gesis

Leibniz Institute
for the Social Sciences

Preregistration of secondary data analysis

Tobias Heycke & Bernd Weiß, 16.07.2019, Zagreb

Leibniz
Association

p-hacking

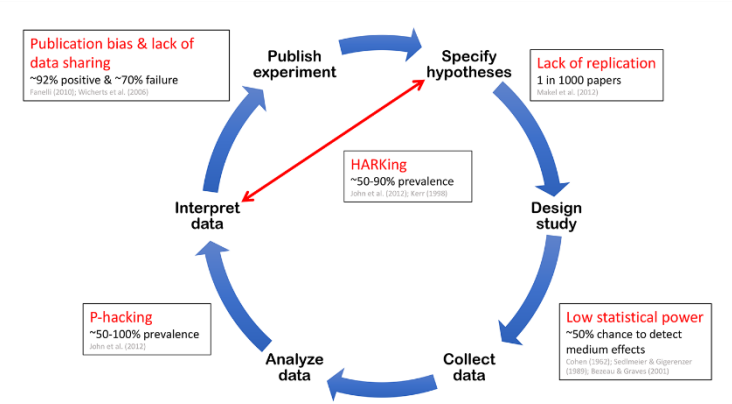
- Trying out multiple statistical analyses until a *p* value smaller than .05 is found and reporting only this analysis

Simmons, Nelson, & Simonsohn (2011), Psychological Science

HARKing

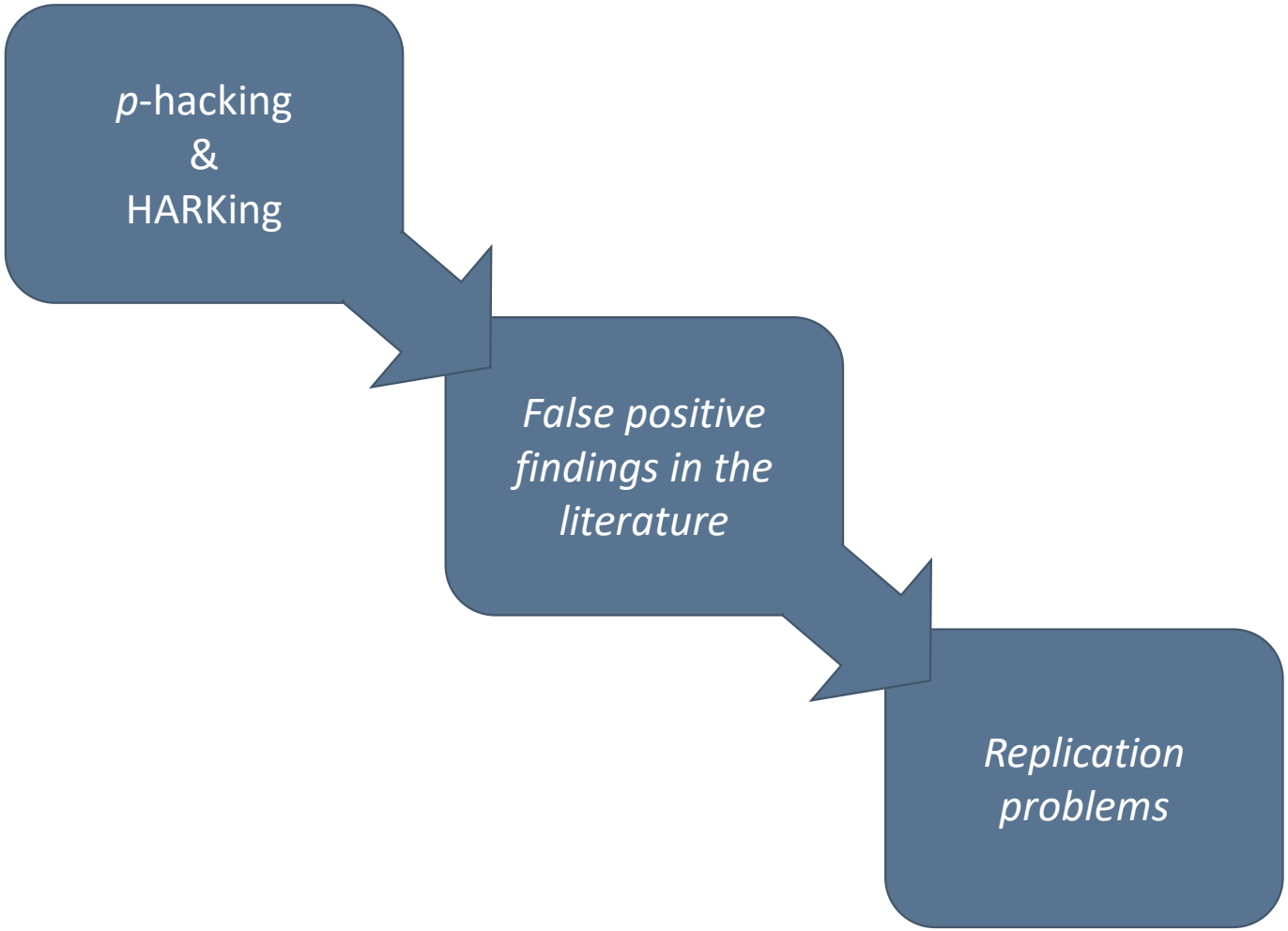
Hypothesizing After the Results are Known (HARKing):

