Replicability: Are we finding real effects?

Why run a replication?

Can ask two kinds of questions: (cf. Simonsohn 2015)

Why run a replication?

Can ask two kinds of questions: (cf. Simonsohn 2015)

1. Is the effect in the original study “real” ?

Why run a replication?

Can ask two kinds of questions: (cf. Simonsohn 2015)

1. Is the effect in the original study “real” ?

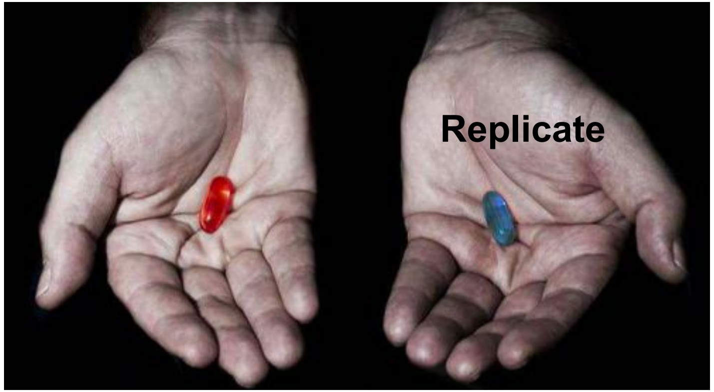
   NHST: 5% of all scientific findings are false positives.
   File drawer effect: > 5% of all **published** results are false positives.

Why run a replication?

Can ask two kinds of questions: (cf. Simonsohn 2015)

1. Is the effect in the original study “real” ?

   NHST: 5% of all scientific findings are false positives.
   File drawer effect: > 5% of all **published** results are false positives.



Why run a replication?

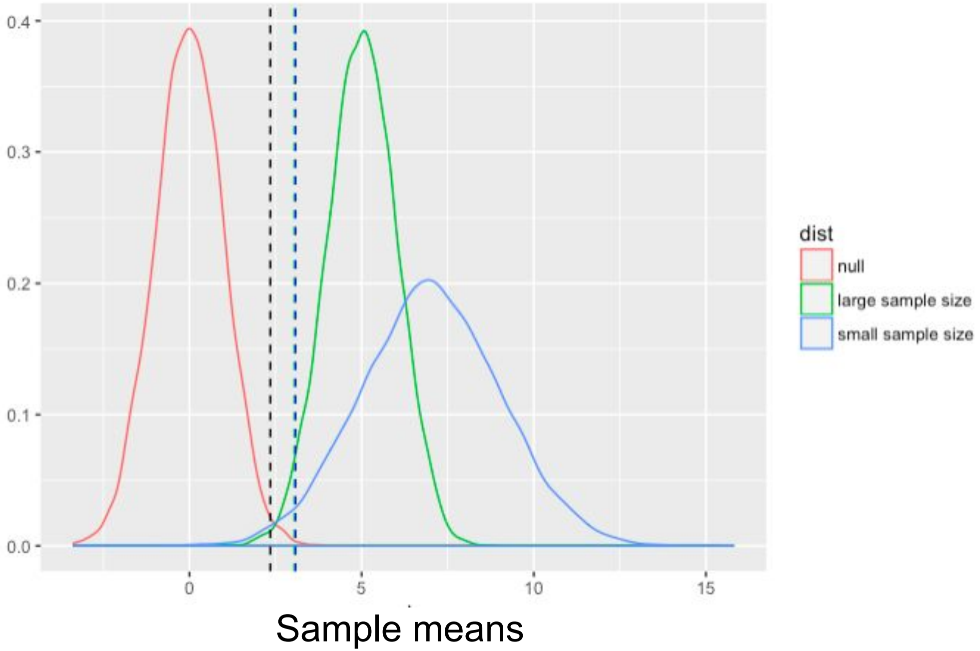
Can ask two kinds of questions: (cf. Simonsohn 2015)

2. Was the original effect a good estimate of the "true" effect?

Why run a replication?

Can ask two kinds of questions: (cf. Simonsohn 2015)

2. Was the original effect a good estimate of the "true" effect?



Studies with smaller sample sizes (i.e. with lower power) are more likely to overestimate the effect size.

Aside on effect sizes and power

- Power: The probability of finding an effect of a particular magnitude ("effect size") given a particular sample size.
  - Power analyses: What is the required sample size to achieve a certain power threshold (usually 0.8) for a given effect size.
  - Underpowered study: Sample size has power less certain threshold.

