

# Are registered reports an effective method to counter p-hacking?



Andreas Holst  
Supervisor: Bram Duyx, PhD

# Replication Crisis in Psychology

Open Science Collaborations (2015) article estimated that only 39% of psychological research can be replicated

- Statistics would predict 5% with significance level at  $\alpha = 0.05$
- Researchers degrees of freedom (Simonsohn et al., 2011)
  - Increases the chances of finding false-positive results and overinflated effect sizes
    - 34 items of potential degrees of freedom (Wicherts et al., 2016)
    - p-hacking - data analysis and eligibility decisions
- Selective publishing of significant results by journals
  - Survival of the fittest
  - Publication bias - most published results are significant.
    - file-drawer problem
  - Researchers Degrees of Freedom + perverse incentives

# Registered reports

- Pre-registered reports sent to journals to be peer-reviewed (Center for Open Science)
  - Before collecting and analyzing data
  - Just 190 reports since 2013
  - Pre-registration requires researchers to disclose their methods in advance



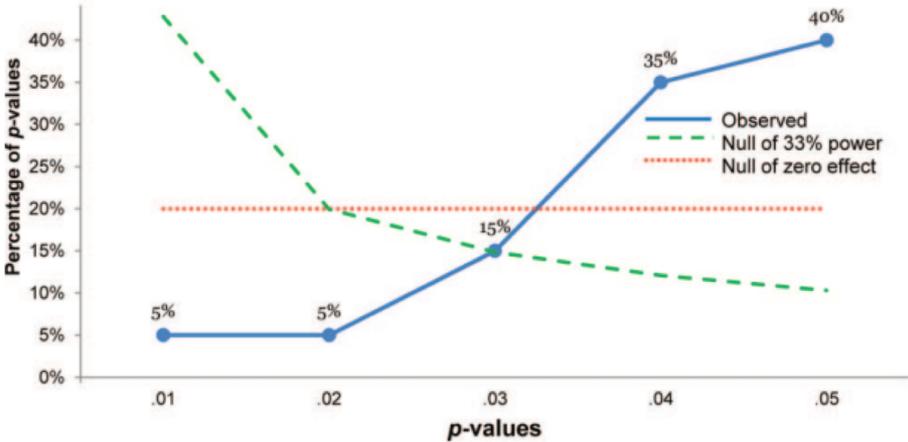
- Journals base their decision of publication on the relevance of research question and quality of the research design, not results.
  - Motivator to follow improved guidelines and disincentivizes p-hacking.

# The p-curve (Simonsohn, Nelson, Simmons)

- Estimating the evidential value of a meaningful set of findings
  - Distribution of statistically significant p-values
    - Avoids the effects of publication bias on the sample
  
- There is evidential value if selective reporting can be ruled out as the sole reason for the results.
  - Estimated by the skewness of the graph
    - Inferred with 3 statistical tests
      - Test for right skew
      - Test for flat right skew with 33% power
      - Power analysis

We expected these experiments to have been *p*-hacked

A



**Statistical Inference**

- 1) Studies contain evidential value  
*(right-skewed)*
- 2) Studies lack evidential value  
*(flatter than 33%)*
- 3) Studies lack evidential value and were intensely *p*-hacked?  
*(left-skewed)*

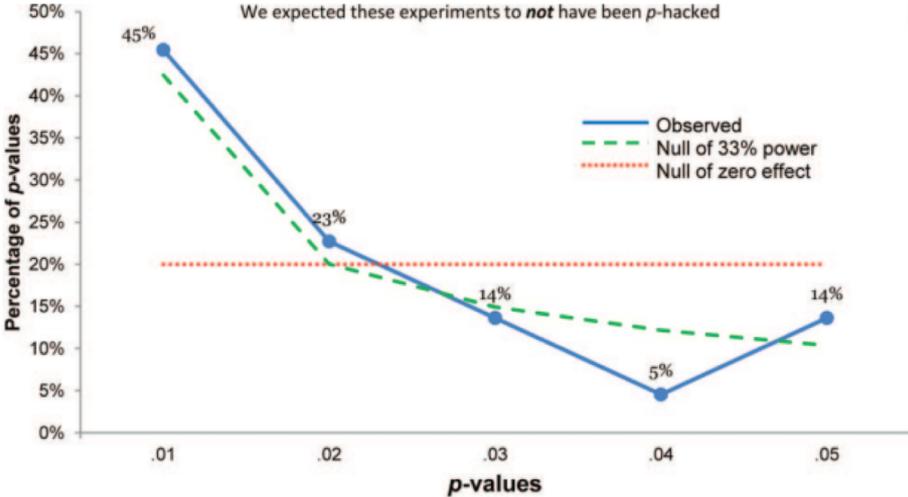
**Results**

1)  $\chi^2(40)=18.3, p=.999$   
 2)  $\chi^2(40)=82.5, p<.0001$   
 3)  $\chi^2(40)=58.2, p=.031$

