Assessing the Robustness of Statistical Results to Data Exclusion in t-tests

Tsz Keung Wong

Tilburg University

Understanding Society

Main Message

1. Data exclusion is common

2. Different types of data exclusion:
   - Randomized data exclusion (MCAR)
   - P-hacking (MNAR)
   - Research misconduct - Extreme data exclusion (MNAR)

3. Potential effect of data exclusion on meta-analysis:
   - Bias in effect size estimates and testing result

Main Message

1. Data exclusion is common

Main Message

1. Data exclusion is common

Questionable Research Practices (QRP): A range of activities that intentionally or unintentionally distort data in favour of a researcher's own hypotheses

John, Loewestein, and Prelec (2012) and Agnoli et al. (2017)

Table 1. Questionable Research Practices (QRPs) and self-admission rates in percentages for US [25] and Italian psychologists.

	US		Italian Association of Psychology	
QRP	**Self-admission rate (*N*)**	**95% CI**	**Self-admission rate (*N*)**	**95% CI**
1. In a paper, failing to report all of a study's dependent measures	63.4 (486)	59.1–67.7	47.9 (219)	41.3–54.6
2. Deciding whether to collect more data after looking to see whether the results were significant	55.9 (490)	51.5–60.3	53.2 (222)	46.6–59.7
3. In a paper, failing to report all of a study's conditions	27.7 (484)	23.7–31.7	16.4 (219)	11.5–21.4
4. Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6 (499)	12.4–18.8	10.4 (221)	6.4–14.4
5. In a paper, "rounding off" a *p* value (e.g., reporting that a *p* value of .054 is less than .05)	22.0 (499)	18.4–25.7	22.2 (221)	16.7–27.7
6. In a paper, selectively reporting studies that "worked"	45.8 (485)	41.3–50.2	40.1 (217)	33.6–46.6
7. Deciding whether to exclude data after looking at the impact of doing so on the results	38.2 (484)	33.9–42.6	39.7 (219)	33.3–46.2
8. In a paper, reporting an unexpected finding as having been predicted from the start	27.0 (489)	23.1–30.9	37.4 (219)	31.0–43.9
9. In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)	3.0 (499)	1.5–4.5	3.1 (223)	0.9–5.4
10. Falsifying data	0.6 (495)	0.0–1.3	2.3 (220)	0.3–4.2

Note: Confidence intervals for US psychologists were computed from data provided by Leslie John.

Main Message

1. Data exclusion is common

2. Different types of data exclusion :
   - Randomized data exclusion (MCAR)
   - P-hacking (MNAR)
   - Research misconduct - Extreme data exclusion (MNAR)

Statistical Result

Two hundred and twenty-four participants (N=224) were recruited via Amazon's Mechanical Turk. When asked whether they had taken the study seriously, **11 participants** indicated that they had not and were therefore removed from analysis, yielding a final sample of **213 participants**

Statistical Result

There was a significant effect of condition, **t(211) = 2.02, p = .045, d = 0.37.** As predicted, participants who had previously read about the immoral act rated the debate and research on automaticity more negatively **(M = 4.29, SD = 1.05)** than those who read about the morally neutral act **(M = 4.68, SD = 1.04).**

Statistical Result

There was a significant effect of condition, **t(211) = 2.02, p = .045, d = 0.37.** As predicted, participants who had previously read about the immoral act rated the debate and research on automaticity more negatively **(M = 4.29, SD = 1.05)** than those who read about the morally neutral act **(M = 4.68, SD = 1.04).**

"How robust is this result to data exclusion?"

Type of Data Exclusion

- Randomized data exclusion
- P-hacking - Selective randomized data exclusion
- Misconduct -Extreme data exclusion

