

Assurance-based sample size determination for replication studies under publication bias



Ondrej Libiger

September 23rd

NCS 2026

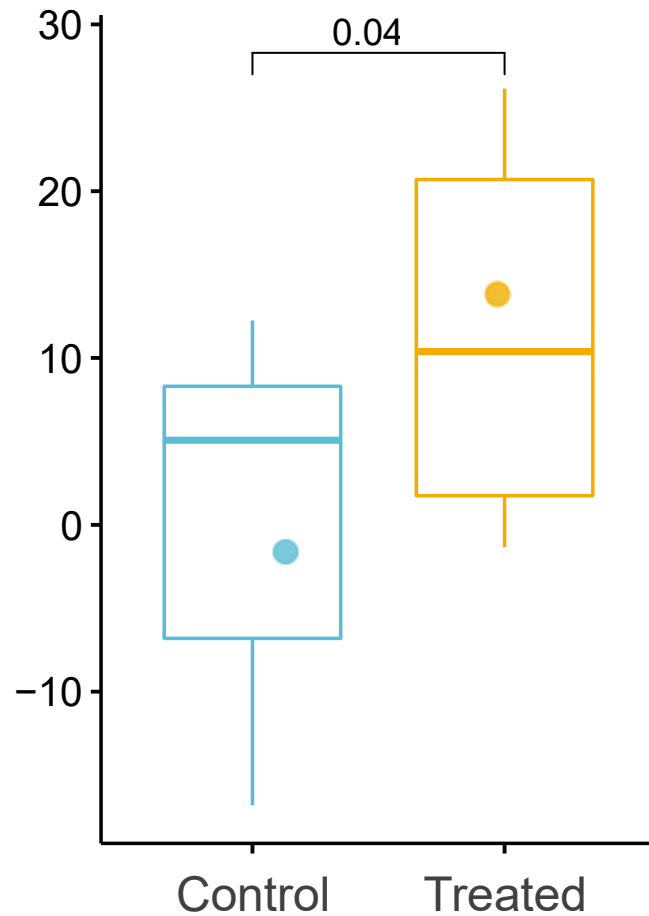
Padova, IT

Summary

1. Replication usually requires a much larger sample size
2. Does this fact contribute to the *replication crisis*?
3. The probability of replicating a real finding from a published underpowered study using conditional 80% power is:
 - surprisingly low
 - consistent with observed low replication rates
4. Frequent and widespread failure to replicate is primarily a consequence of standard statistical and publishing practice
5. Can we find the smallest sample size that will assure successful replication?



