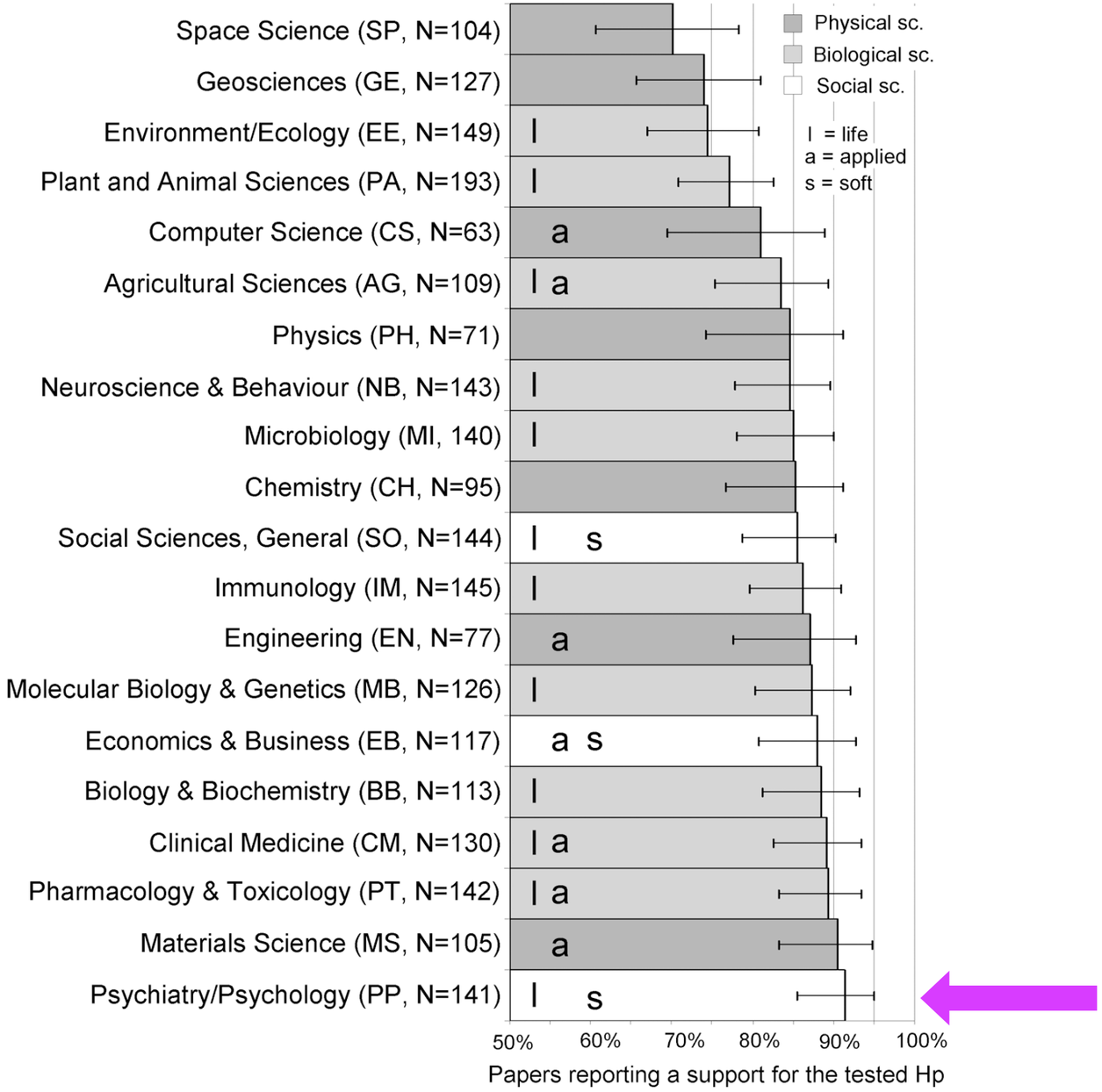
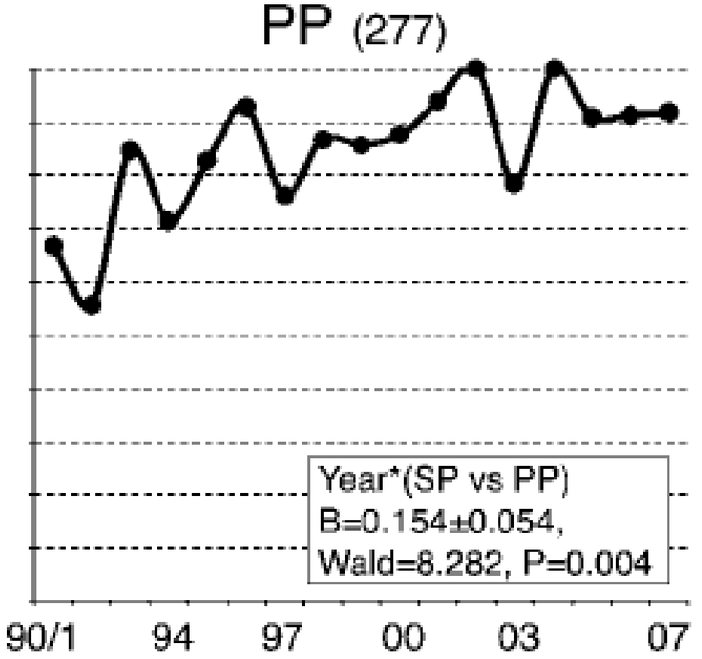
Positive result rates in psychology: Registered Reports compared to the conventional literature