Needed Information for the result t(211) = 2.02

Sample size per group

$N_1 = 41$

$N_2 = 183$

Data exclusion per group

$m_1 = 6$

$m_2 = 5$

Sample size after exclusion

$n_1 = 35$

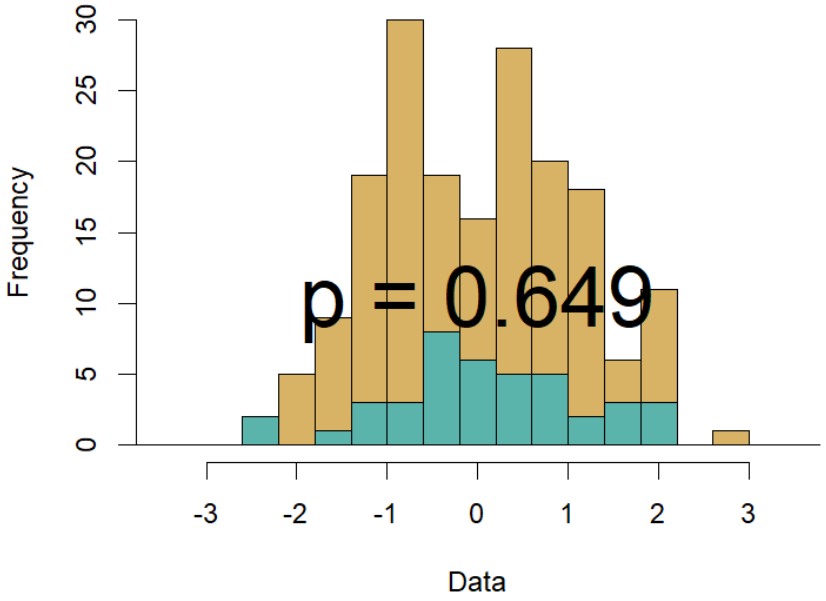
$n_2 = 178$

Randomized data exclusion

Simulating data

$N_1 = 41$

$N_2 = 183$



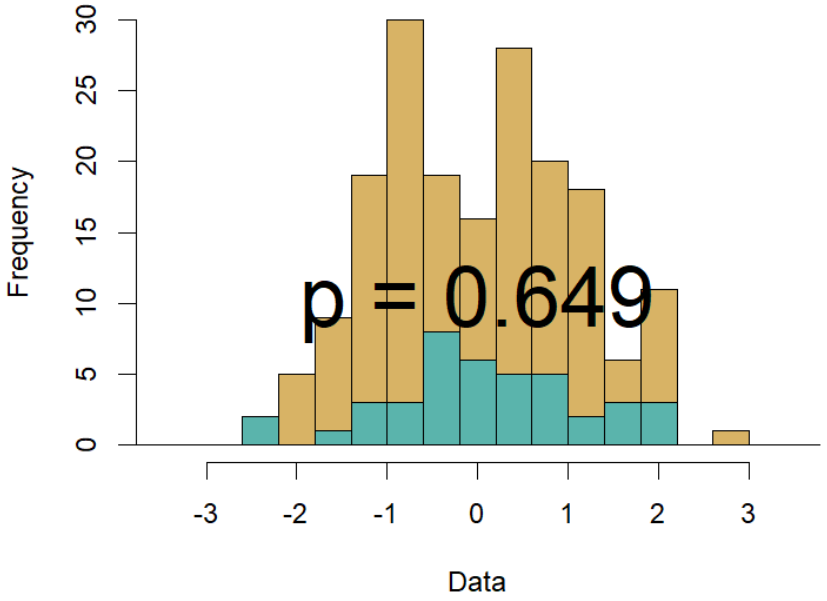
TILBURG UNIVERSITY

Randomized data exclusion

Simulating data

$N_1 = 41$

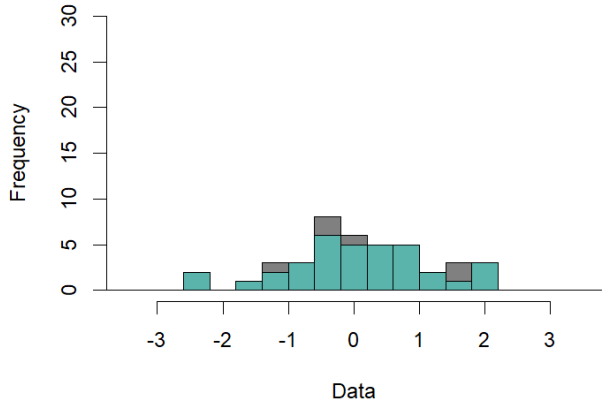
$N_2 = 183$

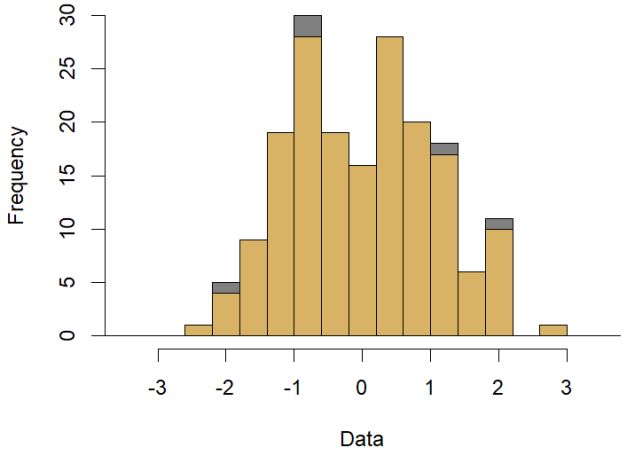


Data exclusion

$m_1 = 6$

$m_2 = 5$





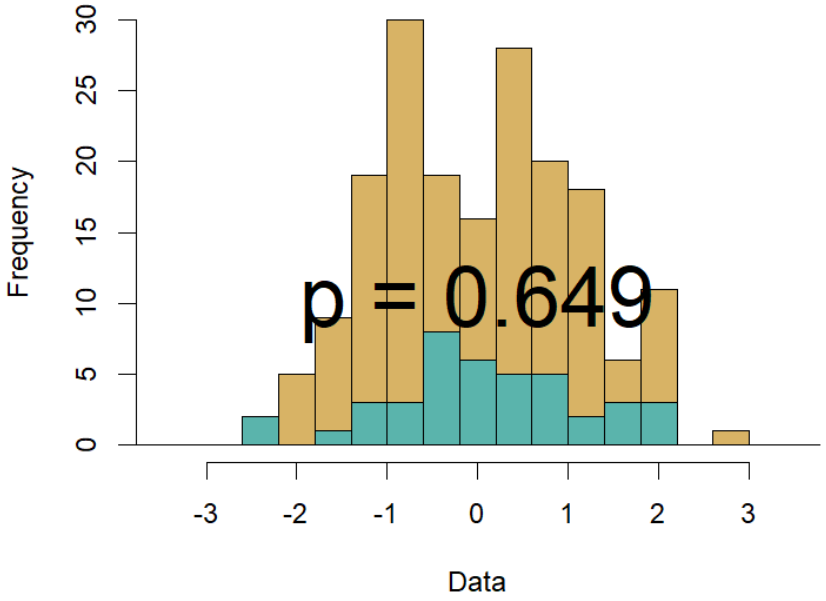
TILBURG UNIVERSITY

Randomized data exclusion

Simulating data

$N_1 = 41$

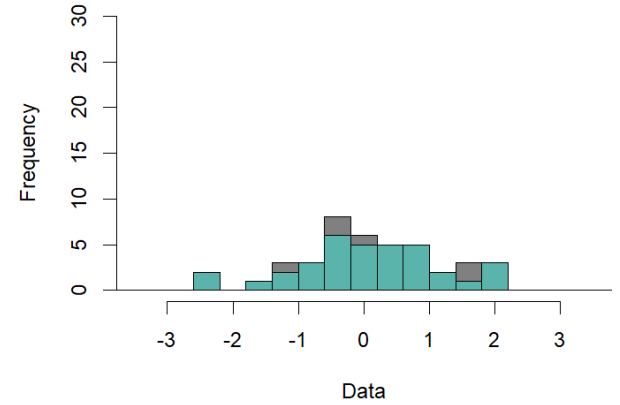
$N_2 = 183$

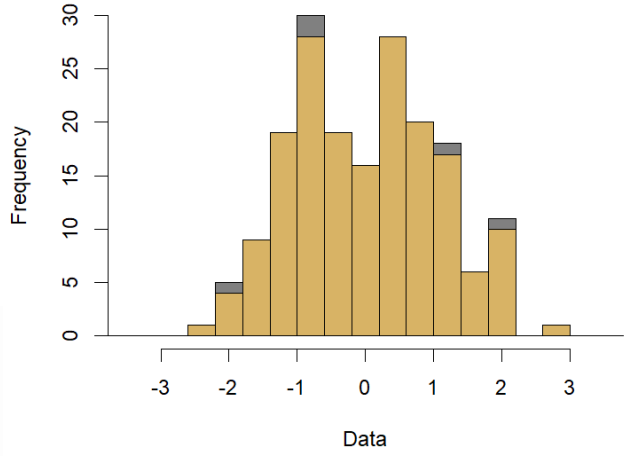


Data exclusion

$m_1 = 6$

$m_2 = 5$

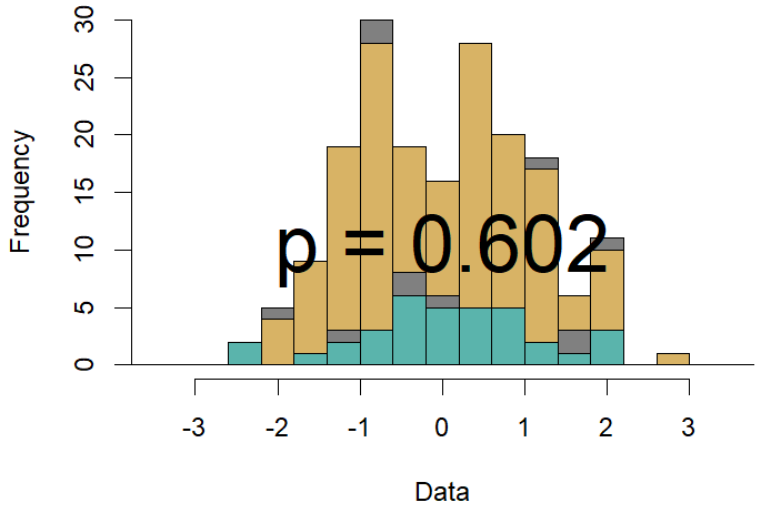




testing

$n_1 = 35$

$n_2 = 178$



TILBURG UNIVERSITY

P(t> 2.02) OR P(p-value < reported p-value .045)

Randomized data exclusion

- Given the null is true: 2%
- Given a just insignificant result:

P-hacking - Selective randomized data exclusion

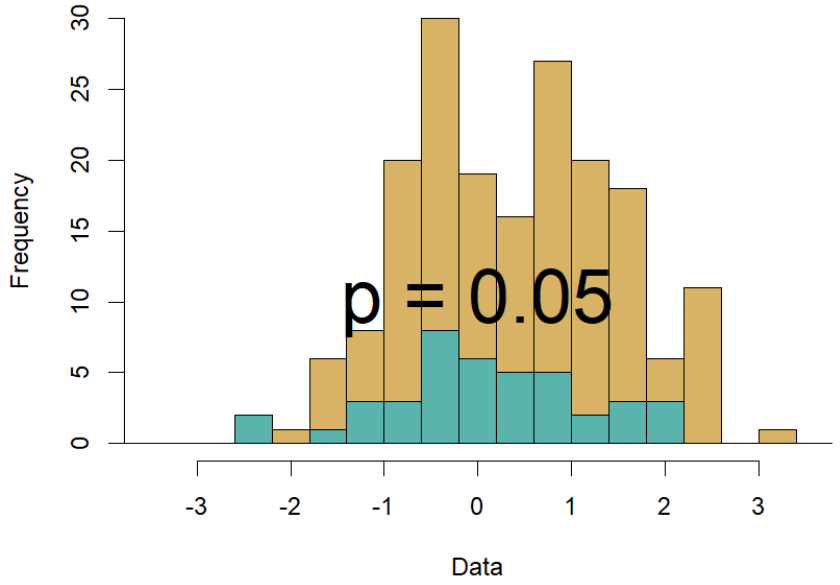
- Given the null is true:
- Given a just insignificant result:

Misconduct -Extreme data exclusion

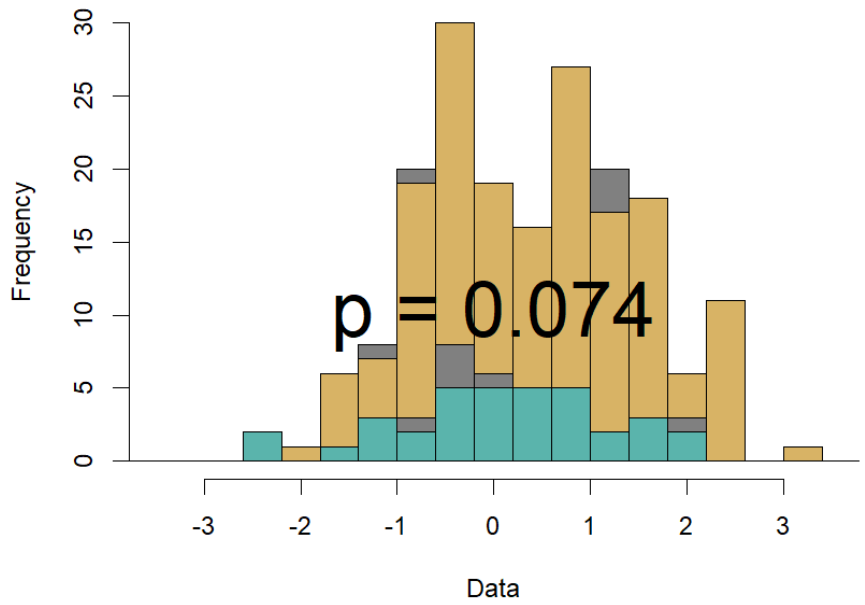
- Given the null is true:
- Given a just insignificant result:

Randomized data exclusion

Simulating data



Excluding and testing the data



TILBURG UNIVERSITY

P(t> 2.02) OR P(p-value < reported p-value .045)

Randomized data exclusion

- Given the null is true: 2%
- Given a just insignificant result: 4%

P-hacking - Selective randomized data exclusion

- Given the null is true:
- Given a just insignificant result:

Misconduct -Extreme data exclusion

- Given the null is true:
- Given a just insignificant result:

P(t> 2.02) OR P(p-value < reported p-value .045)

Randomized data exclusion

- Given the null is true: 2%
- Given a just insignificant result: 4%

P-hacking - Selective randomized data exclusion

- Given the null is true:
- Given a just insignificant result:

Misconduct -Extreme data exclusion

- Given the null is true:
- Given a just insignificant result:

Selective randomized data exclusion

Goal: testing the mean difference between DV1 and DV2

DV1	DV2	Education level
1	2	1
3	1	1
2	8	2
3	4	2
2	1	3
7	8	3

Selective randomized data exclusion

Goal: testing the mean difference between DV1 and DV2

DV1	DV2	Education level
~~1~~	~~2~~	~~1~~
~~3~~	~~1~~	~~1~~
2	8	2
3	4	2
2	1	3
7	8	3

TILBURG UNIVERSITY

Selective randomized data exclusion

Goal: testing the mean difference between DV1 and DV2

DV1	DV2	Education level
1	2	1
3	1	1
2	8	2
3	4	2
2	1	3
7	8	3

Tilburg University

Selective randomized data exclusion

Simulating data

$N_1 = 41$

$N_2 = 183$

Exclusion

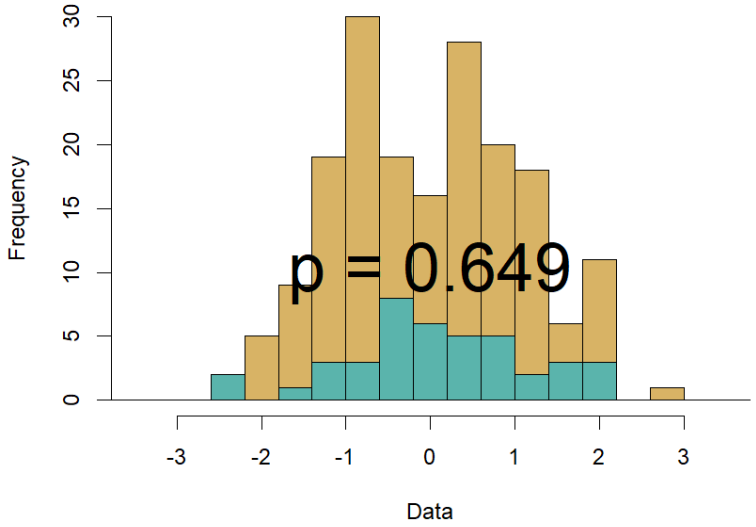
$m_1 = 6$

$m_2 = 5$

Selecting and testing

$n_1 = 35$

$n_2 = 178$



TILBURG UNIVERSITY

Selective randomized data exclusion

Simulating data

$N_1 = 41$

$N_2 = 183$

Exclusion

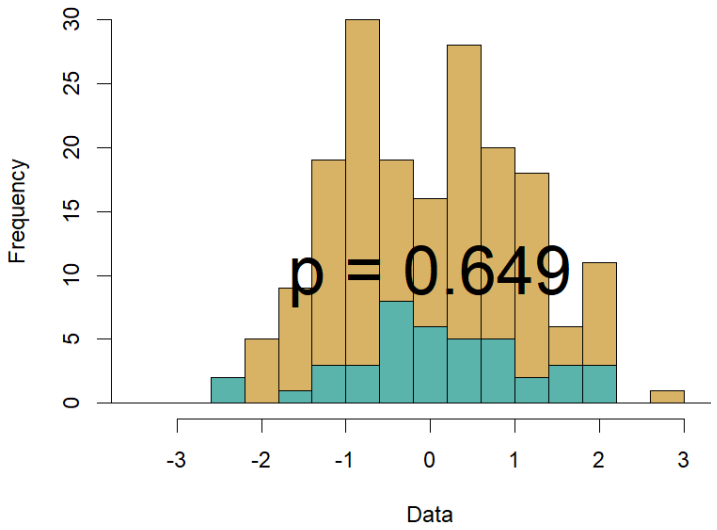
$m_1 = 6$

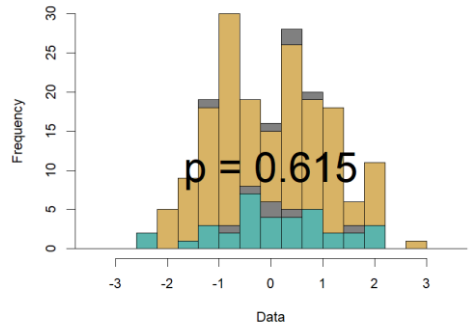
$m_2 = 5$

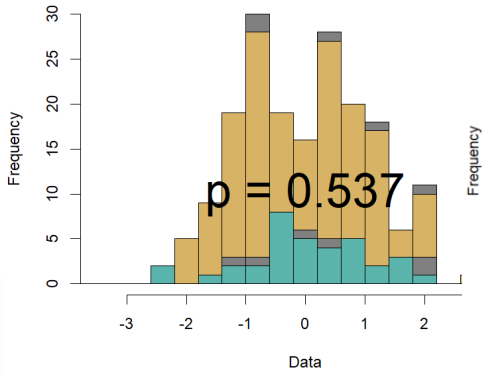
Selecting and testing

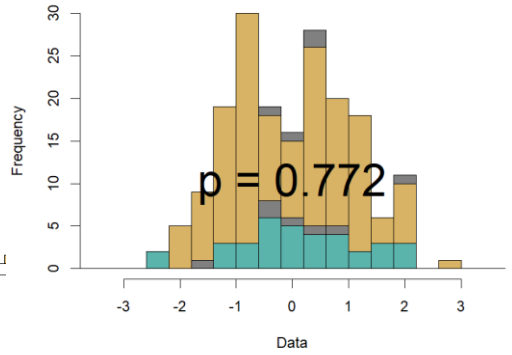
$n_1 = 35$

$n_2 = 178$









TILBURG UNIVERSITY

Selective randomized data exclusion

Simulating data

$N_1 = 41$

$N_2 = 183$

Exclusion

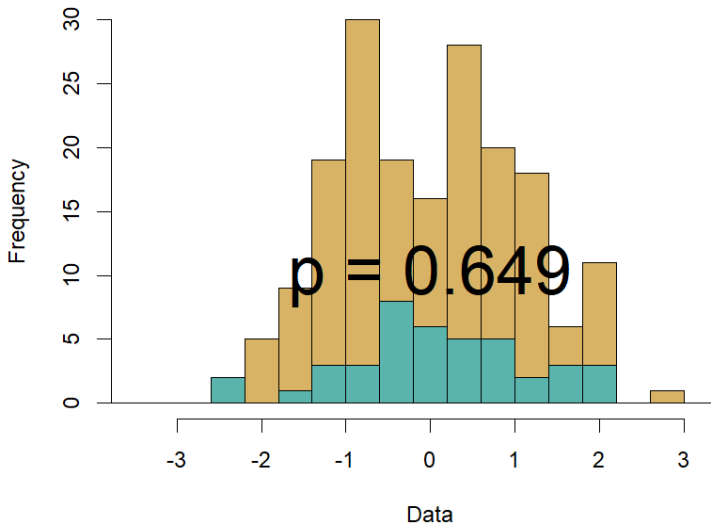
$m_1 = 6$

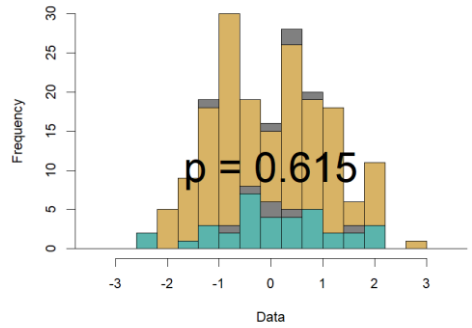
$m_2 = 5$

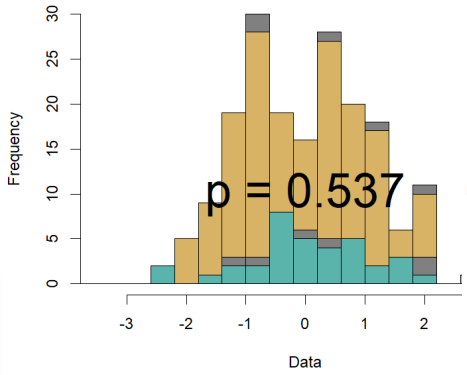
Selecting and testing

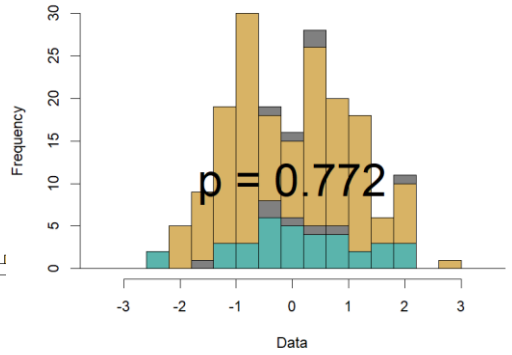
$n_1 = 35$

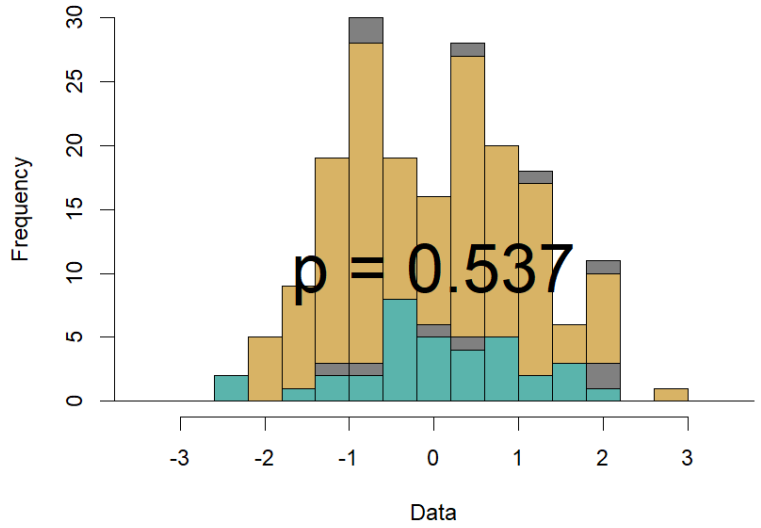
$n_2 = 178$











TILBURG UNIVERSITY

P(t> 2.02) OR P(p-value < reported p-value .045)

Randomized data exclusion

- Given the null is true: 2%
- Given a just insignificant result: 4%

P-hacking - Selective randomized data exclusion