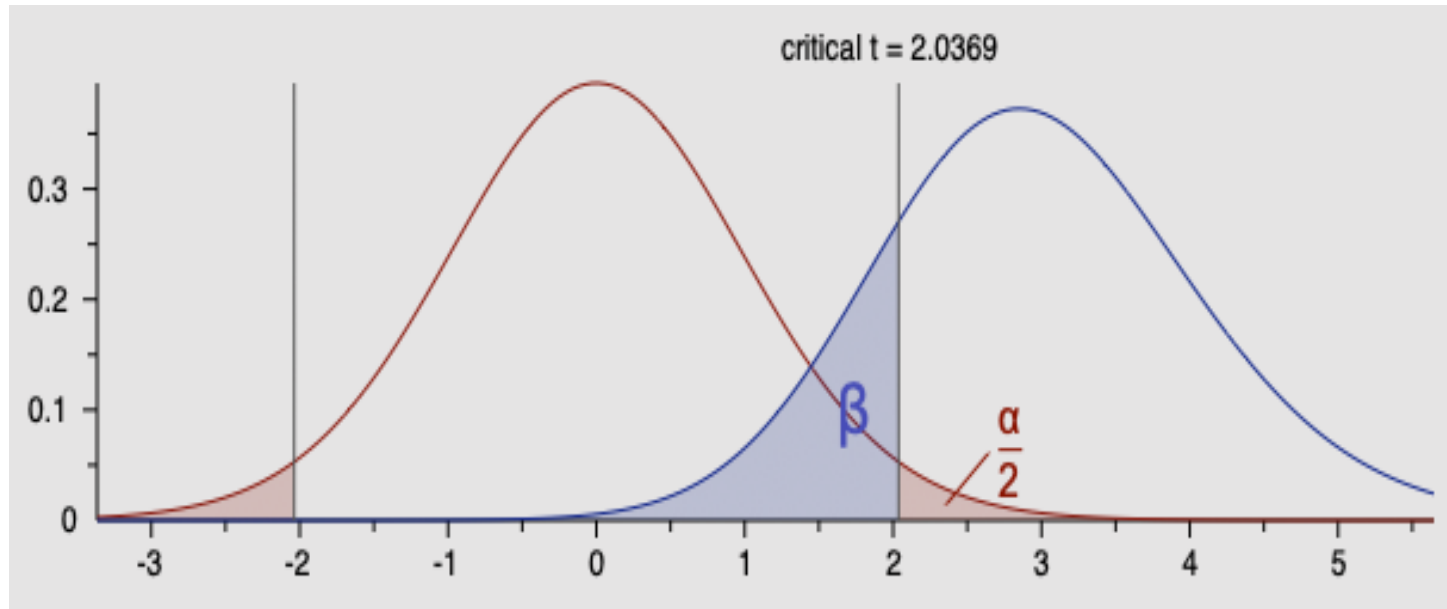
How many samples are needed to replicate this finding?



- $n = 10$
- *Control* sample mean = 0
- *Treated* sample mean = 10
- sample SD = 10