- presenting post-hoc hypotheses (usually based on statistically significant results) as a priori hypotheses



Kerr (1998), Personality and Social Psychology Review

The Big Picture



Preregistration

- “When you preregister your research, you're simply specifying your plan in advance, before you gather data”
- Commitment is usually accomplished by posting it to an independent registry
- Forms that can be filled out (e.g., osf.io/prereg)

Benefits Preregistration

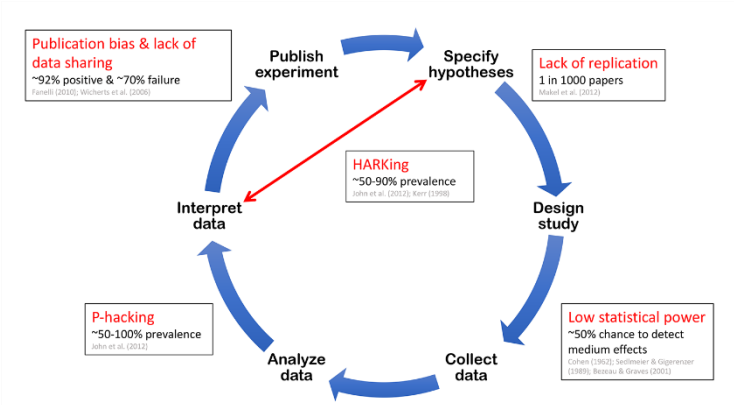
- Distinguish between confirmatory and exploratory analyses (the HARKing problem)
- Restrict researchers degrees of freedom (*p* hacking problem)

Preregistration before data collection

- "When you preregister your research, you're simply specifying your plan in advance, before you gather data"
- Many researchers in the social sciences depend on large data sets and cannot collect data themselves (i.e., research based on secondary data analysis)

P-hacking and HARKing still possible

- Many variables in social sciences data sets
- Therefore easy to find (supposedly) meaningful results



Three levels of secondary-data preregistration

1. Data publicly available
2. Data needs to be requested
3. Data collected but not available yet

1. Data publicly available

- Writing preregistration when aware of results is scientific misconduct
- Scientists need to give an estimate how well they know the data
  - See https://osf.io/ne3bw/ for a first template
  - Sign the form
- Other (more detailed) preregistration templates are already available (https://osf.io/x4gzt/)

2. Data need to be requested

- Use the templates and forms (see previous page)
- Additionally: data distributing institutions could certify when data access was granted to requester
- Example form (CC 0 license):
  - https://osf.io/6yguf/
- Signed form can be uploaded to public repository

3. Data collected but not available yet

- Code book should be made available as soon as data collection started
- Announce code book and approximate (earliest) time of data publication
- Use form to report knowledge (of previous waves)

Synthetic practice data

For all (but especially 2 and 3):

- Provide practice data to write statistical code before seeing the data

Create practice data

A first script can be found here:

```
https://gist.githubusercontent.com/TobiasHeycke/da27cab
493643e2284f7a8c8a60a9080/raw/040d8d4627796ece652d59924
39cdc715cb1d308/synthdata.R
```

Outlook

Facilitating preregistration of secondary data analysis – who should be involved?