- Given the null is true: 9%
- Given a just insignificant result:

Misconduct -Extreme data exclusion

- Given the null is true:
- Given a just insignificant result:

P(t> 2.02) OR P(p-value < reported p-value .045)

Randomized data exclusion

- Given the null is true: 2%
- Given a just insignificant result:

P-hacking - Selective randomized data exclusion

- Given the null is true: 9%
- Given a just insignificant result: 25%

Misconduct -Extreme data exclusion

- Given the null is true:
- Given a just insignificant result:

P(t> 2.02) OR P(p-value < reported p-value .045)

Randomized data exclusion

- Given the null is true: 2%
- Given a just insignificant result:

P-hacking - Selective randomized data exclusion

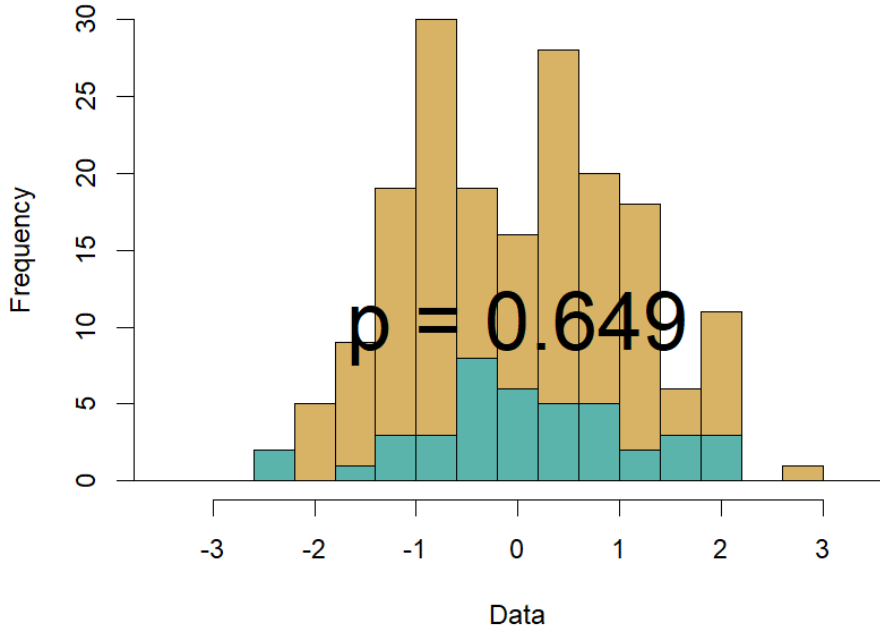
- Given the null is true:
- Given a just insignificant result:

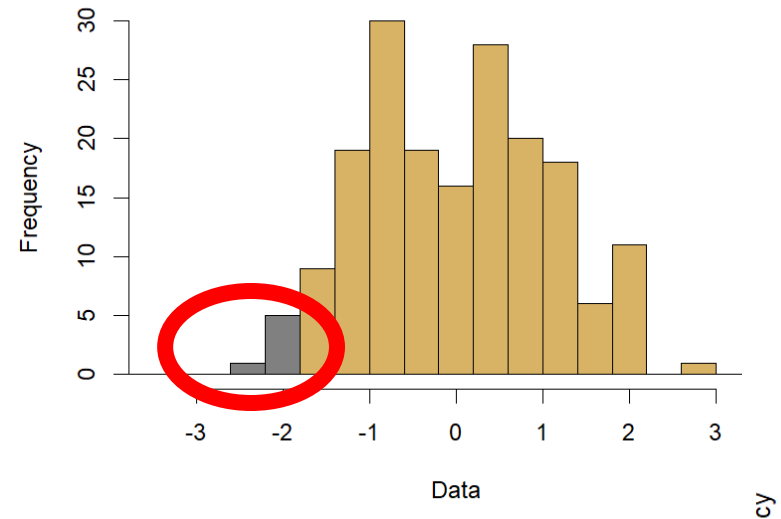
Misconduct -Extreme data exclusion

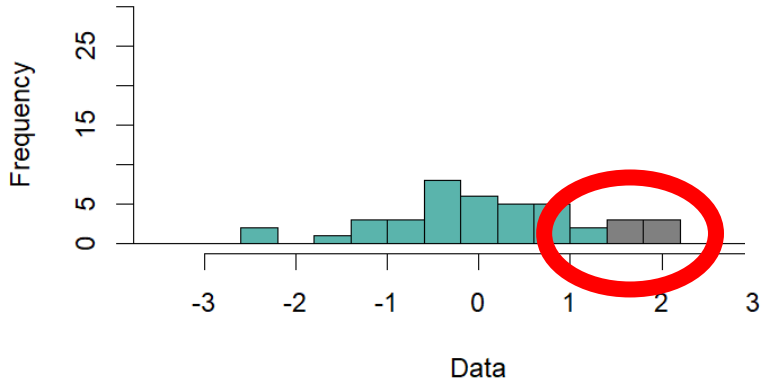
- Given the null is true:
- Given a just insignificant result:

Extreme data exclusion

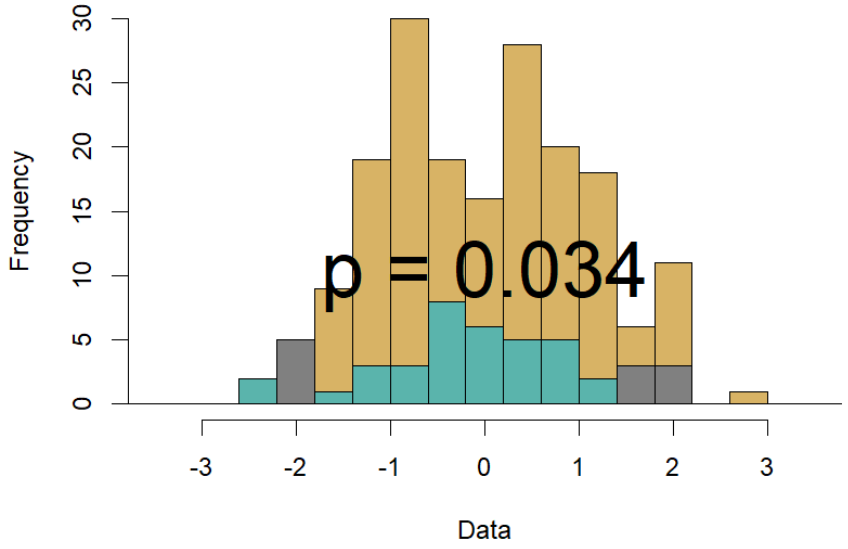
$N_1 = 41$
$N_2 = 183$







$n_1 = 35$
$n_2 = 178$



TILBURG UNIVERSITY

P(t> 2.02) OR P(p-value < reported p-value .045)

Randomized data exclusion

- Given the null is true: 2%
- Given a just insignificant result:

P-hacking - Selective randomized data exclusion

- Given the null is true: 9%
- Given a just insignificant result: 25%

Misconduct - Extreme data exclusion

- Given the null is true: 47%
- Given a just insignificant result: 100%

P(t> 2.02) OR P(p-value < reported p-value .045)

Randomized data exclusion

- Given the null is true: 2%
- Given a just insignificant result: 4%

P-hacking - Selective randomized data exclusion

- Given the null is true: 9%
- Given a just insignificant result: 25%

Misconduct - Extreme data exclusion

- Given the null is true: 47%
- Given a just insignificant result: 100%

Original result from the authors

Before exclusion:

With all participants, t(222) = 1.78, p = .077

After exclusion:

There was a significant effect of condition, t(211) = 2.02, p = .045, d = 0.37.