What's the **minimum necessary sample size** to detect a statistically significant difference in the same direction with 80% power?

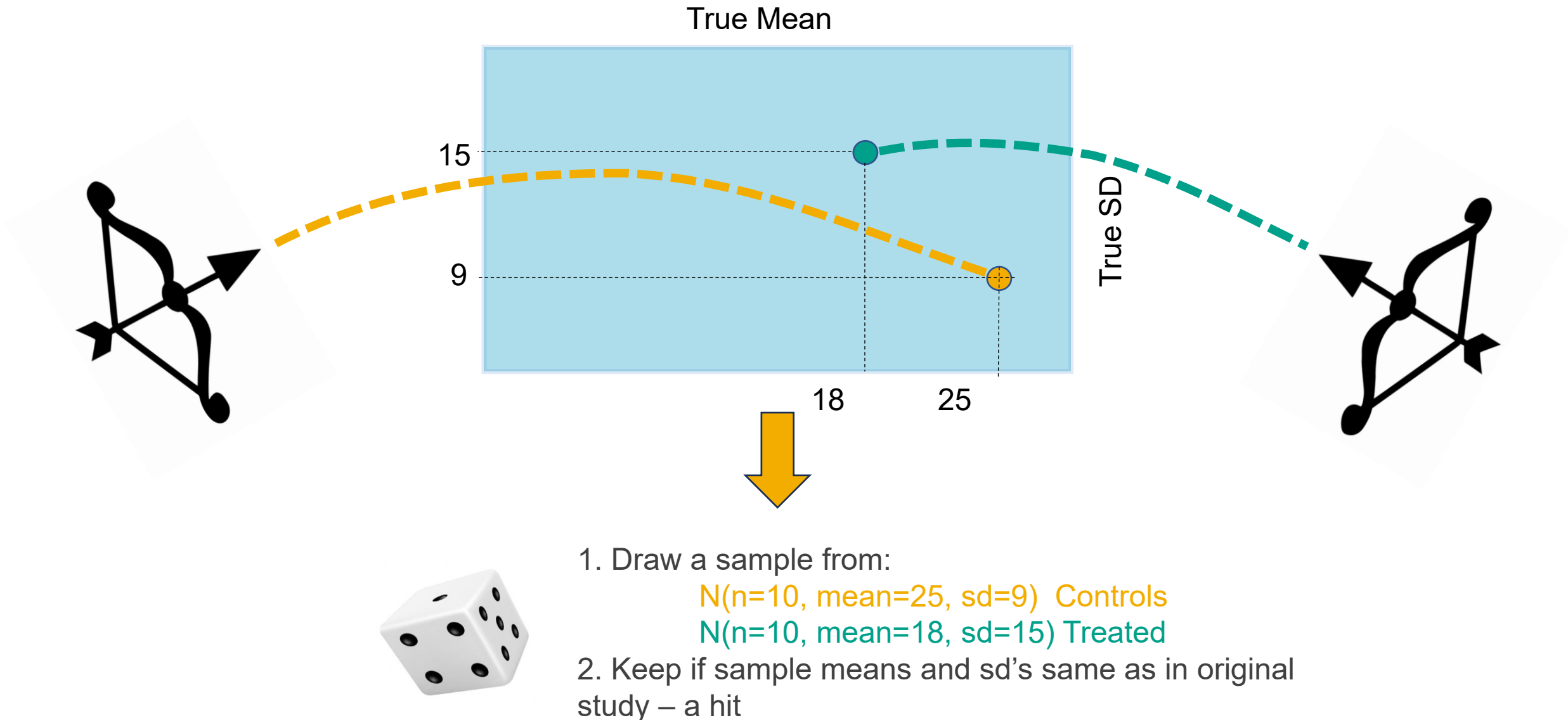
Conditional power assumes that the estimates are correct



