The statistical fundamentals of (non-)replicability

Prof Jeff Miller
Department of Psychology
University of Otago

Supported by

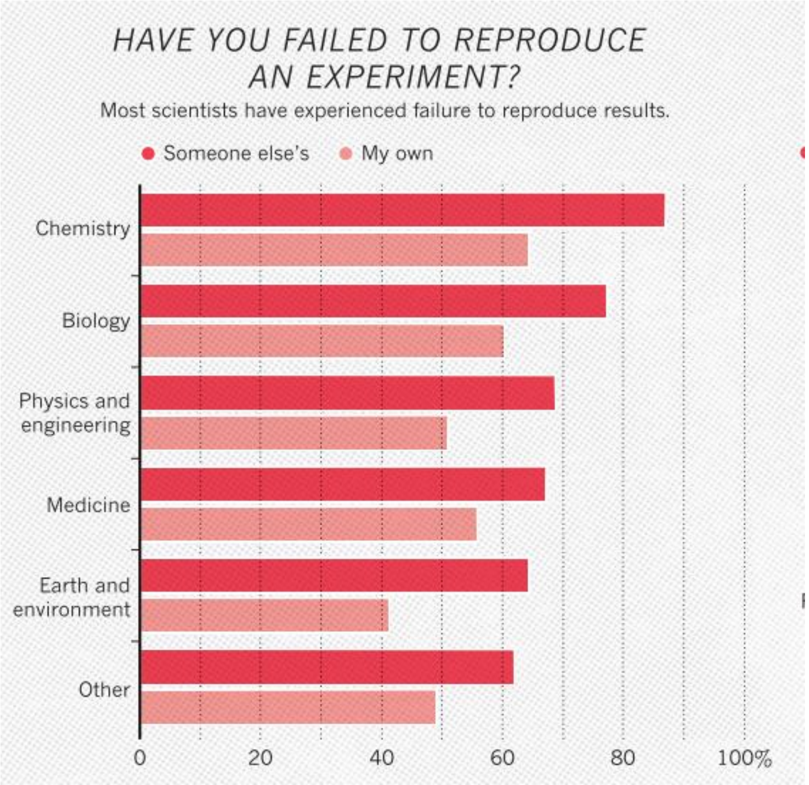
Alexander von Humboldt
Stiftung/Foundation

Presentation at University of Canterbury Business School Workshop
“Reproducibility and Integrity in Scientific Research”, 26-10-2018
Slides available at https://tinyurl.com/y8oau559

Replication

The ideal: *“Replicability of findings is at the heart of any empirical science”*
Asendorpf et al. (2013, *European Journal of Personality*)

The reality:



IS THERE A CRISIS?	
Yes, a significant crisis	52%
Yes, a slight crisis	38%
Don‘t know	7%
No, there is no crisis	3%

(Baker, 2016, *Nature*)