Assessing the robustness of statistical results

Sample –

- 55 papers from Journal of Personality and Social Psychology in 2014

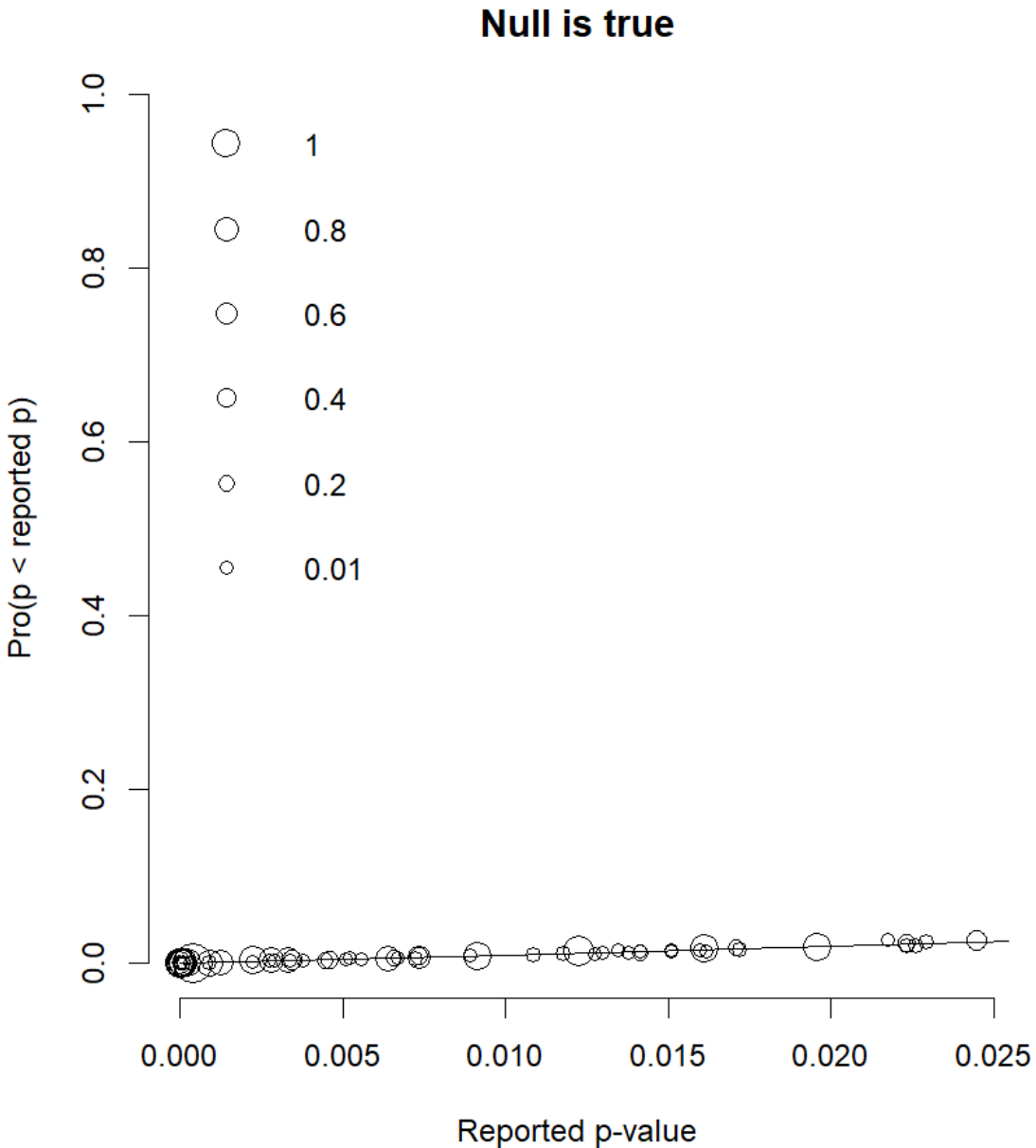
Targeted statistical result

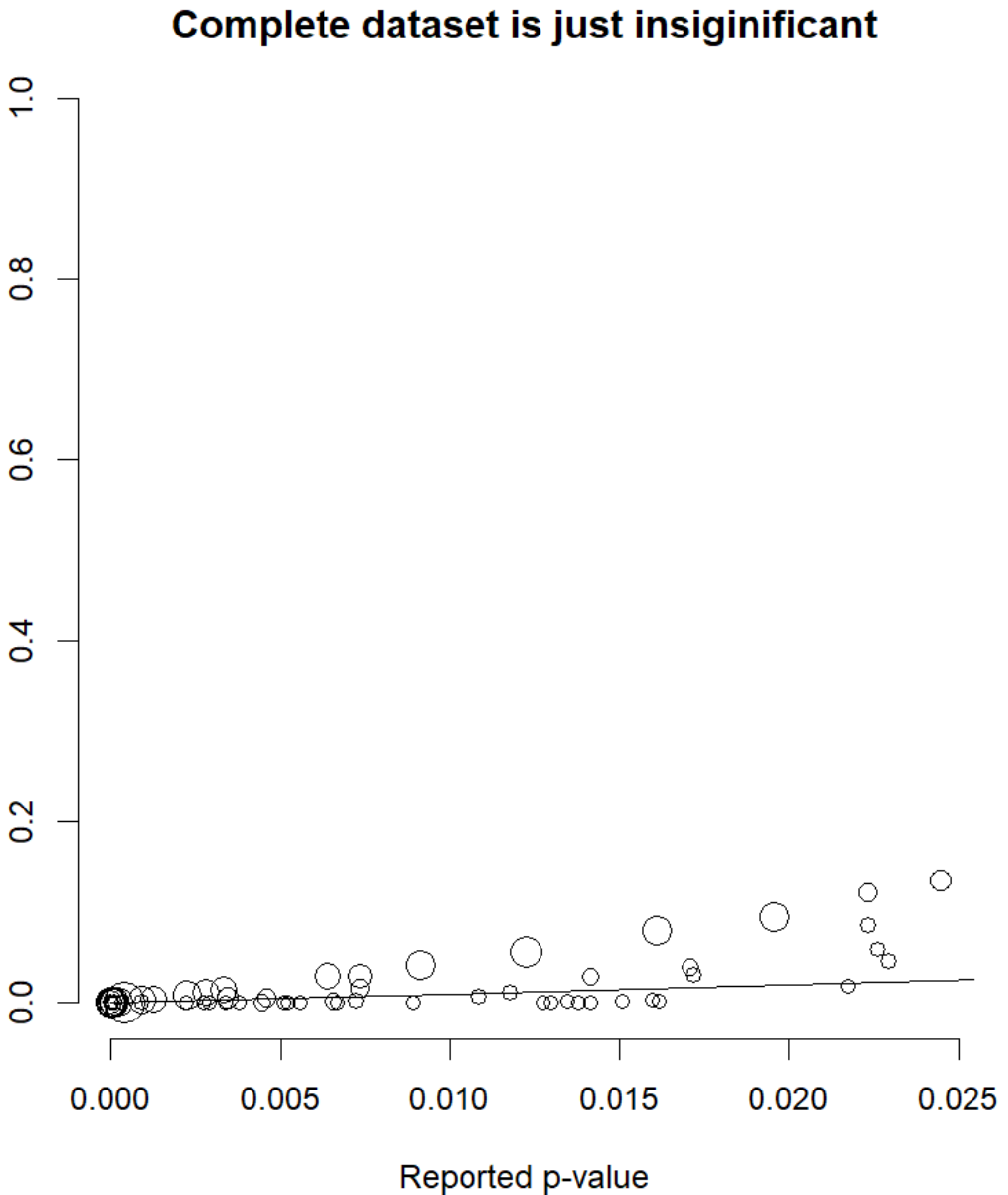
- One-sample or independent t-test
- Significant results
- Data exclusion

Selected cases:

- 77 are the targeted results

Randomized Data Exclusion

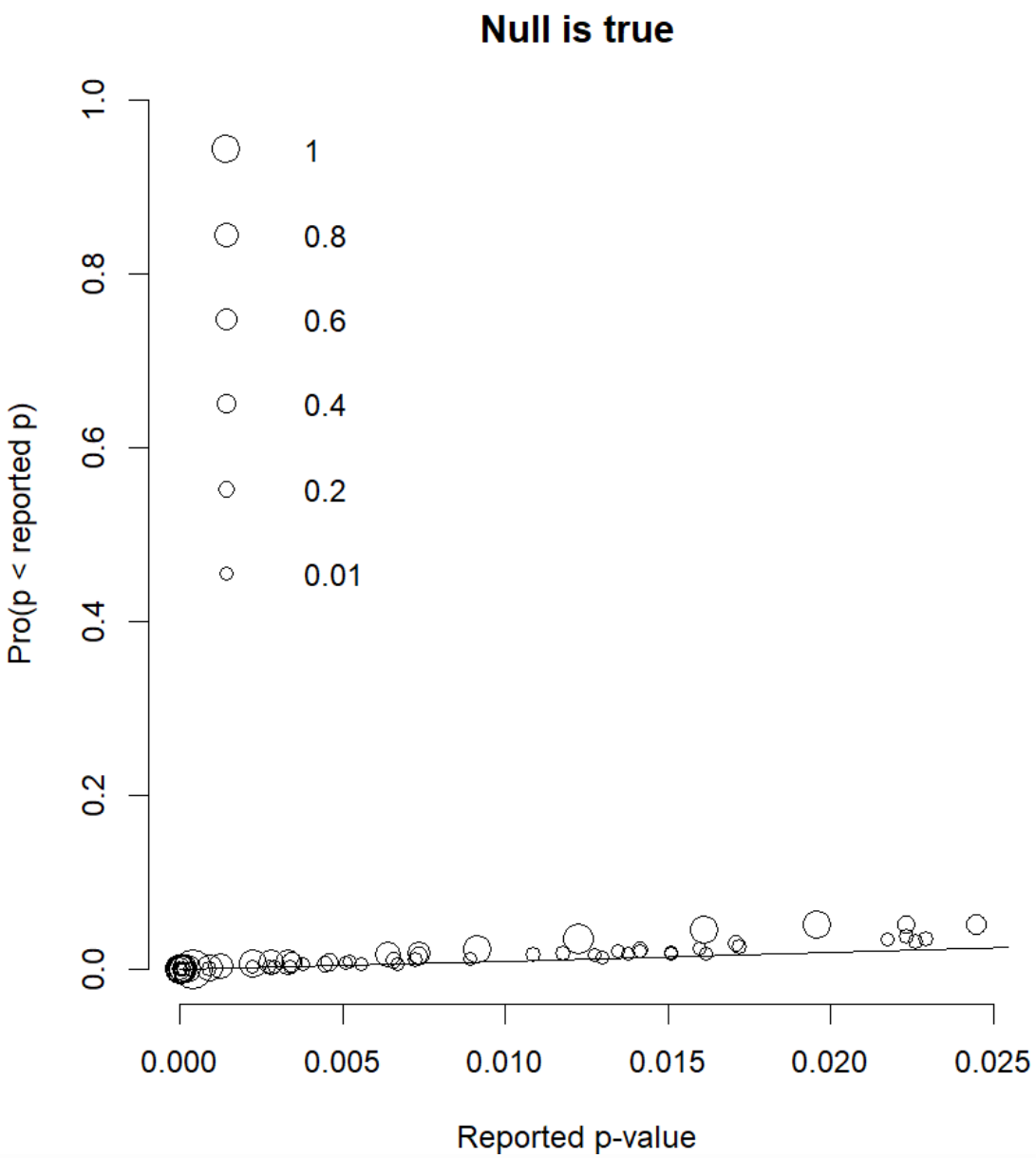


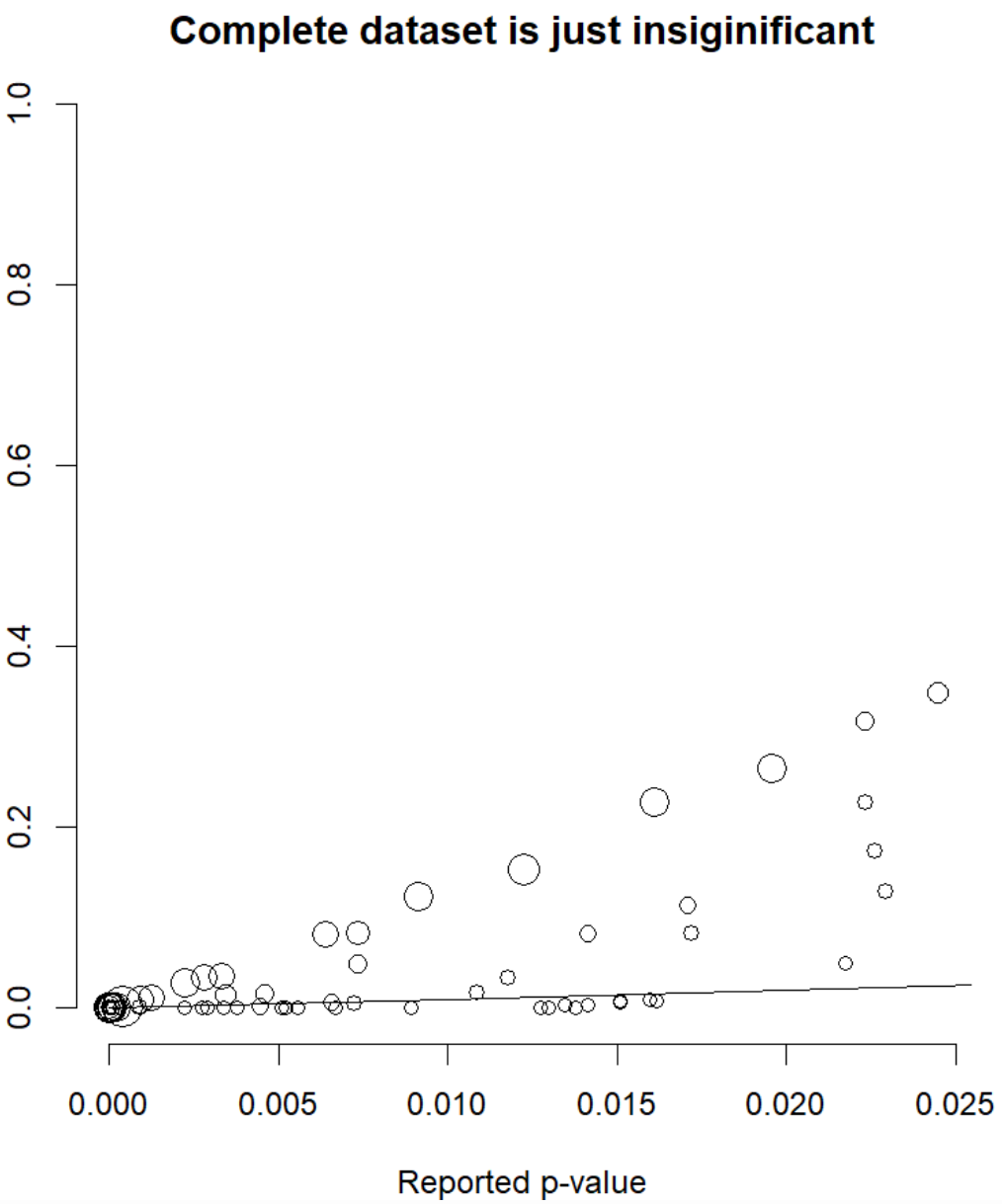


Main conclusion

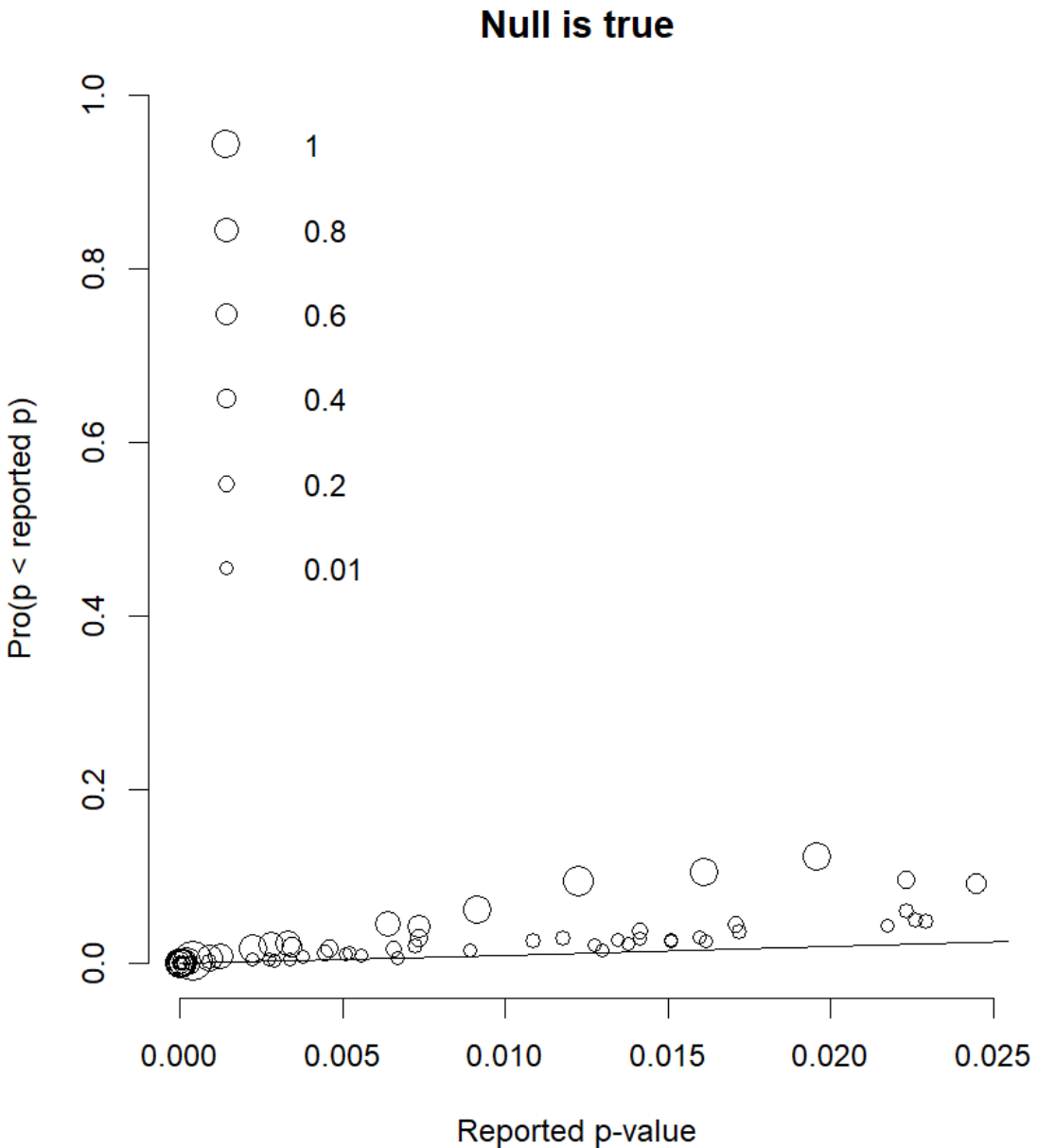
1. Statistical results are generally robust to Missing Completely At Random.