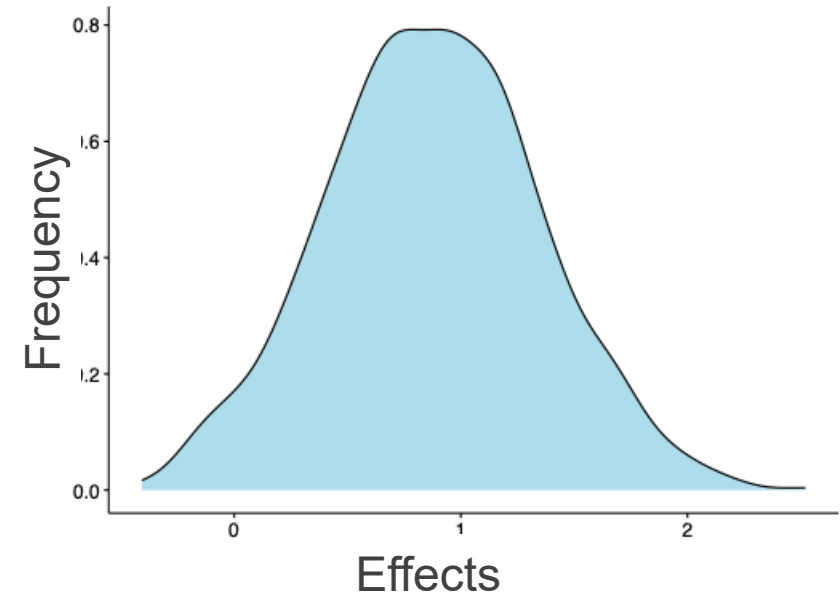
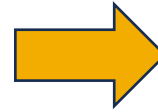
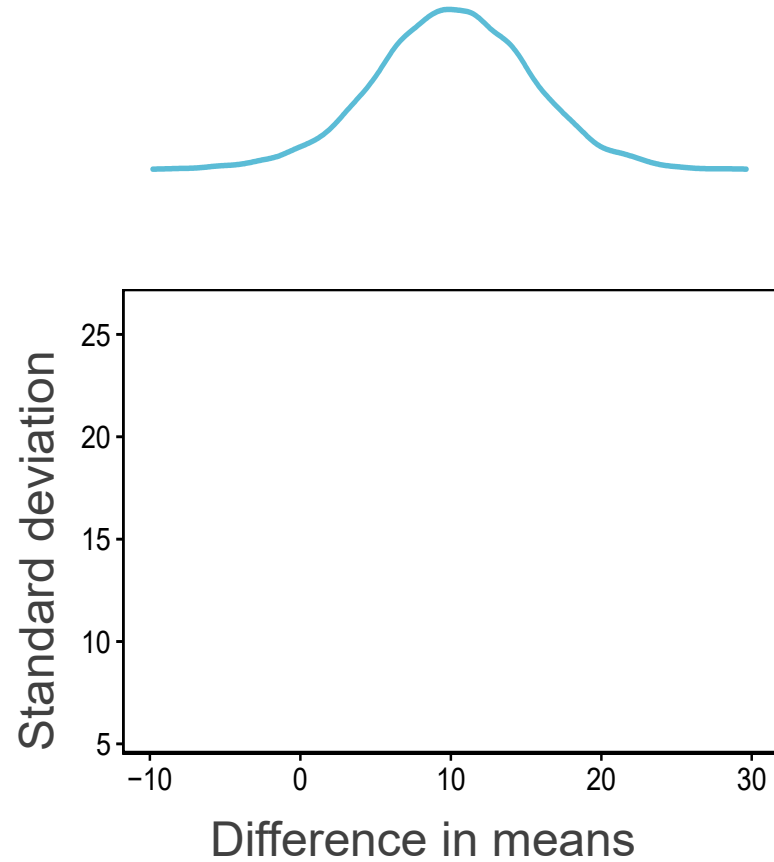
Original study: $n=10$