- Data suppliers
- Journals
- Authors/Scientists

Thank you for your attention

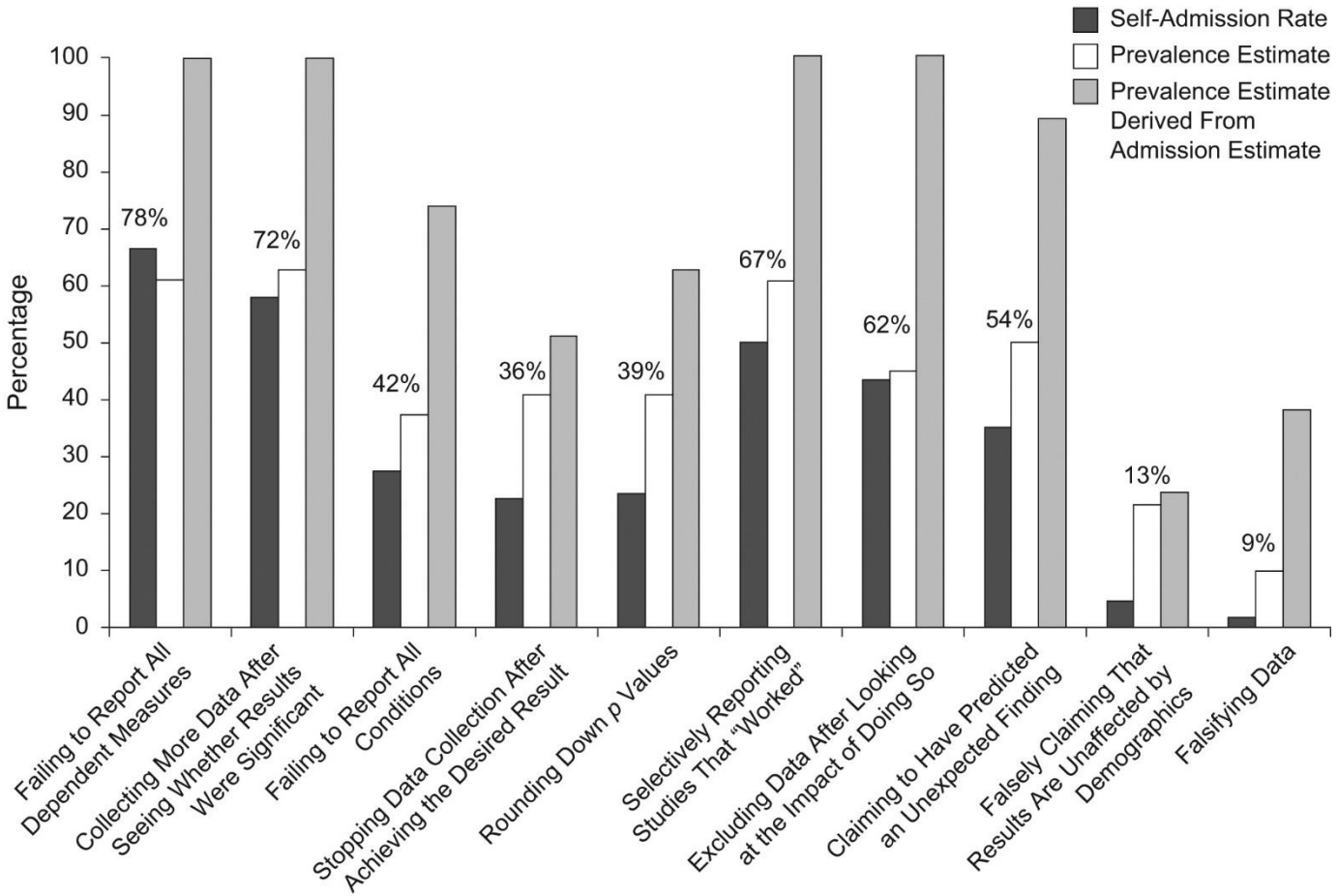
Slides: https://osf.io/cqb47/

gesis Leibniz Institute for the Social Sciences

Leibniz Association