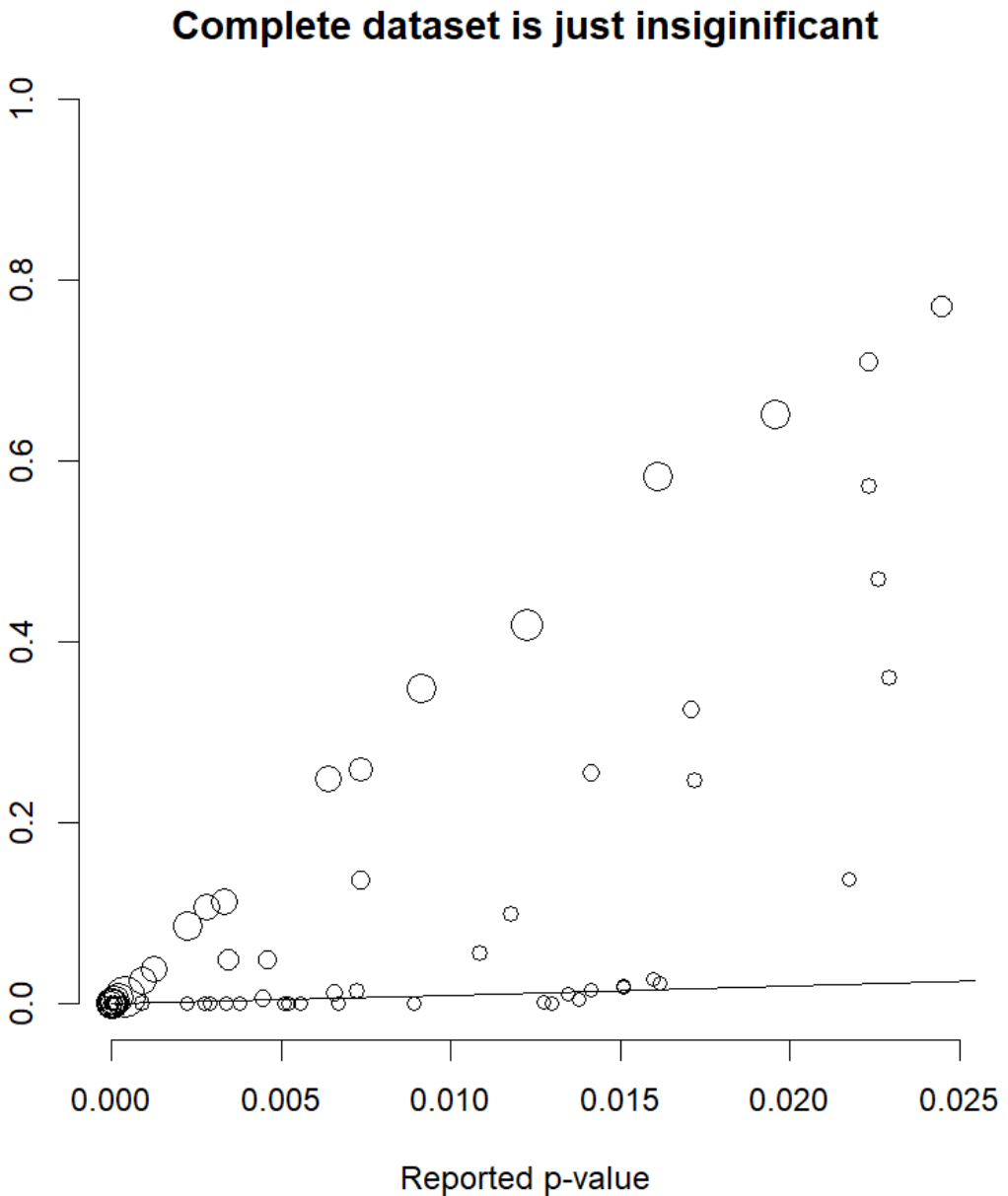
P-hacking (k =3)


Null is true
Pro(p < reported p)
1.0
0.8
0.6
0.4
0.2
0.0
1
0.8
0.6
0.4
0.2
0.01
0.000
0.005
0.010
0.015
0.020
0.025
Reported p-value


Complete dataset is just insiginificant
1.0
0.8
0.6
0.4
0.2
0.0
0.000
0.005
0.010
0.015
0.020
0.025
Reported p-value

P-hacking (k =10)