Replication & random variability

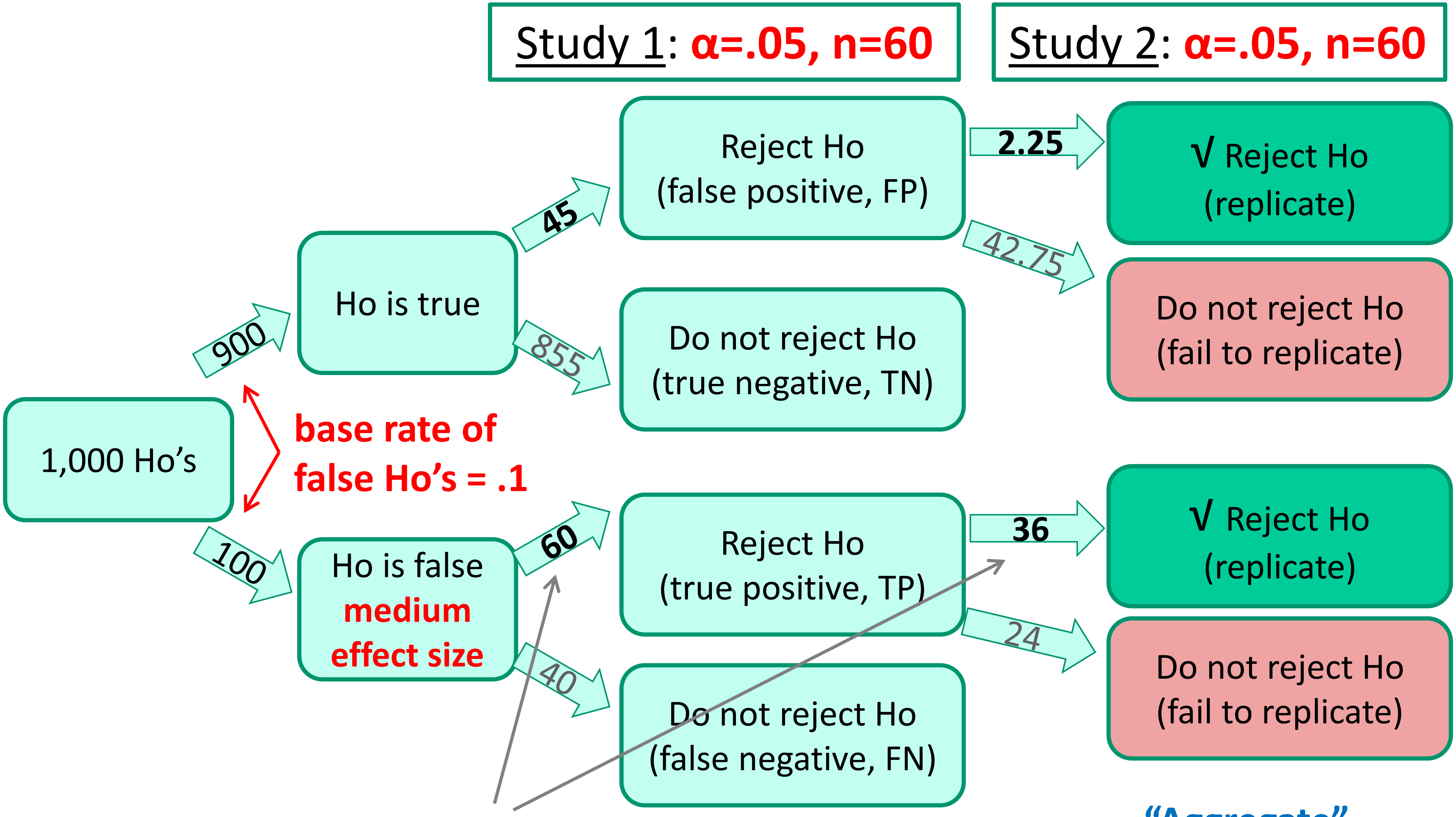
- Replications will only be successful with some probability, $p_{rep} < 100\%$.
- Statistical models can be used to study p_{rep}:
  - what p_{rep} values should we expect?
  - what factors affect p_{rep}?
  - how can we increase p_{rep}?
- "replication": an equivalent study with a statistically significant effect in the same direction as the original study.

Hypothesis testing

	Decision reached from data	
True state of world	**“do not reject Ho”**	**“reject Ho”**
Ho is true	true negative (TN)	false positive (FP) or type I error
Ho is false	false negative (FN) or type II error	true positive (TP)

	Conditional probability of decision	
True state of world	**“do not reject Ho”**	**“reject Ho”**
Ho is true	$1 – \alpha = .95$	$\alpha = .05$
Ho is false	β	$1 – \beta$ = power

A model of replication probability



Study 1: **α=.05, n=60**

Study 2: **α=.05, n=60**

1,000 Ho's

base rate of false Ho's = .1

900

Ho is true

45

Reject Ho (false positive, FP)

855

Do not reject Ho (true negative, TN)

2.25

√ Reject Ho (replicate)

42.75

Do not reject Ho (fail to replicate)

100

Ho is false **medium effect size**

60

Reject Ho (true positive, TP)

40

Do not reject Ho (false negative, FN)

36

√ Reject Ho (replicate)

24

Do not reject Ho (fail to replicate)

power = .60

Rate of false positives:

$$R_{fp} = \frac{45}{45+60} = .43$$

"Aggregate" Replication probability:

$$p_{rep} = \frac{2.25+36}{45+60} = .36$$

Slides at https://tinyurl.com/y8oau559

Individual p_{rep}

An individual researcher from the previous slide might say:

> "I'm not interested in aggregate results for a whole field, but only in p_{rep} for my particular effect. Based on these calculations, if I repeat my study 100 times, should I expect about 36% significant results?"