Conditional power: $n=17$

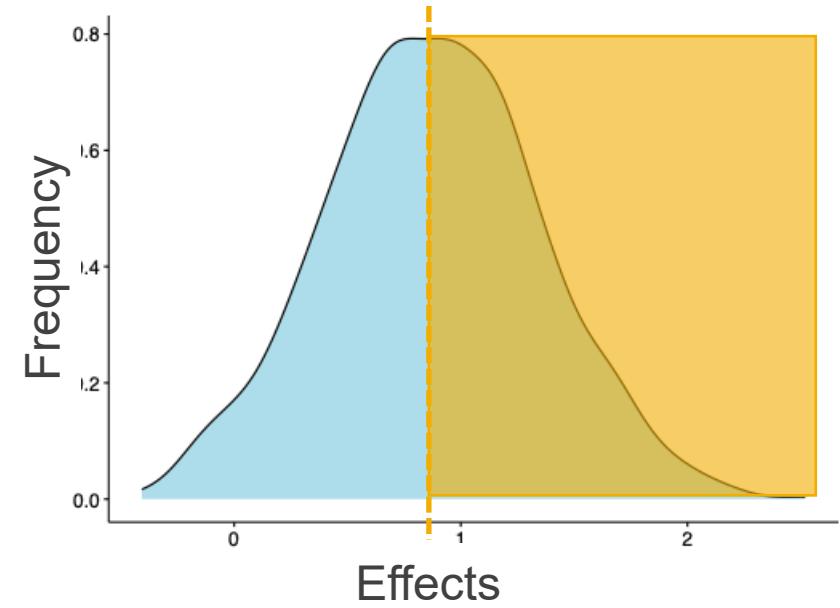
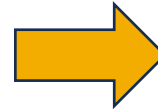
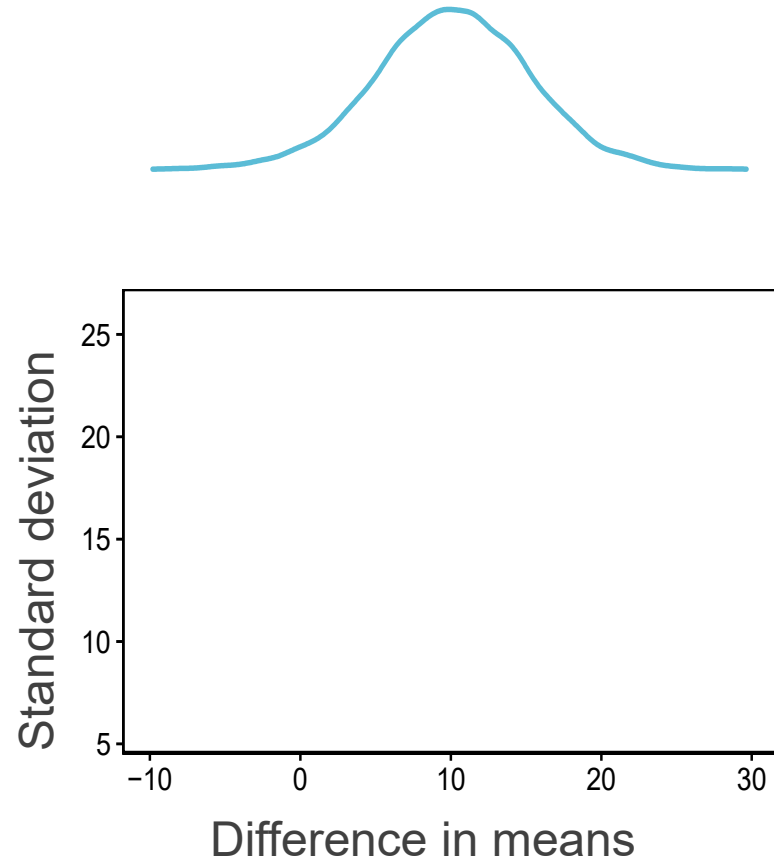
Simulations generate a distribution of possible *true* effects