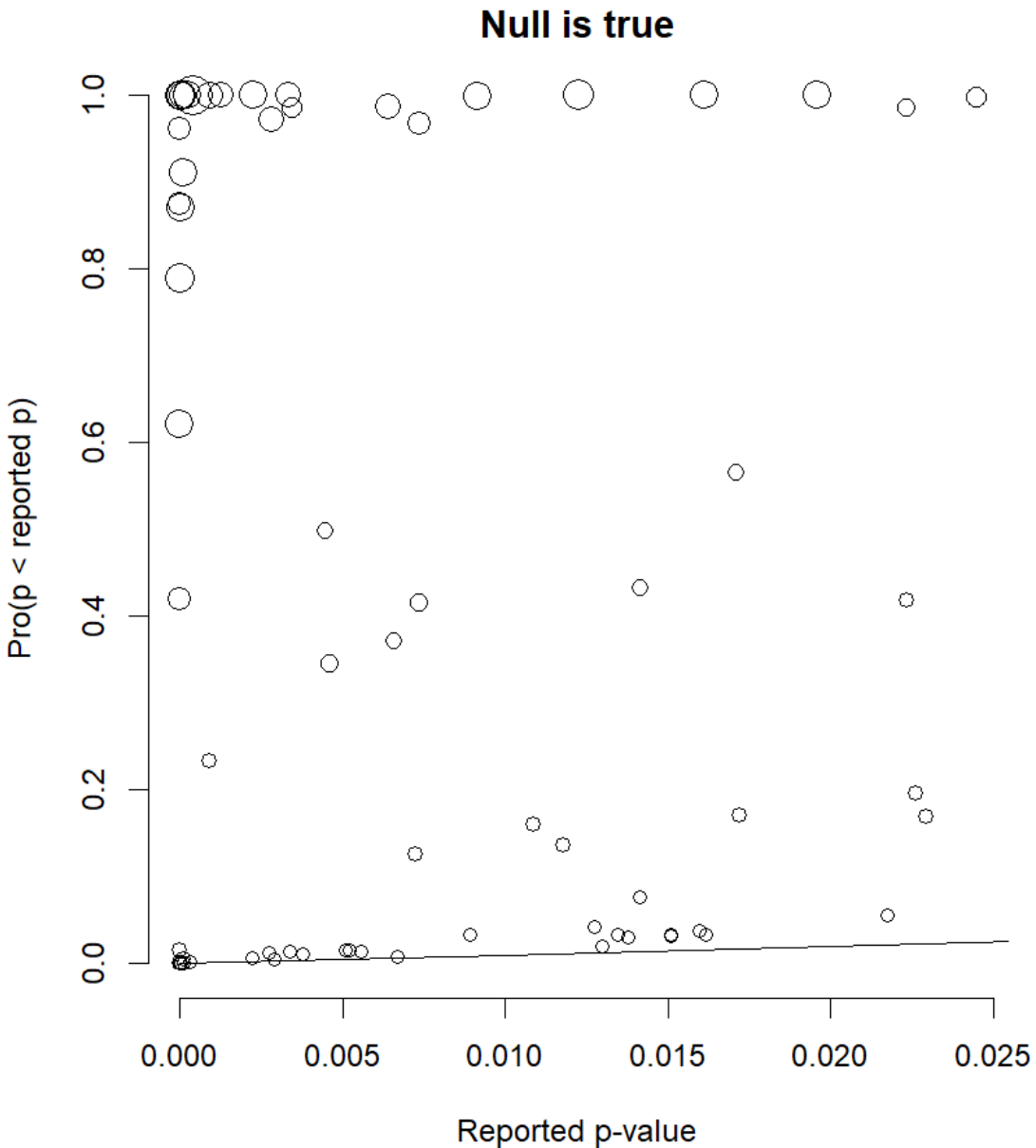


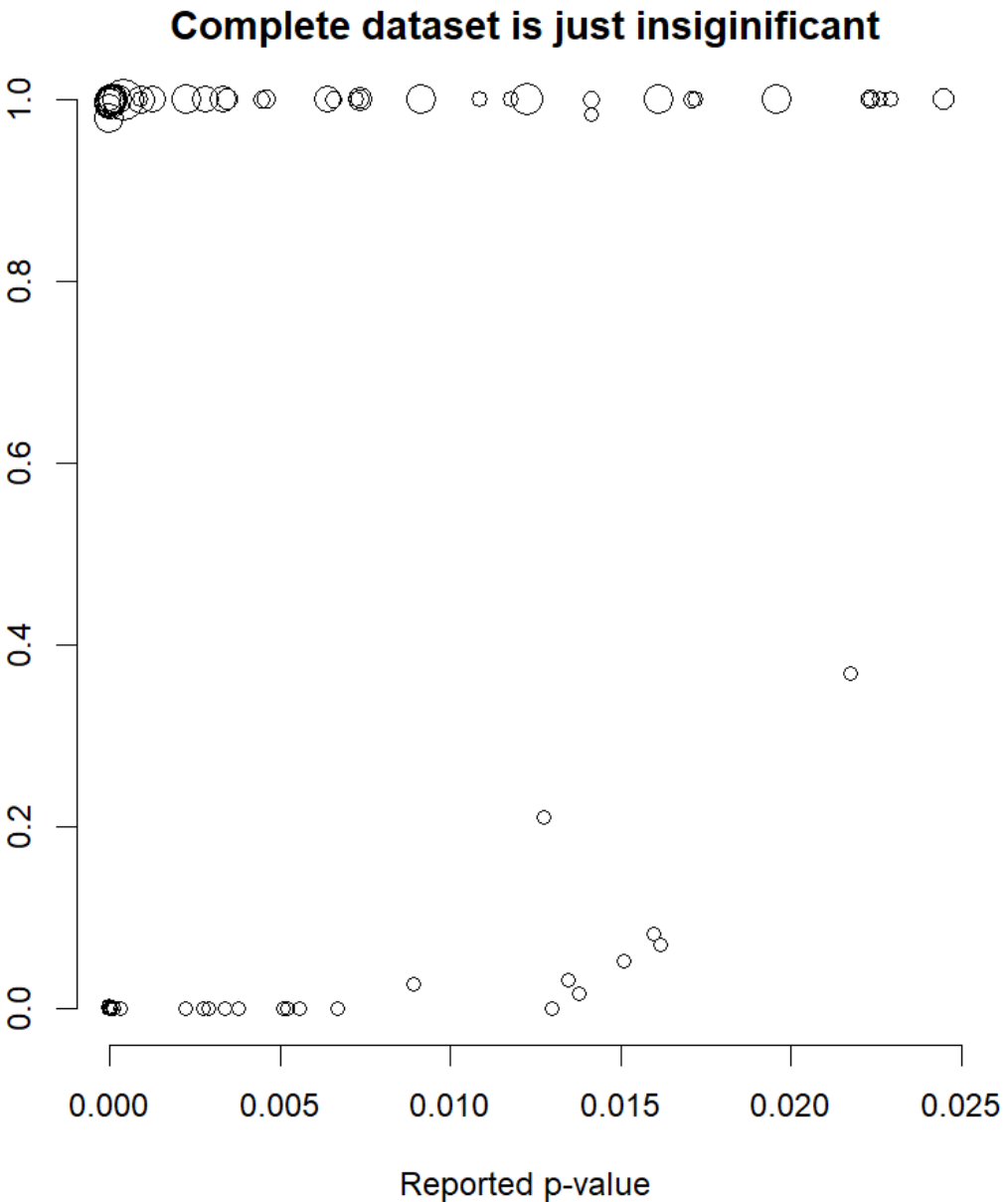
Main conclusion

1. Statistical results are generally robust to Missing Completely At Random.

2. With p-hacking, the robustness depends on (1) the proportion of data being excluded and (2) the severity of the p-hacking strategy.

Research Misconduct





Main conclusion

1. Statistical results are generally robust to Missing Completely At Random.

2. With p-hacking, the robustness depends on (1) the proportion of data being excluded and (2) the severity of the results.

3. With research misconduct, significant result are almost and always guaranteed

Tilburg University

In the context of meta-analysis

If the probabilities are large, caution should be taken.

Future research- how data exclusion influence :

- Statistical inference in meta-analysis?
- the effect size estimation in meta-analysis ?

Needed Information for the result t(211) = 2.02

$N_1 = 41$
$N_2 = 183$

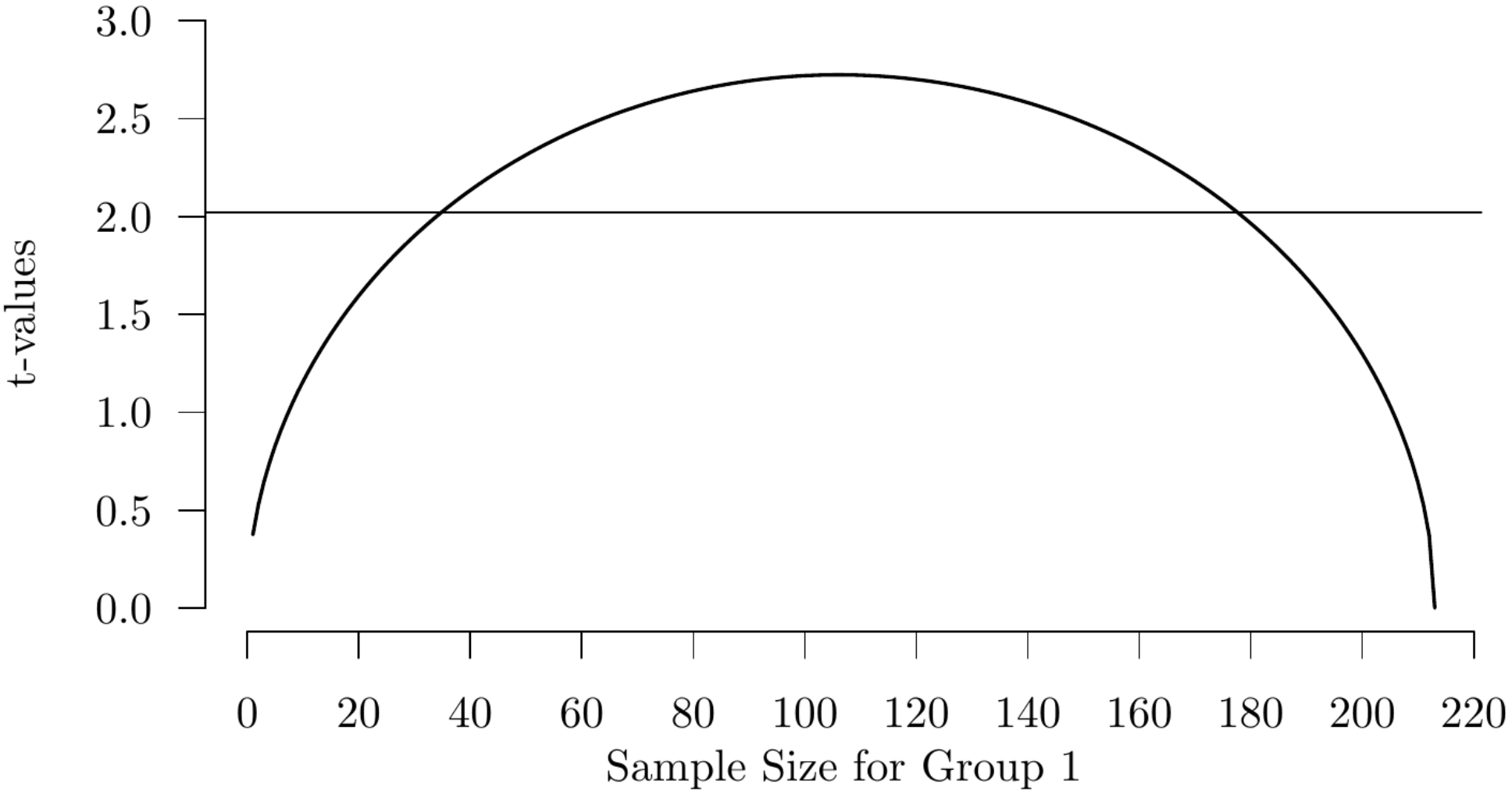
$m_1 = 6$
$m_2 = 5$

$n_1 = 35$
$n_2 = 178$

Statistical Result

There was a significant effect of condition, **t(211) = 2.02, p = .045, d = 0.37.** As predicted, participants who had previously read about the immoral act rated the debate and research on automaticity more negatively **(M = 4.29, SD = 1.05)** than those who read about the morally neutral act **(M = 4.68, SD = 1.04).**

t-value consistency



t-value inconsistency