Aside on effect sizes and power

- Power: The probability of finding an effect of a particular magnitude (“effect size”) given a particular sample size.
  - Power analyses: What is the required sample size to achieve a certain power threshold (usually 0.8) for a given effect size.
  - Underpowered study: Sample size has power less certain threshold.
- In the context of power analyses, effect sizes are usually specified in terms of standardized mean differences

$$\text{Effect size} = \frac{\text{Mean(Group 1)} - \text{Mean(Group 2)}}{\text{Combined standard deviation}}$$

Cohen’s estimates: 0.2 - small, 0.5 - average, 0.8 - large

Using statistical significance to evaluate success

Rerun the experiment with sample size large enough to get adequate power.

Replication is successful if you find a statistically significant effect.

Using statistical significance to evaluate success

Rerun the experiment with sample size large enough to get adequate power.
Replication is successful if you find a statistically significant effect.

Caveat:

- Power is calculated based on effect size in original study
  - Effect size could be an overestimate if the original study was underpowered.
  - Might need more power than expected.

Camerer et al (2018) study

- Replicated 21 studies in Nature and Science between 2010 and 2015
  - All studies had experimental vs control comparison with at least one significant result and were run on accessible populations.
  - Replicated one finding from every study.
- Two stage process:
  - Stage 1: 90% power to detect 75% of effect size
  - Stage 2 (if stage 1 doesn’t replicate): 90% power to detect 50% of effect size

Statistical significance criterion

- 12 studies (57%) replicated when powered to detect 75% of the effect size
- 2 additional studies replicated when powered to detect 50% of the effect size.

This is a much higher percentage than the 36% in the Reproducibility Project Psychology (RPP) which replicated 100 studies in psychology.

Statistical significance criterion

- 12 studies (57%) replicated when powered to detect 75% of the effect size
- 2 additional studies replicated when powered to detect 50% of the effect size.

This is a much higher percentage than the 36% in the Reproducibility Project Psychology (RPP) which replicated 100 studies in psychology.

If we want to use this criterion, base the power analysis on 50% of the effect size.

Small telescope approach

"It is generally very difficult to prove that something does not exist; it is considerably easier to show that a tool is inadequate for studying that something." (Simonsohn 2015)

Small telescope approach

"It is generally very difficult to prove that something does not exist; it is considerably easier to show that a tool is inadequate for studying that something." (Simonsohn 2015)

Basic idea:

d_O : Effect size that would have 33% power with original sample size

d_R : Effect size of the replication.

Test significance of $d_R < d_O$

Small telescope approach

"It is generally very difficult to prove that something does not exist; it is considerably easier to show that a tool is inadequate for studying that something." (Simonsohn 2015)

Basic idea:

d_O : Effect size that would have 33% power with original sample size

d_R : Effect size of the replication.

Even if the effect was real, the original study had very low chances of finding it.

Test significance of $d_R < d_O$

Small telescope approach

"It is generally very difficult to prove that something does not exist; it is considerably easier to show that a tool is inadequate for studying that something." (Simonsohn 2015)

Basic idea:

d_O : Effect size that would have 33% power with original sample size

d_R : Effect size of the replication.

Even if the effect was real, the original study had very low chances of finding it.

Test significance of $d_R < d_O$

One study had a significant effect in the right direction but was too small to have been found by the original study.

Combining data: Meta-analytic estimate

- Combine original effect and replication effect
- 16 studies had significant meta-analytic effect ($p < 0.05$)
  - Something not significant in the replication but in the same direction as the original can add up and become significant
- More stringent alpha level recommended for meta-analysis. With this the same 13 were significant ($p < 0.005$)

Combining data: Meta-analytic estimate

- Combine original effect and replication effect
- 16 studies had significant meta-analytic effect ($p < 0.05$)
  - Something not significant in the replication but in the same direction as the original can add up and become significant
- More stringent alpha level recommended for meta-analysis. With this the same 13 were significant ($p < 0.005$)

Similarly confidence intervals can be plotted (with combined sd). 14 studies fell in the 95% confidence interval

Criticism about NHST approaches

These approaches are useful to think about what a successful replication can tell us. But if we fail to find a significant effect we cannot reason directly about whether or not the effect exists.