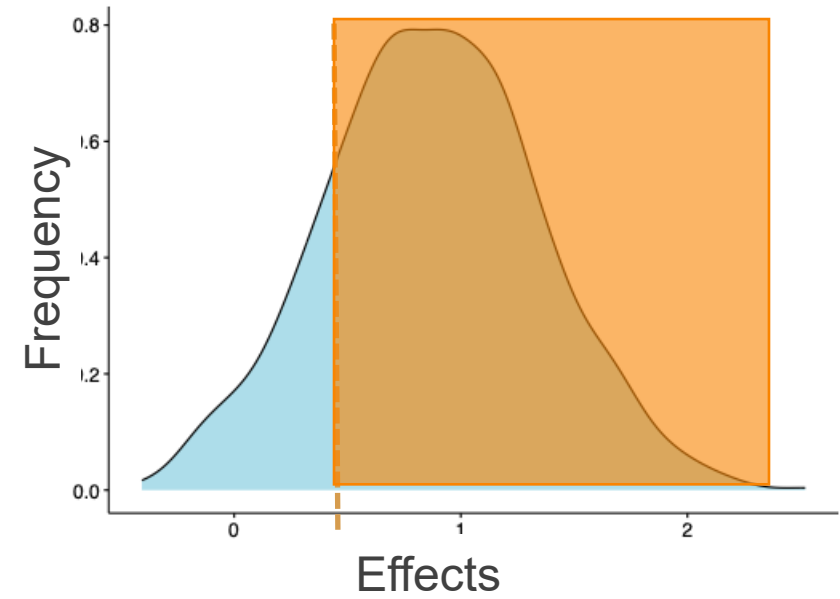
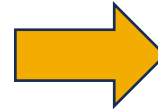
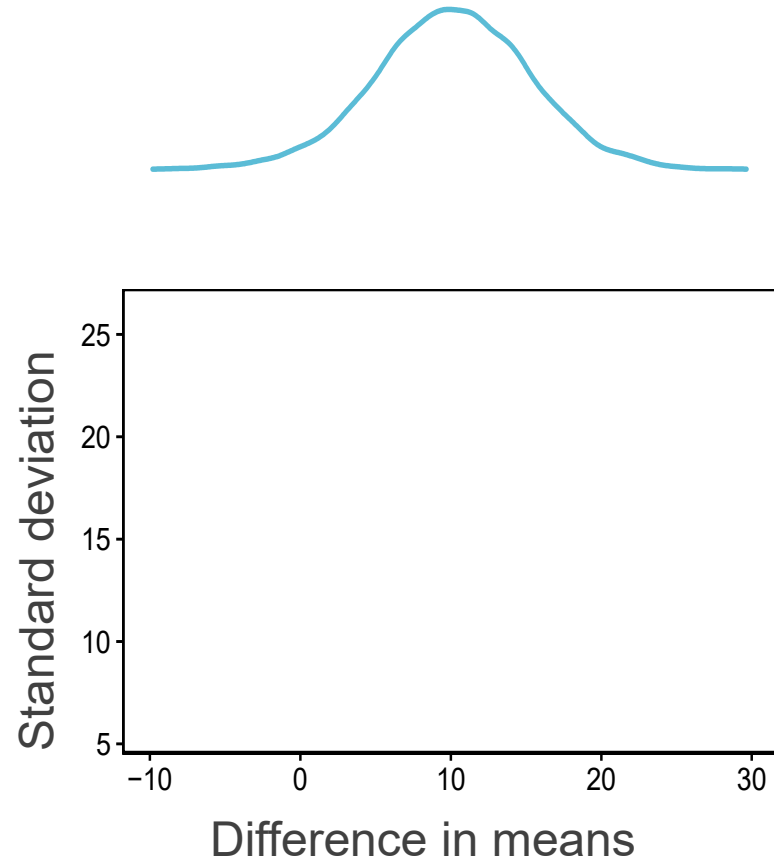
Assurance accounts for all possible *true* effects