Mitchell Schijen **Anne Scheel** **Daniël Lakens**

a.m.scheel@tue.nl
@AnneMScheel

Eindhoven University of Technology

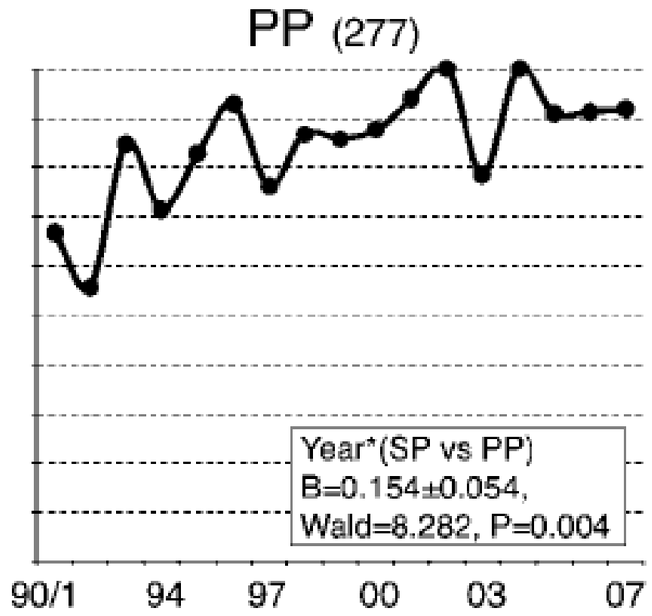


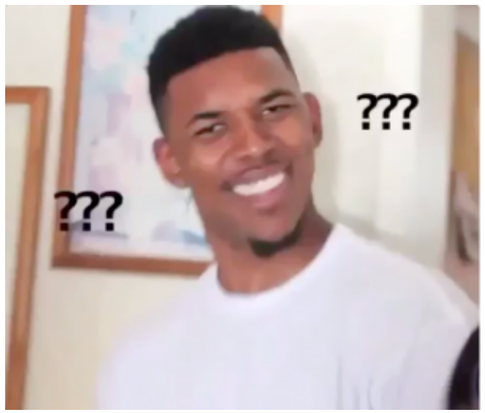


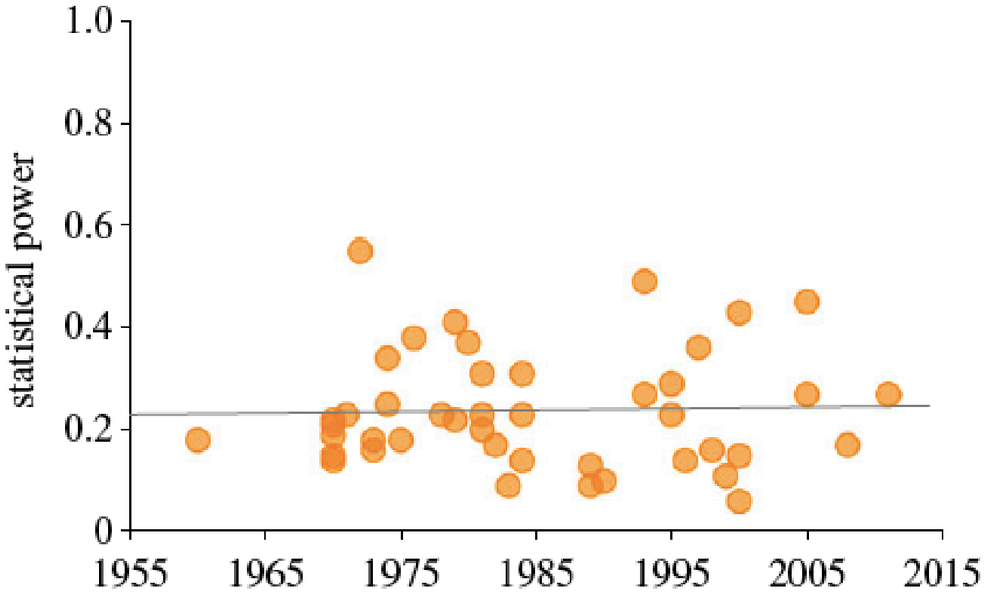
Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891-904.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of sciences. *PLOS ONE, 5(4)*: e10068.

Low power, high success rate







Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891-904.

Smaldino, P. E. & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*, 160384.

