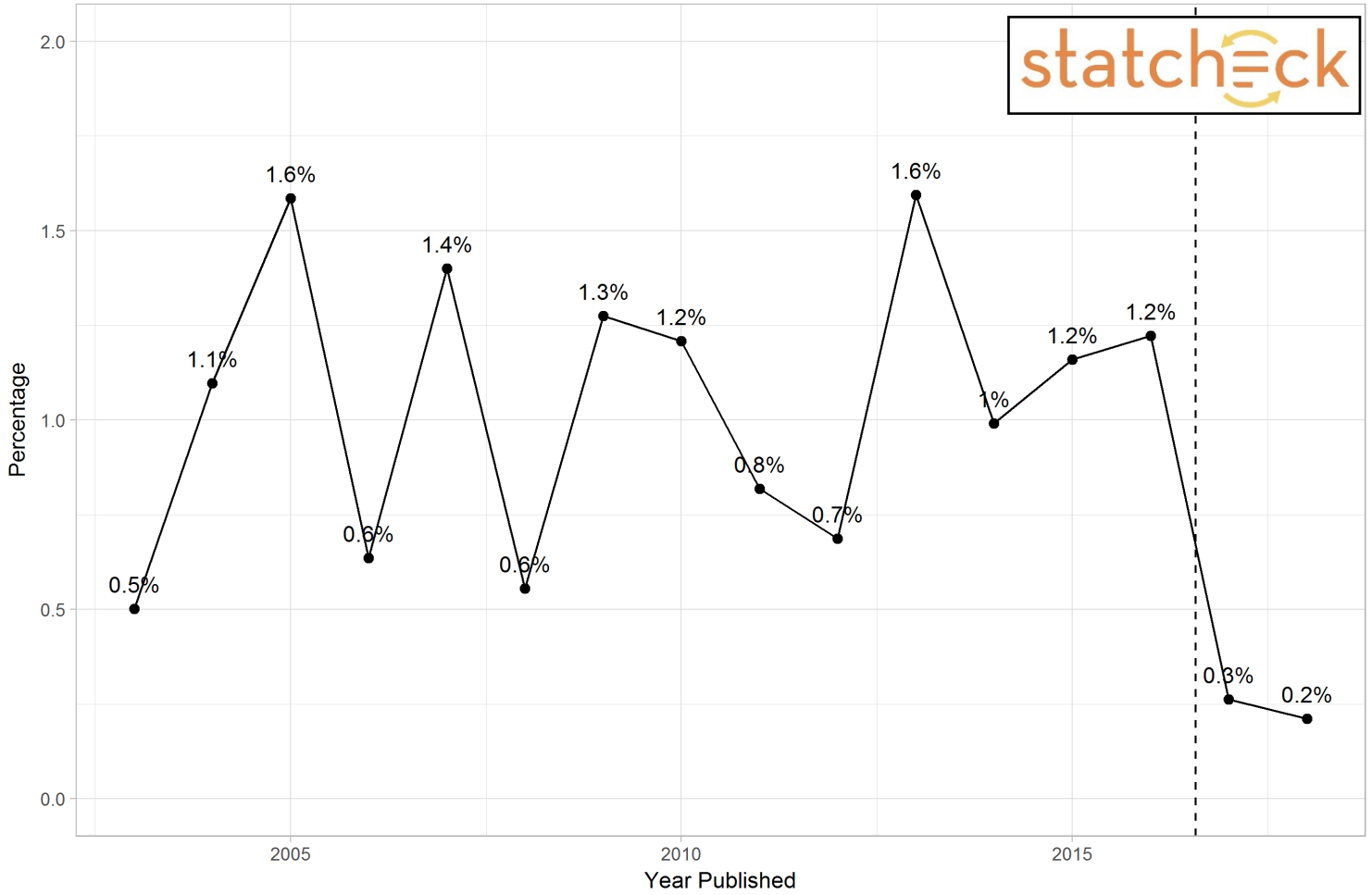
(Epskamp & Nuijten, 2014)