Answer: **No!**

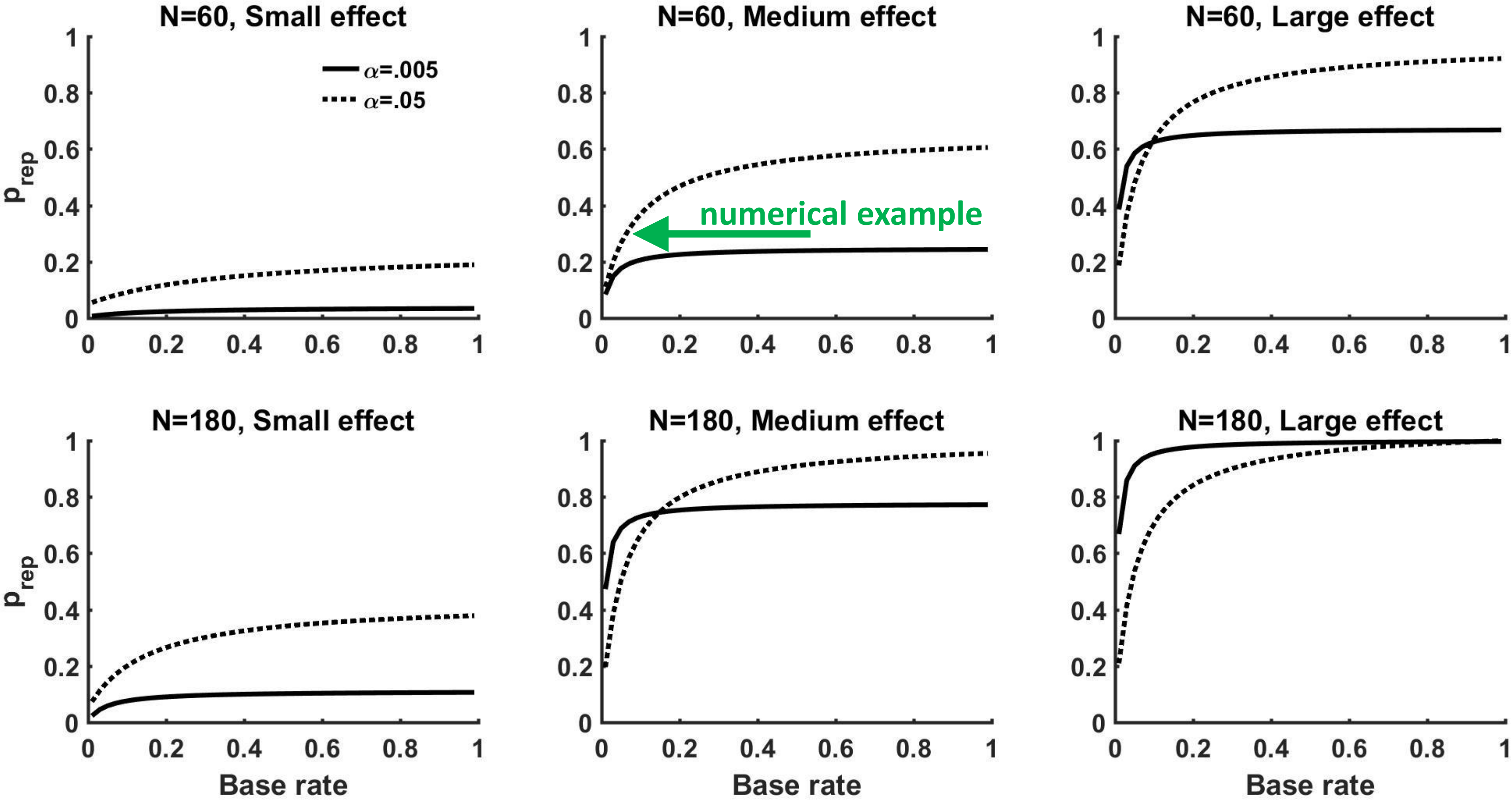
- If your effect is real, you will get about 60% significant results (individual p_{rep} = .60).
- If not, you will get about 5% significant results (individual p_{rep} = .05).