Two possible explanations

1) Psychologists only test true hypotheses with very large effect sizes

2) **Bias**: negative results don't get published
   - file-drawering
   - *p*-hacking
   - other questionable research practices
   - coerced overfitting (through reviewers & editors)

One proposed remedy: Registered Reports

preregistration + publication decision ahead of time

 reduces questionable research practices **and** publication bias

The present study

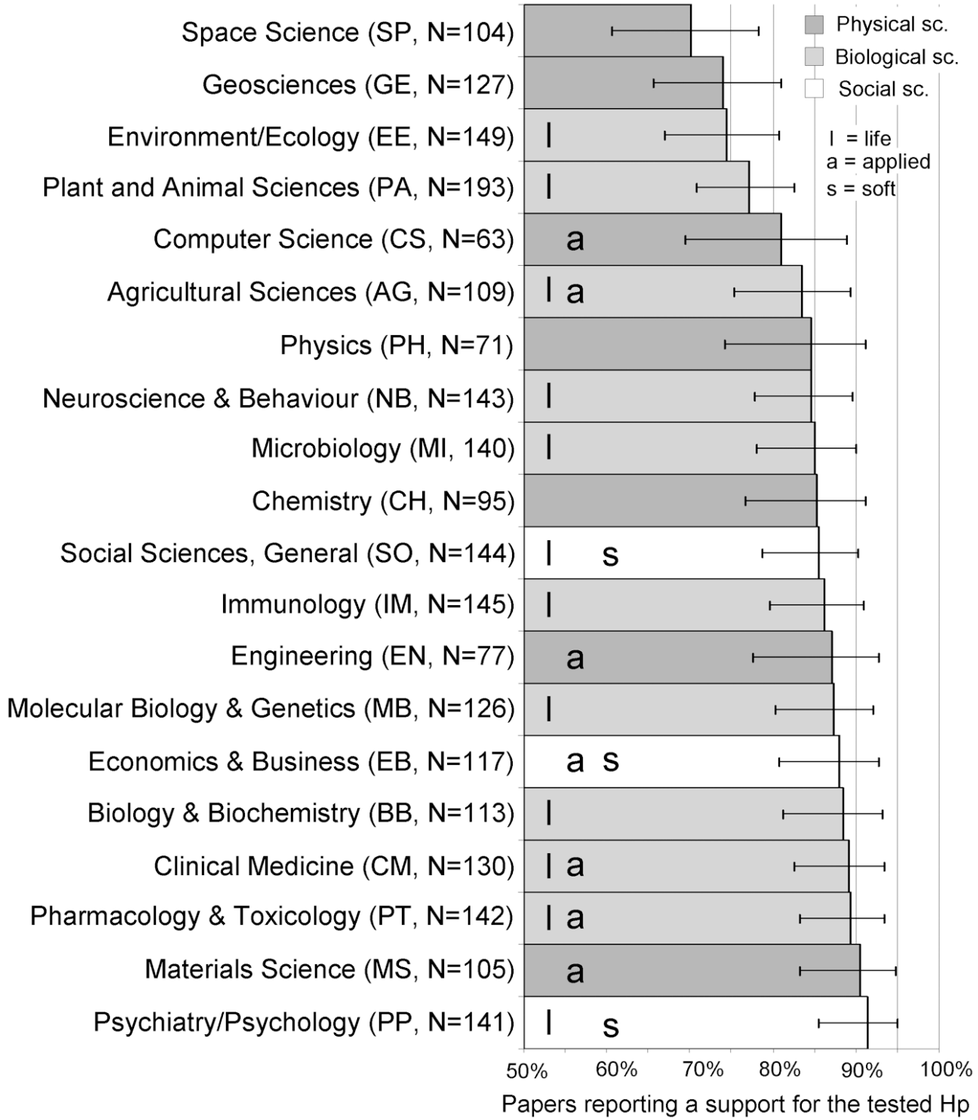
Bachelor end project by Mitchell Schijen