Bayes factor

$$\text{Bayes Factor} = \frac{P(\text{data} \mid \text{model1}) * P(\text{model1})}{P(\text{data} \mid \text{model2}) * P(\text{model2})}$$

BF < 1 : Supports model2

BF > 3 : Substantial evidence for model1

BF > 20 : Strong evidence for model1

Bayes factor

$$\text{Bayes Factor} = \frac{P(\text{data} \mid \mathbf{H_A}) * P(\mathbf{H_A})}{P(\text{data} \mid \mathbf{H_0}) * P(\mathbf{H_0})}$$

The studies that failed to replicate had Bayes factor < 1 providing evidence for the null hypothesis

BF < 1 : Supports model2

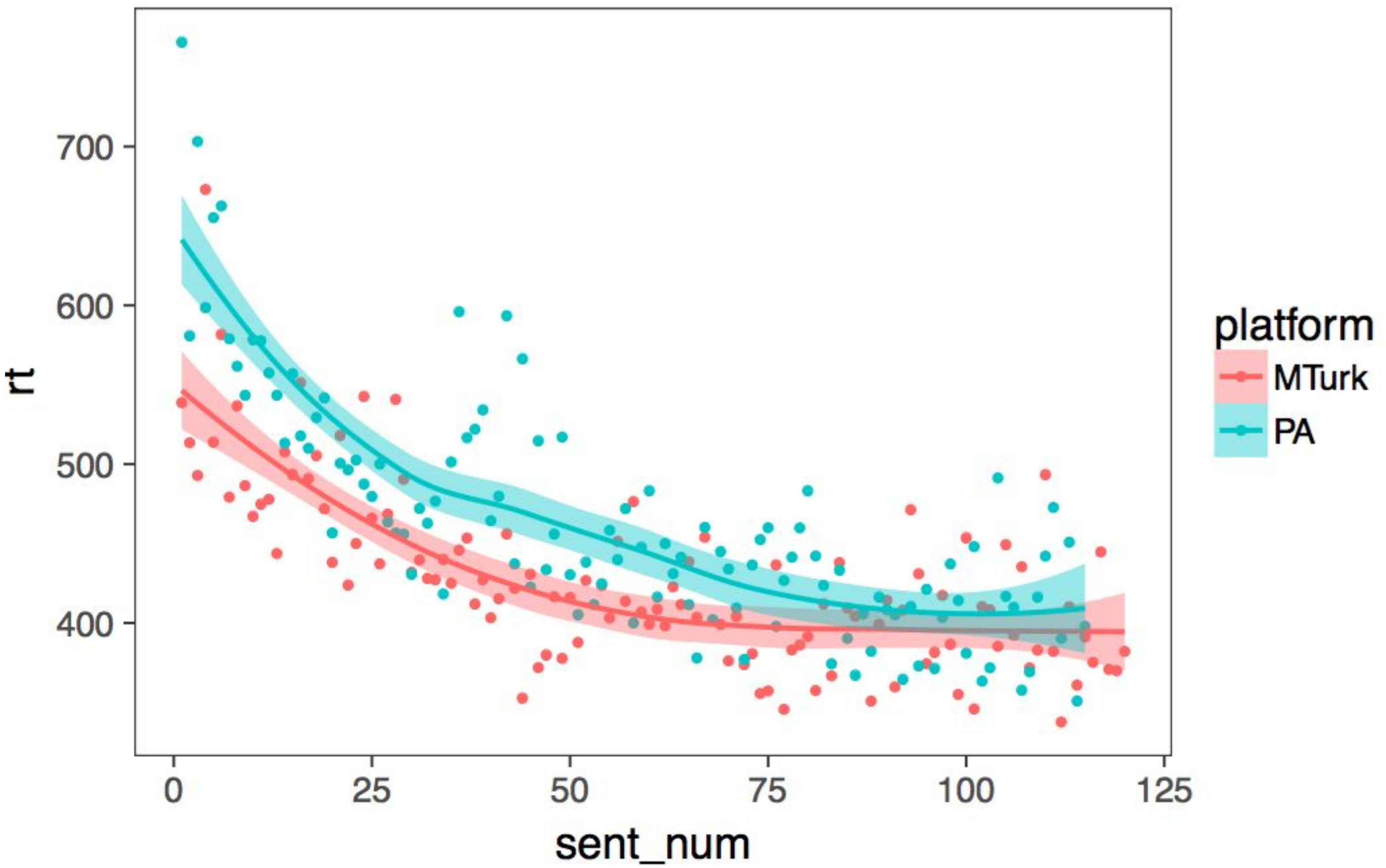
BF > 3 : Substantial evidence for model1