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


TILBURG UNIVERSITY

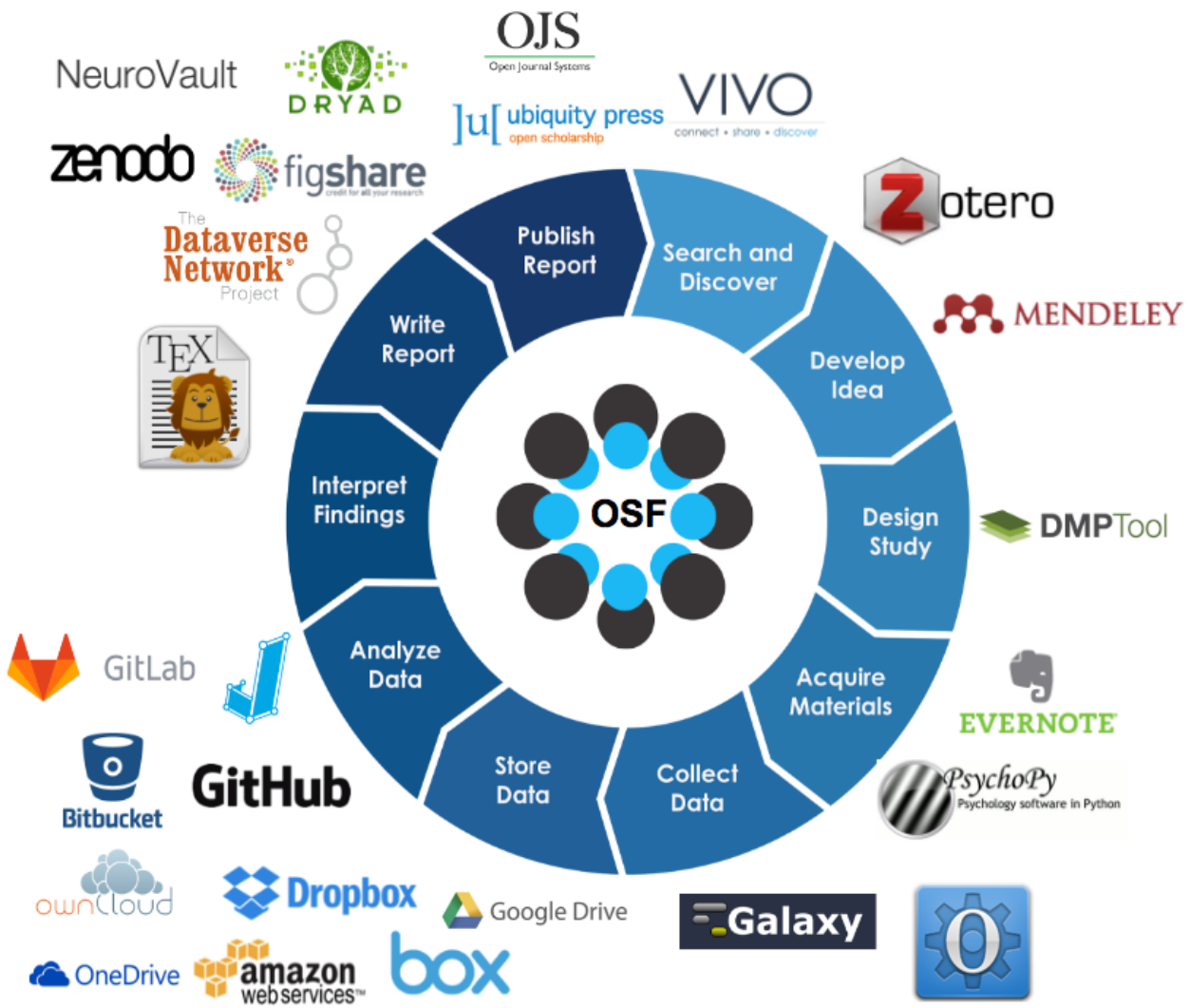
# Preventing reporting errors


% grossly inconsistent *p*-values that can change the conclusion



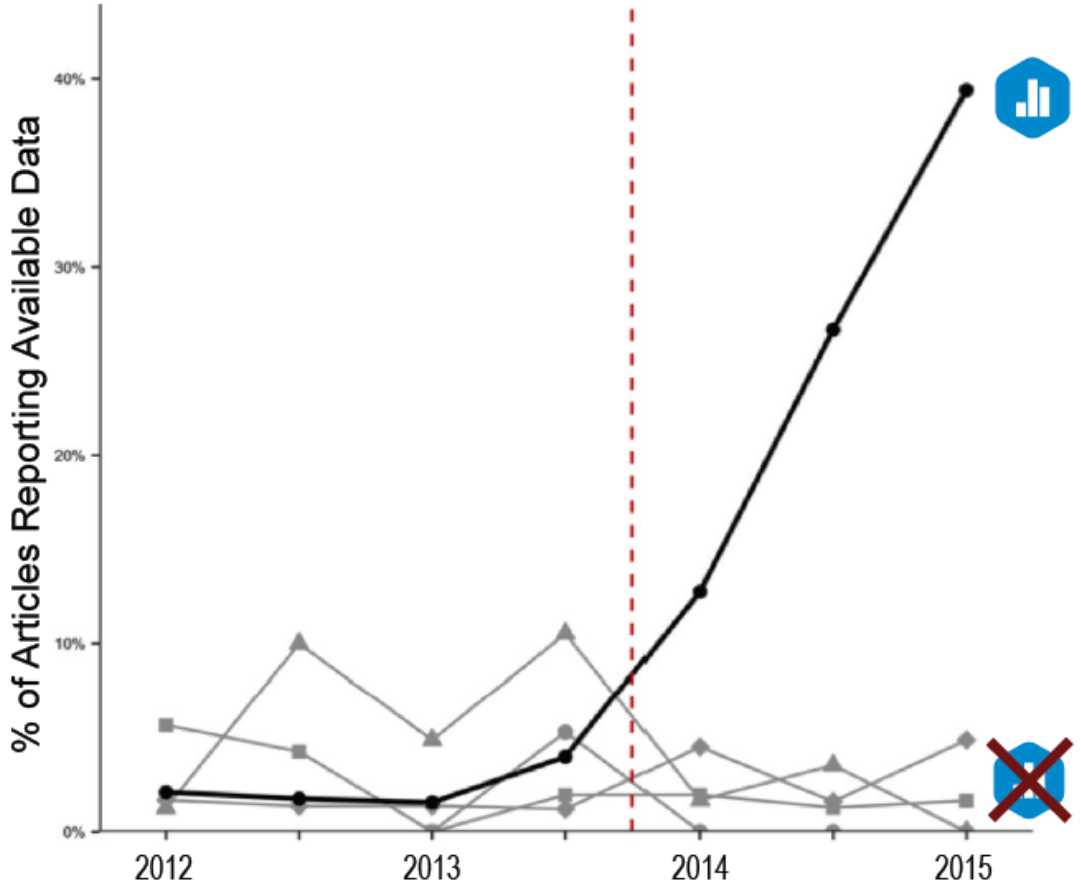

TILBURG UNIVERSITY

# Open Science Framework



# Badges


PREREGISTERED
OPEN DATA
OPEN MATERIALS


% of Articles Reporting Available Data
40%
30%
20%
10%
0%
2012
2013
2014
2015

# Preregistration and Registered Reports

- Preregistration
  - Clear distinction between two modes of research:
    - Confirmatory testing (data is collected to test predictions)
    - Exploratory analysis (data is used to generate predictions that could be tested in the future)
  - Restrict opportunistic use of researcher degrees of freedom

# Preregistration and 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

# From theory to practice

- Preregistration
  - Preregistration Challenge
  - Preregistration badges
    - 38 journals award badges

PREREGISTRATION CHALLENGE


- Registered reports
  - 125 journals offer this format

# … 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.

# … and to Research

Example 1:

"*We will use the PANAS to measure affect*":

Options:

- Use subscale ('positive affect' or 'negative affect' items)
- Only use items that correlate with dependent variable
- Use sum score, weighted mean score, factor score
- Delete items with lower reliability

# … and to Research

Example 2:

"*participants who did not participate seriously will be excluded*"

Options:

- Define criteria for 'participating seriously' anyway you like

- Also use other criteria (looked tired, etc.)