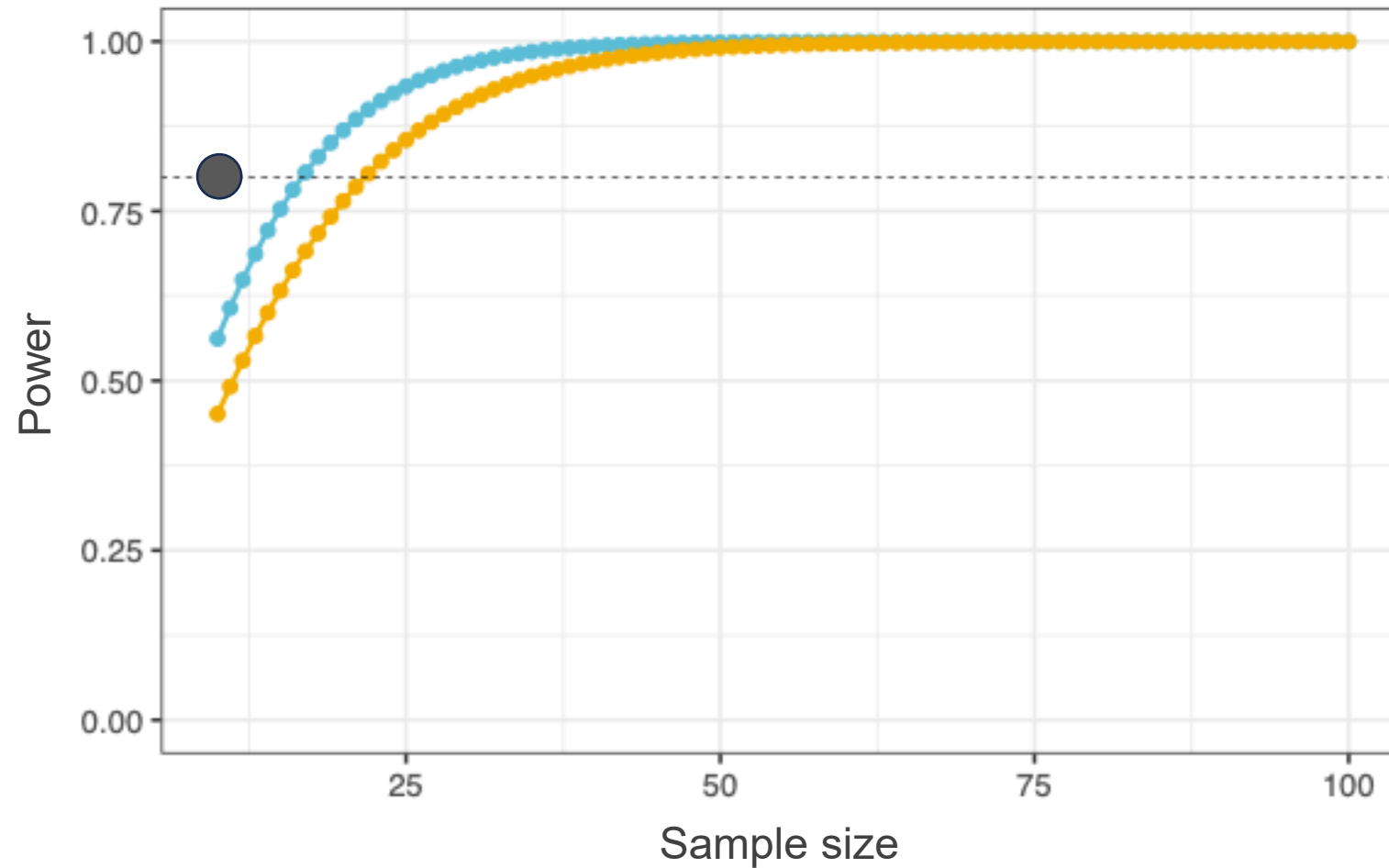
Assurance accounts for all possible *true* effects