Positive result rates in Registered Reports vs. conventional reports

Goals:

1) Assess one indicator of publication bias in RRs using ‘quick & dirty’ method
2) Replicate Fanelli (2010) with a current sample
3) Examine validity of Fanelli’s search criteria (qualitative)

‘Positive‘ results increase down the hierarchy of sciences (Fanelli, 2010)



“The sentence **‘test* the hypothes*’** was used to search all 10837 journals in the Essential Science Indicators database, which classifies journals univocally in 22 disciplines. When the number of papers retrieved from one discipline exceeded 150, papers were selected using a random number generator. … By **examining the abstract** and/or full-text, it was determined whether the authors of each paper had concluded to have found a positive (full or partial) or negative (null or negative) support. If more than one hypothesis was being tested, **only the first one to appear** in the text was considered. ”

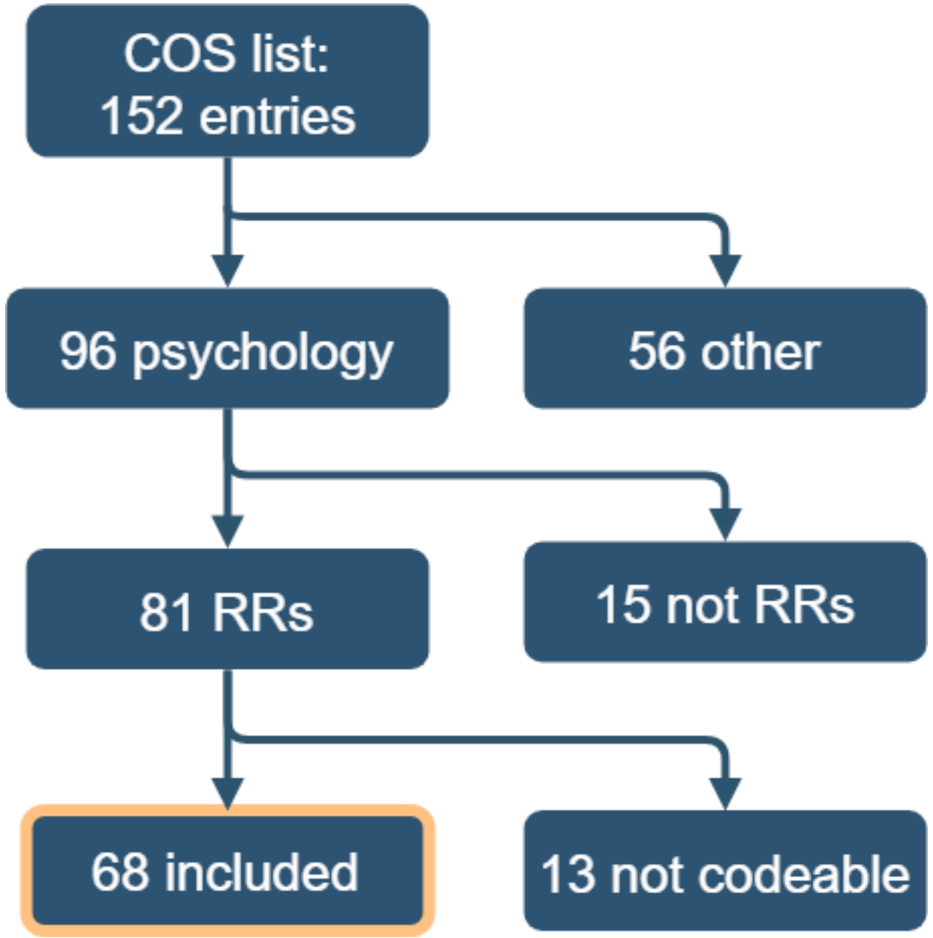
‘Positive‘ results increase down the hierarchy of sciences (Fanelli, 2010)