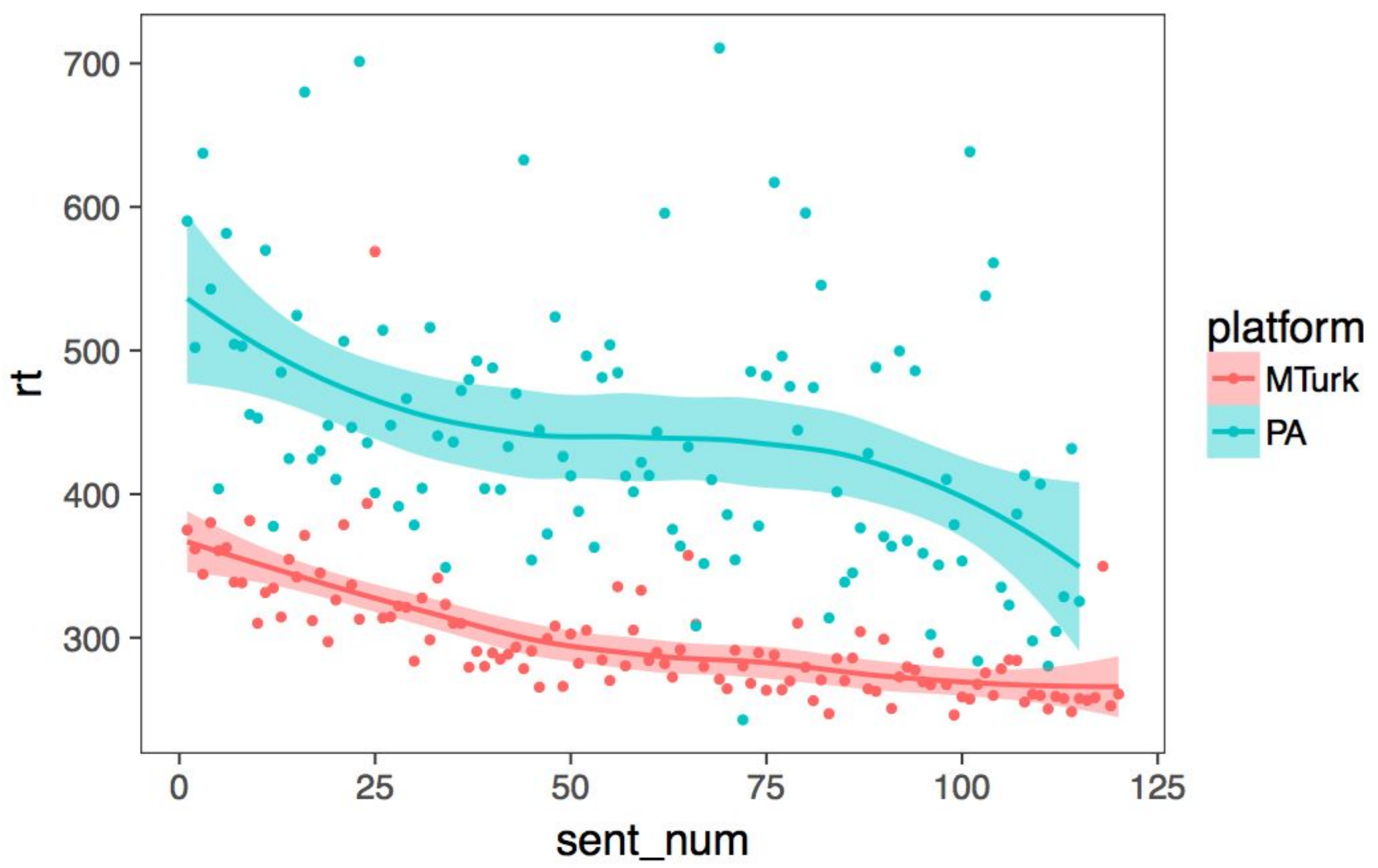
BF > 20 : Strong evidence for model1

Summary

- The effect size in studies - particularly those that are underpowered - can be exaggerated. So design replication studies such that they have the power to find 50% of the original effect size.
- Different approaches to looking at replication success resulted in the same conclusions.
- Studies that failed to replicate did not show any evidence for the effect.
    - These were probably false positives
- People were able to predict which studies would not replicate - failure to replicate not due to chance alone.

Replication assumes that there are no systematic differences in the procedures. This is a reasonable assumption to make. But some of our data suggest that there are baseline differences between crowdsourcing platforms.


rt
700
600
500
400
0
25
50
75
100
125
sent_num
platform
MTurk
PA



rt
700
600
500
400
300
0
25
50
75
100
125
sent_num
platform
MTurk
PA