- Exclude from one analysis but not another

# SIPS


SIPS
SOCIETY FOR THE
IMPROVEMENT OF
PSYCHOLOGICAL SCIENCE

# Psychological Science Accelerator

A globally distributed network of psychological science laboratories (currently over 300), representing over 45 countries on all six populated continents, that coordinates data collection for democratically selected studies.

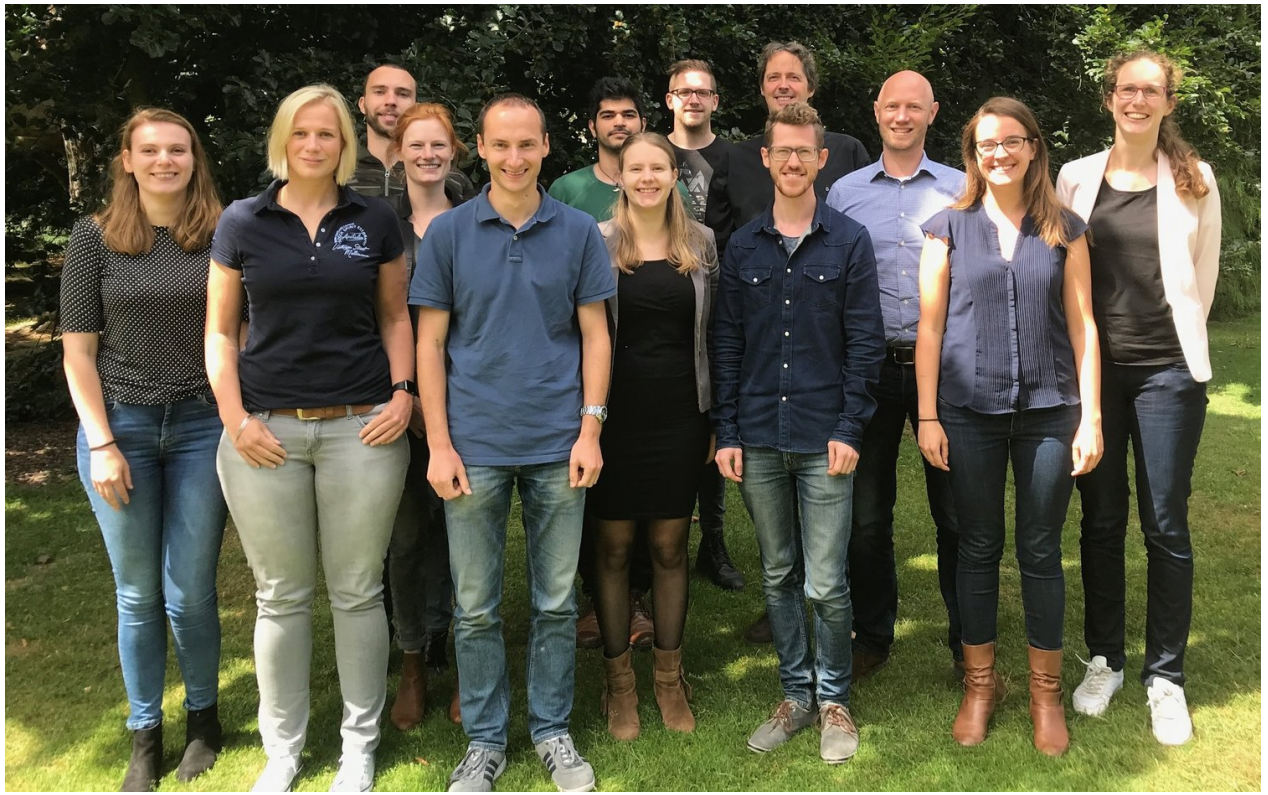
# Meta-Research Center

- Continue with meta-research
- Investigate solutions
  - Comparison of preregistrations with their publication
- Develop guidelines for researchers
  - Questions
  - Training
- Improve science!

META-RESEARCH CENTER

Tilburg School of Social and Behavioral Sciences

# Thank you



TILBURG UNIVERSITY

META-RESEARCH CENTER

Tilburg School of Social and Behavioral Sciences