- psychology journals
- ‘test* the hypothes*’
- first hypothesis only
- positive (full/partial) vs. negative (null/negative) support

“The sentence **‘test* the hypothes*’** was used to search all 10837 journals in the Essential Science Indicators database, which classifies journals univocally in 22 disciplines. When the number of papers retrieved from one discipline exceeded 150, papers were selected using a random number generator. … By **examining the abstract** and/or full-text, it was determined whether the authors of each paper had concluded to have found a positive (full or partial) or negative (null or negative) support. If more than one hypothesis was being tested, **only the first one to appear** in the text was considered.”

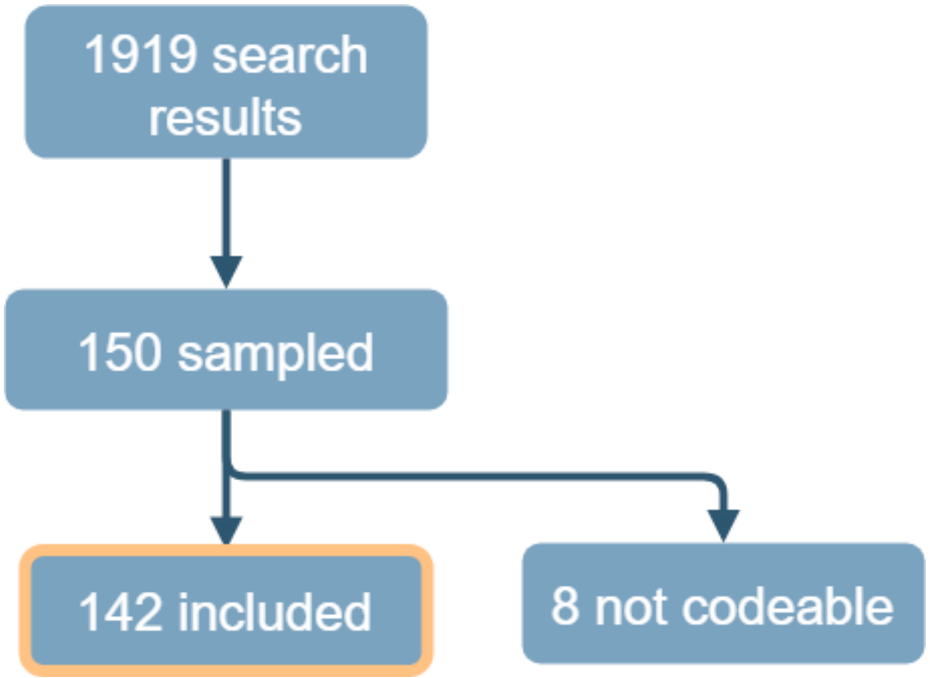
RRs

RR database curated by the Center for Open Science (Nov. 2018)