The observed *p*-curve includes 20 significant *p*-values, an additional 3 were *p*>.05  
 Of those 20 *p*-values, 3 are *p*<.025, binomial test for right-skew: *p*>.999; for left-skew: *p*=.0013

We expected these experiments to **not** have been *p*-hacked

B



**Statistical Inference**

- 1) Studies contain evidential value  
*(right-skewed)*
- 2) Studies lack evidential value  
*(flatter than 33%)*
- 3) Studies lack evidential value and were intensely *p*-hacked?  
*(left-skewed)*

**Results**

1)  $\chi^2(44)=94.2, p<.0001$   
 2)  $\chi^2(44)=43.2, p=.507$   
 3)  $\chi^2(44)=27.2, p=.978$

The observed *p*-curve includes 22 significant *p*-values, an additional 3 were *p*>.05  
 Of those 22 *p*-values, 16 are *p*<.025, binomial test for right-skew: *p*=.026; for left-skew: *p*=.991.

# Study aim & Hypothesis

This study investigates if registered reports are an effective way to counter p-hacking using the p-curve.

- Allows to avoid the effect of publication bias for more accurate estimations

Hypothesis:

1. The p-curve associated with registered reports has a significant result for right skew.
2. The p-curve associated with C-group has a significant result for flat right skew expected at 33% power.

# Methods

## Quasi-experimental Design

### Confirmatory research

### Independent variable - publication type

- categorical, nominal
  - registered reports
  - normal publication

### Dependent variable - Evidential value

- categorical, ordinal
  - Set of studies contain evidential value
  - Set of studies needs further investigation
  - Lack of any evidential value
    - Set of studies were probably p-hacked

### Inclusion criteria:

- Only psychological research
  - Confirmatory
  - Experimental
  - Continuous dependent variable
- Inclusion criteria of p-values
  - uniform distribution under the null hypothesis
  - test relevant hypothesis
  - statistically independent of other p-values

### Exclusion criteria:

- journals publishing only one publication type. (matching)
- simultaneous recording devices.

# Methods

Selecting p-values:

1. Identify hypothesis and study design
2. Identify the appropriate statistical test
3. Report the result of interest
4. Recompute the precise p-value(s)
5. Report robustness results.

Following this process with every study in the sample will result in a standardized “p-curve disclosure table”

Matching algorithm:

1. Identify suitable independent p-values for the p-curve associated with registered reports in a public Center for Open Science Database.
  - a. Total of 190 studies
2. Find the articles in their original journal
3. Find a p-value for C-group keeping all publishing related variables constant besides publishing type
  - a. RR - center
  - b. First above article, then under
  - c. Does it match inclusion-exclusion criteria?

# Statistical analysis

	<b>Binomial Test</b> <i>(Share of results <math>p &lt; .025</math>)</i>	<b>Continuous Test</b> <i>(Aggregate with Stouffer Method)</i>	
		<b>Full p-curve</b> <i>(<math>p</math>'s <math>&lt; .05</math>)</i>	<b>Half p-curve</b> <i>(<math>p</math>'s <math>&lt; .025</math>)</i>
1) Studies contain evidential value. <i>(Right skew)</i>	$p = .0352$	$Z = -3.94, p < .0001$	$Z = -3.38, p = .0004$
2) Studies' evidential value, if any, is inadequate. <i>(Flatter than 33% power)</i>	$p = .9344$	$Z = 1.83, p = .9664$	$Z = 3.74, p = .9999$
		<b>Statistical Power</b>	
Power of tests included in $p$ -curve <i>(correcting for selective reporting)</i>		Estimate: 73%	
		90% Confidence interval: (38% , 92%)	

# Limitations and Questions

Only continuous dependent variable underlying the p-value

Only experimental designs

Confirmatory research

Is the p-curve a valid measure?

Is it an accurate measure?

- Can the p-curve distinguish well enough between the levels of the dependent variable?

Is there a difference in power analysis and a test for left skew?

Thank you for listening!