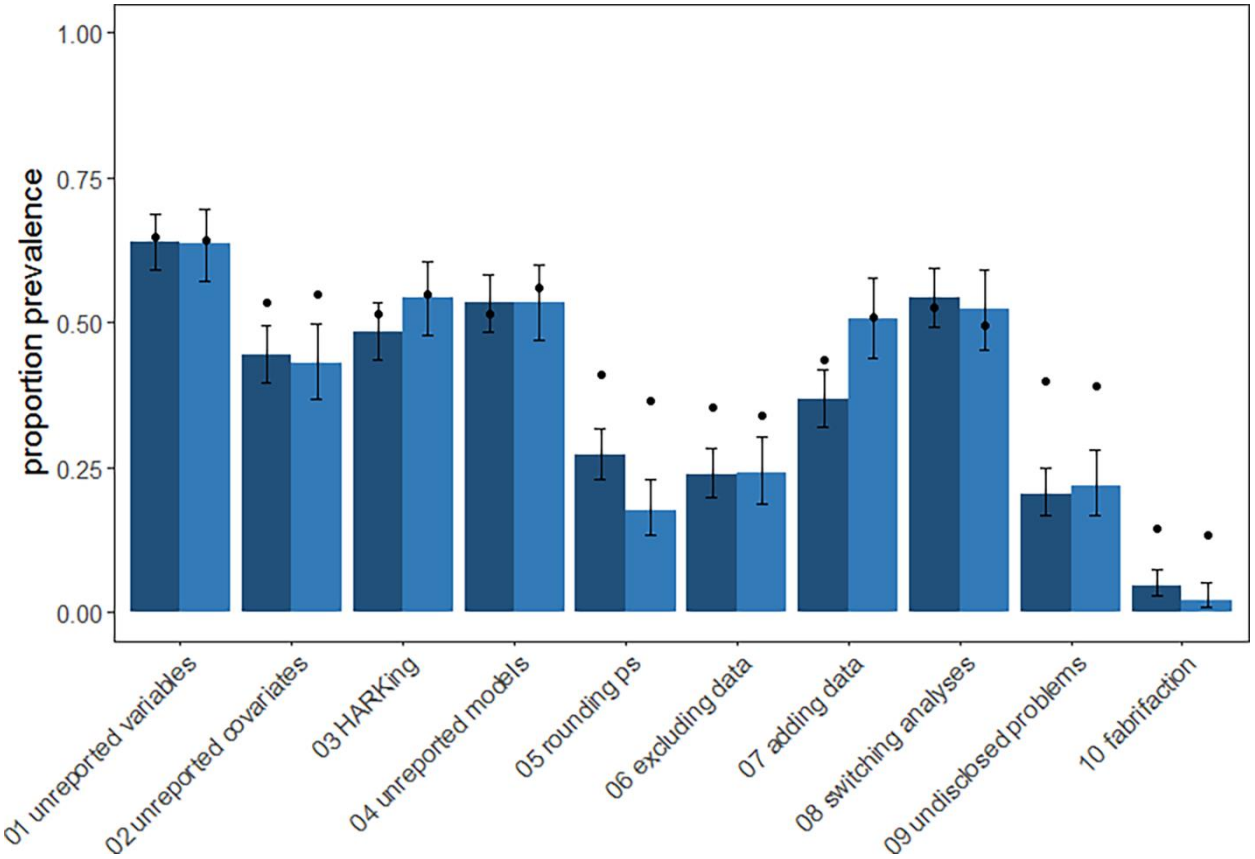
Contact:
tobias.heycke@gesis.org @TobiasHeycke
bernd.weiss@gesis.org @berndweiss

Usage QRPs (Psychology)



John, Loewenstein, & Prelec (2012), Psychological Science

Usage QRPs (Ecology and Evolution)