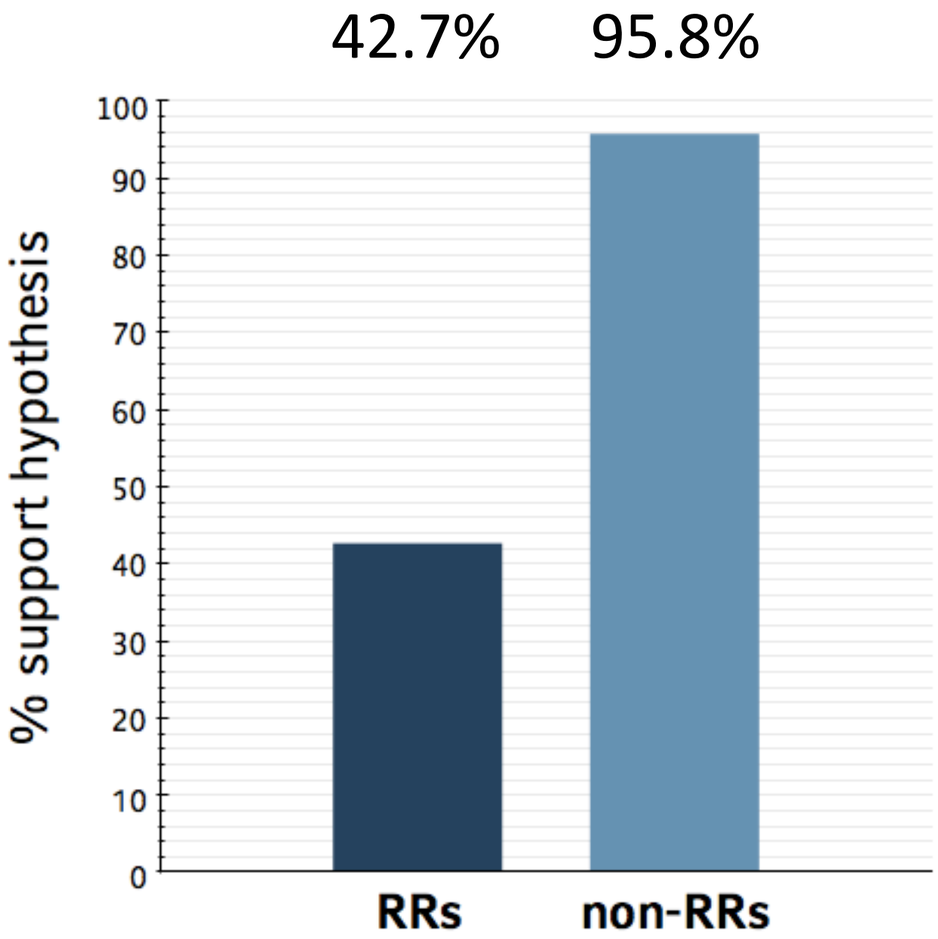


non-RRs

searched 633 psychology/psychiatry journals from ESI database (2013-2018) for **'test* the hypothes*'**



Confirmatory results



H_0: RRs ≥ non-RRs

 reject ($p < .0001$)

H_1: Difference > 6%

 accept ($p > .999$)

Inter-rater reliability:

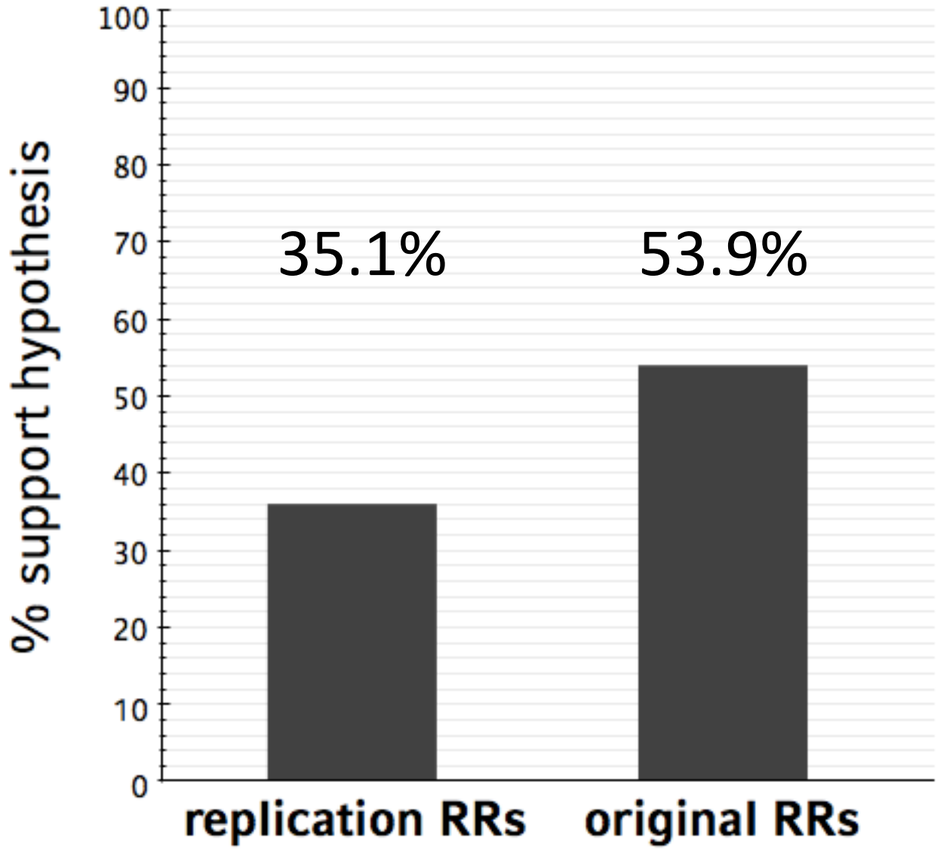
45/210 double-coded

Cohen's kappa = .933

disagreements resolved by discussion

Comparison to Fanelli (2010): **91.5%**

Exploratory results: Replications



42/68 RRs contain replication

26/68 RRs contain only original work

Fisher's exact test: $p = .139$

No non-RR was a replication

Caution: replication status was coded very superficially (not per hypothesis)!

Qualitative results:

Hypothesis introductions

‘test* the hypothes*’ almost never used

“examine”	“has examined the hypothesis”
	“to examine whether” (2x)
	“to examine”
	“we examined whether” (2x)
	“to critically examine and replicate”
	“our goal was to examine”
	“A large, controversial literature has examined the hypothesis that”
“test* … whether”	“this study tested whether”
	“to test”
	“testing whether” (2x)
	“we tested whether”
	“to test whether [hypothesis]”
“test* … hypothes*”	“we tested the following hypotheses” (2x)
	“we test the hypothesis that”
	“we tested the hypothesis”
	“to empirically test”
	“the present study provides a critical test of”
	“This study tested whether”
	“to test an … hypothesis”
“we hypothesized”	“we hypothesized” (6x)
	“we hypothesized that”
	“hypothesized”
“we predicted”	“we predicted”
	“we predicted that” (2x)
“the hypothes* predicted”	“the hypotheses predicted”
	“we had three predictions”

Qualitative results:

Hypothesis introductions (replications)

Replications often not phrased as hypothesis tests

“sought to replicate”	“we sought to replicate” “we sought to replicate the finding” “we sought to replicate these effects” “this study sought to replicate the findings of …”
“we performed a(n) replication”	“we performed a … replication” “we performed a(n) … replication”
“we report the results of … replication*”	“we report the results of … replication” “we report the results of … replications”
“we conducted … replication*”	“we conducted … replications” “we conducted … replications” “we conducted a replication” “we conducted replications”
“replicat*”	“we replicated” “to critically examine and replicate” “we aim at replicating” “the present work includes … attempts to replicate” “replications of … [the] idea that”
Did not use “replication’/’replicate”	12x

Coding difficulties: Example

Hypothesis:
"A large, controversial literature has examined the hypothesis that the attractiveness of potential partners predicts romantic desire more strongly for men than for women."

Finding:
"The sex difference emerged with objective assessments of attractiveness from independent raters (approximately $q = .13$, a small effect) but not with participants' own assessments of attractiveness ($q = .00$)."

Conclusions

Positive result rate in RRs drastically lower than in non-RRs

Fanelli's search phrase may not represent hypothesis tests in psychology very well
 should be replicated with other search terms

Method is relatively easy to use; non-RR results replicate

Limitations:

- Result says little about causes
- Only one hypothesis per paper
- Coding may depend on expertise

Thank you!

Co-authors:

Mitchell Schijen & Daniël Lakens

Contact me:

a.m.scheel@tue.nl

@AnneMScheel