Aggregate p_{rep} = weighted average of individual p_{rep}'s

$$.36 = \frac{45}{105} \times .05 + \frac{60}{105} \times .60$$

Slides at https://tinyurl.com/y8oau559

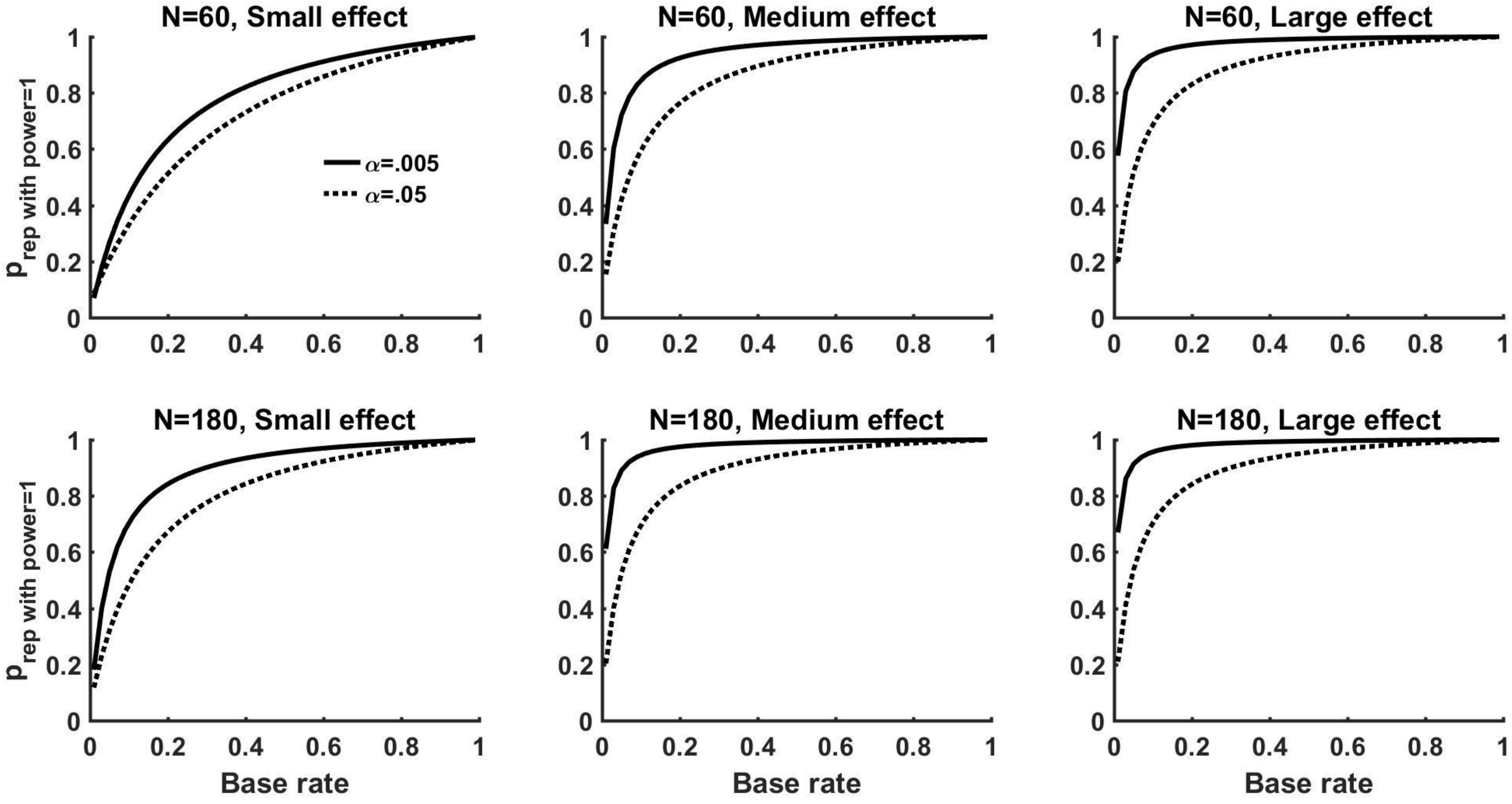
Aggregate p_{rep}: exact replications



numerical example

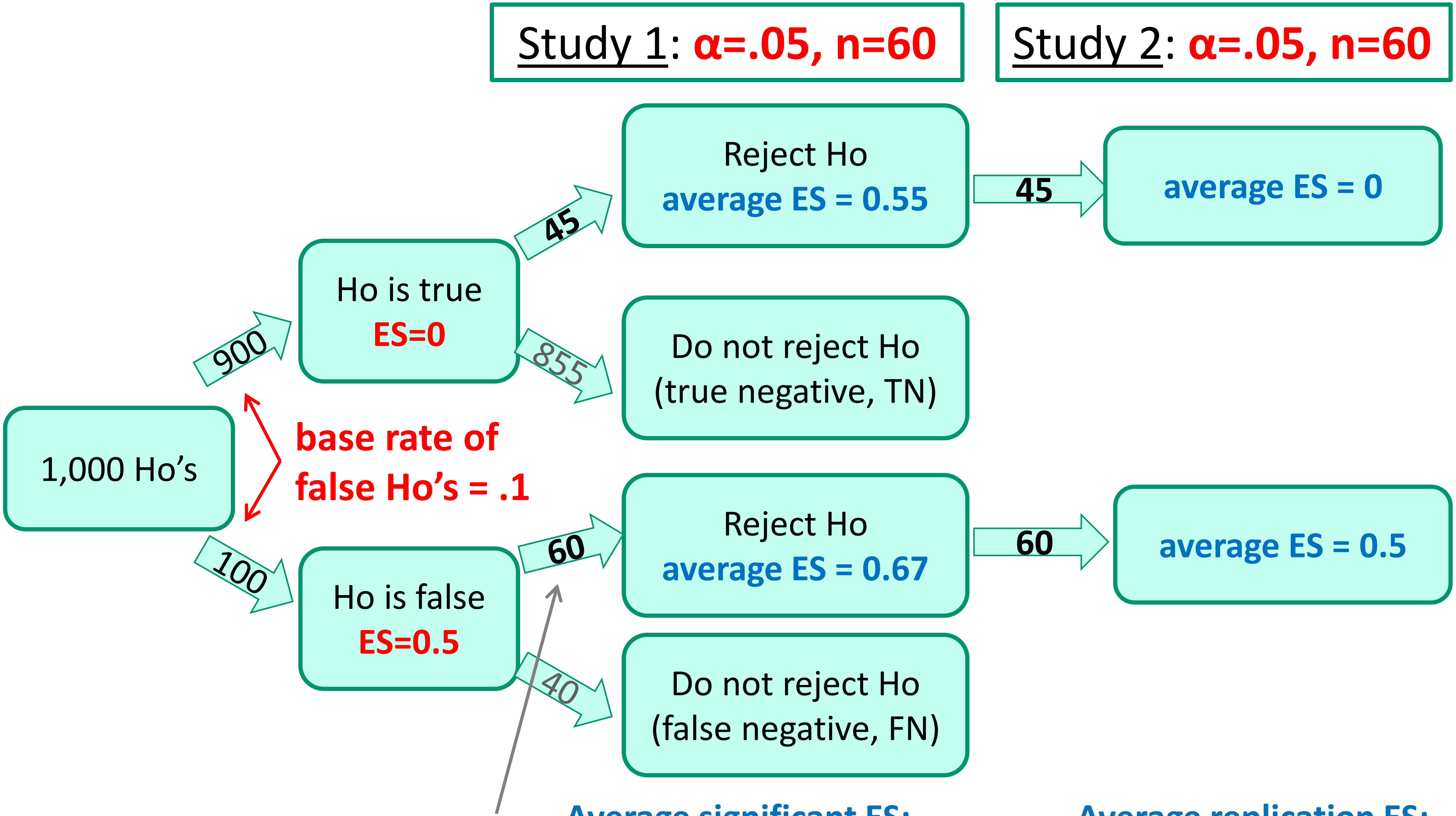
Slides at https://tinyurl.com/y8oau559

Aggregate p_{rep}: power of study 2 = 1



Slides at https://tinyurl.com/y8oau559

A model of replication effect size (ES)



Study 1: α=.05, n=60

Study 2: α=.05, n=60

1,000 Ho’s

900

Ho is true
ES=0

45

Reject Ho
average ES = 0.55

45

average ES = 0

855

Do not reject Ho
(true negative, TN)

base rate of
false Ho’s = .1

100

Ho is false
ES=0.5

60

Reject Ho
average ES = 0.67

60

average ES = 0.5

40

Do not reject Ho
(false negative, FN)

power = .60

Average significant ES:

$$ES = \frac{45 \times 0.55 + 60 \times 0.67}{45 + 60}$$

$$ES = 0.62$$

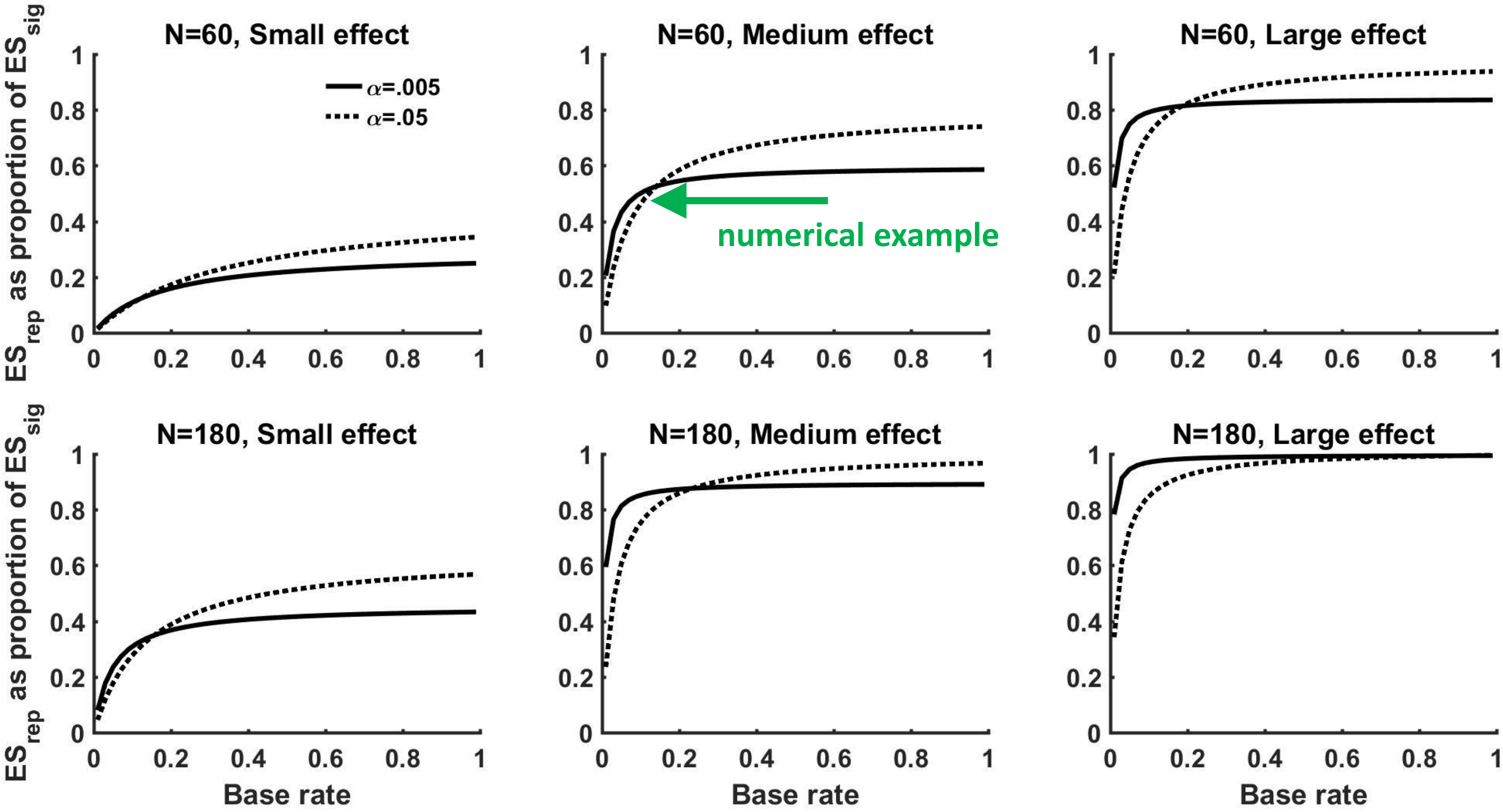
Average replication ES:

$$ES = \frac{45 \times 0 + 60 \times 0.5}{45 + 60}$$

$$ES = 0.29$$

Slides at https://tinyurl.com/y8oau559

Effect sizes of replications (“decline effects”)



Slides at https://tinyurl.com/y8oau559

Conclusions

- Replication failures are inevitable

  …even with exact replications & best practices

- What should we do?
  - Adjust expectations about replicability
  - Be skeptical about one-off results
  - Improve base rates by improving theories
  - Fine-tune α and sample size

    …to maximize research *payoff* based on a cost/benefit analysis of TP, FP, TN, FN

    …may *not* maximize replicability

Slides at https://tinyurl.com/y8oau559