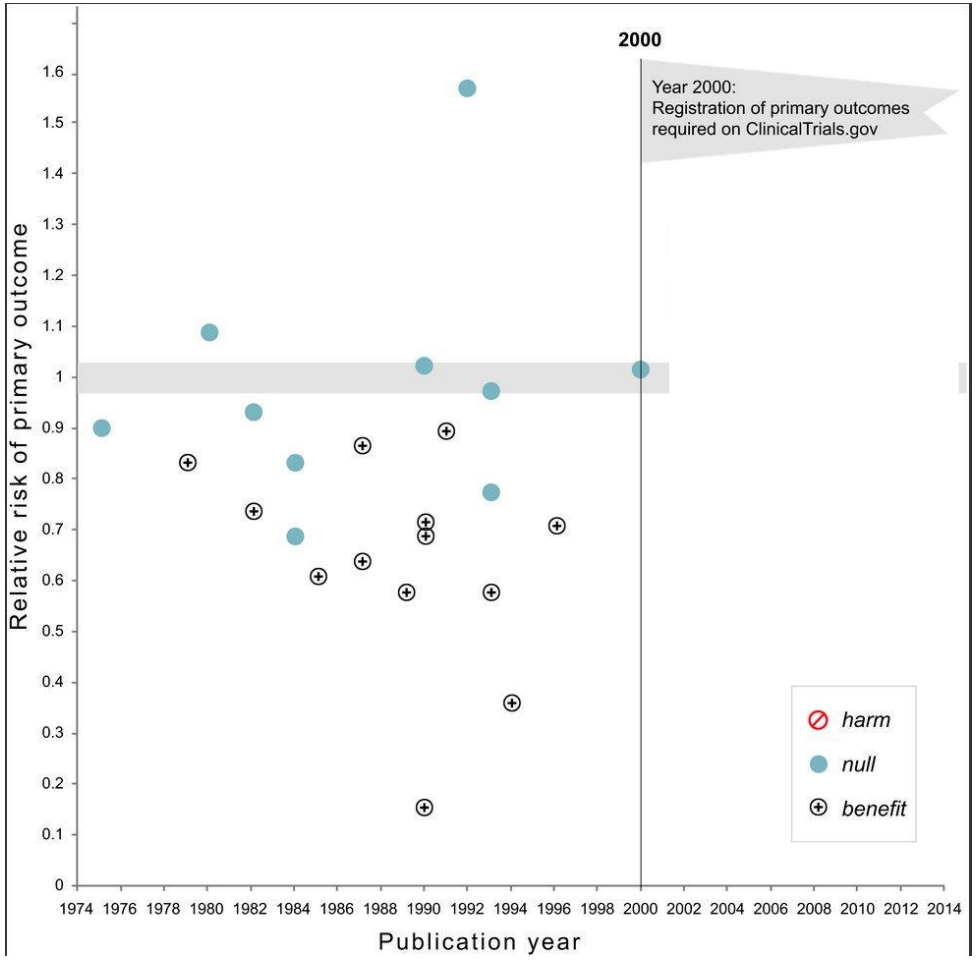


Fraser, Parker, Nakagawa, Barnett, & Fidler (2018). PLOS ONE

Registered Reports

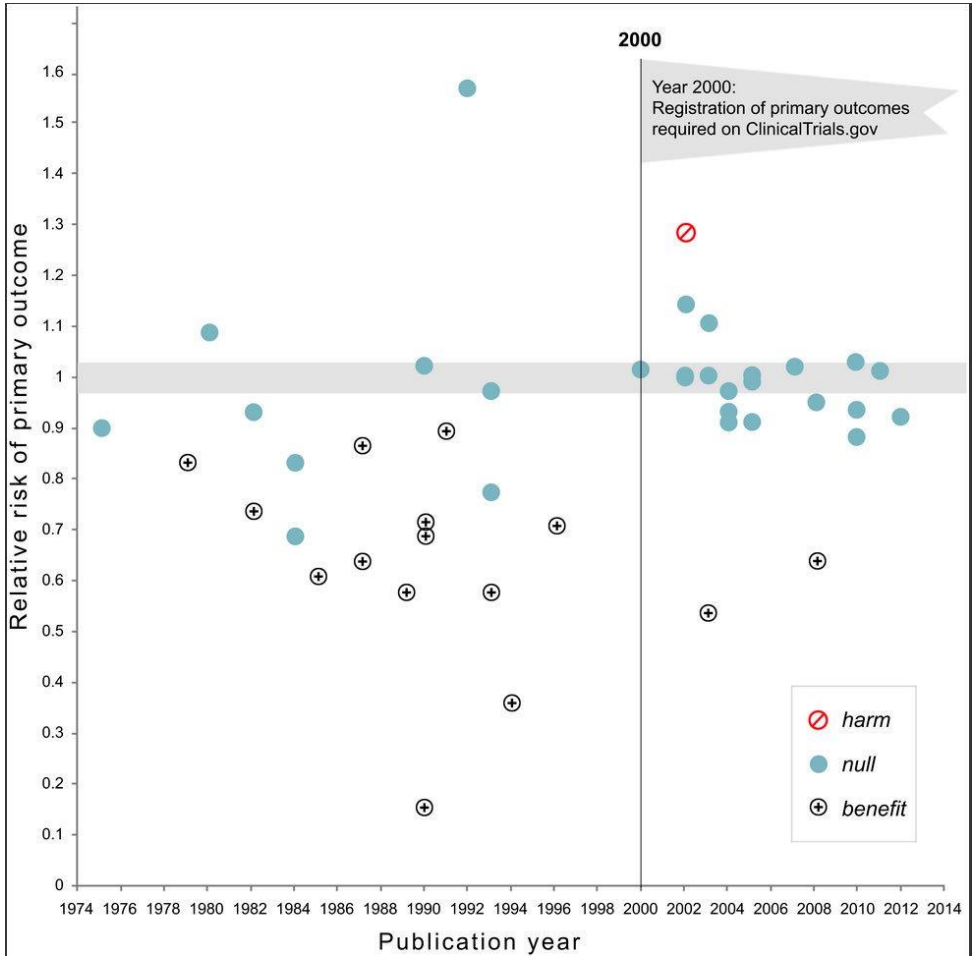
See cos.io/rr for more information and participating journals (N = 203, 13.07.2019)

Result of registered reports



Kaplan & Irvin (2015). PLOS ONE

Result of registered reports



Kaplan & Irvin (2015). PLOS ONE