Assurance accounts for all possible *true* effects



Assurance comes at the cost of larger required sample size

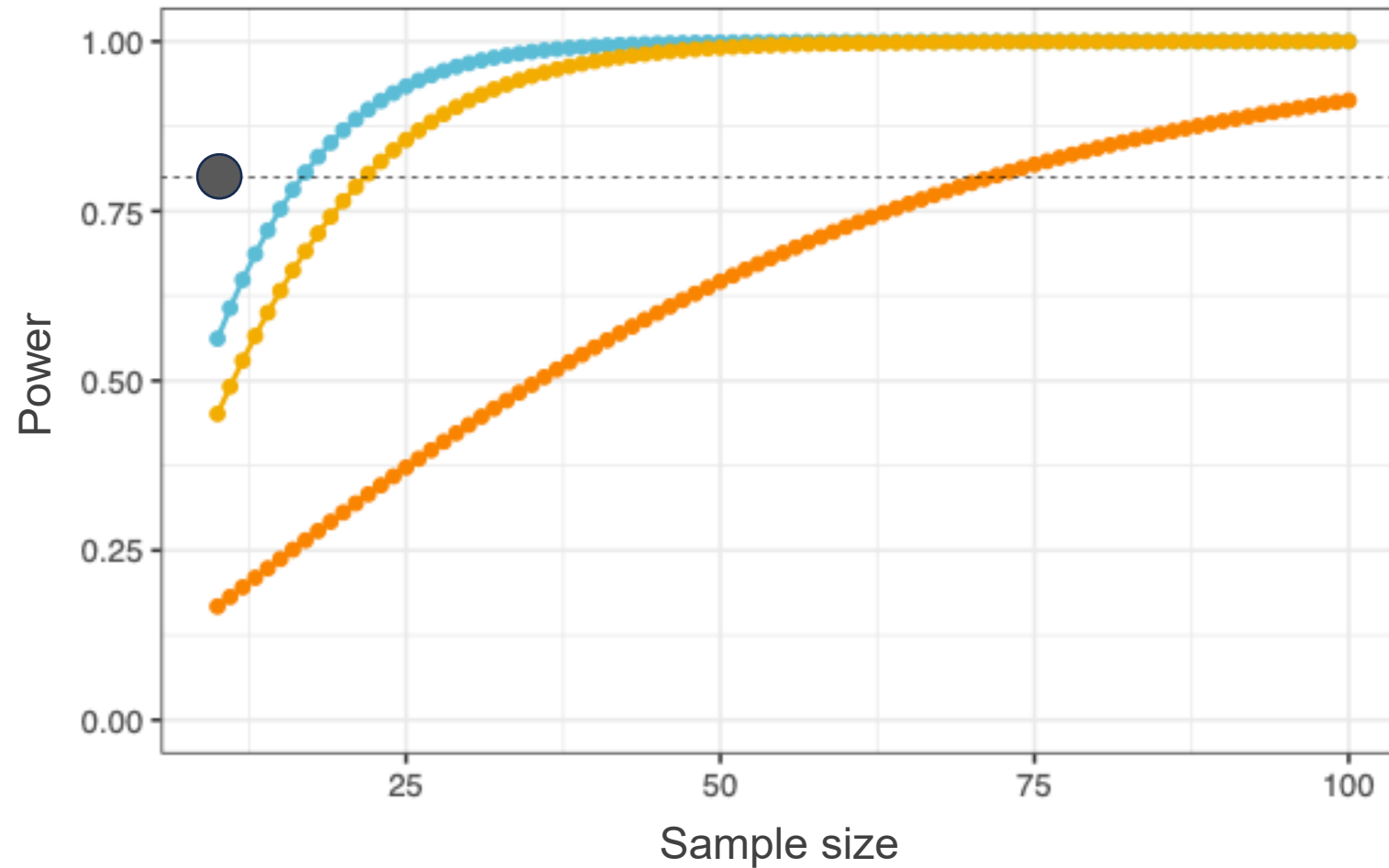


Original study: n=10

Conditional power: n=17

50% assurance of 80%: n=22

Achieving a high level of assurance is costly



Original study: n=10

Conditional power: n=17

50% assurance of 80%: n=22

80% assurance of 80%: n=72

Published results are often subject to publication bias

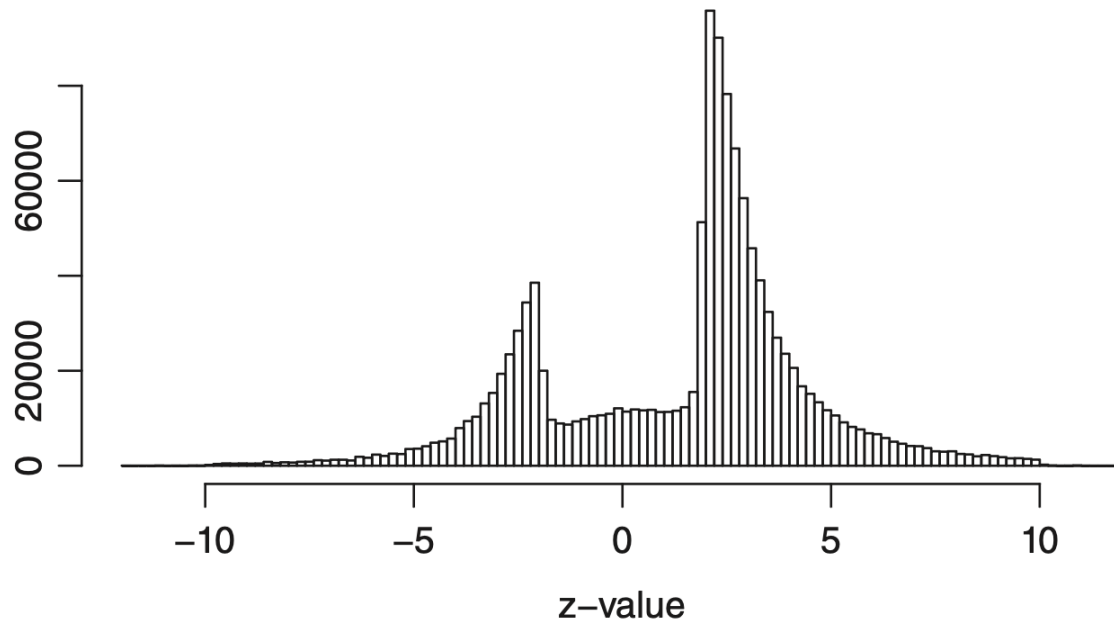
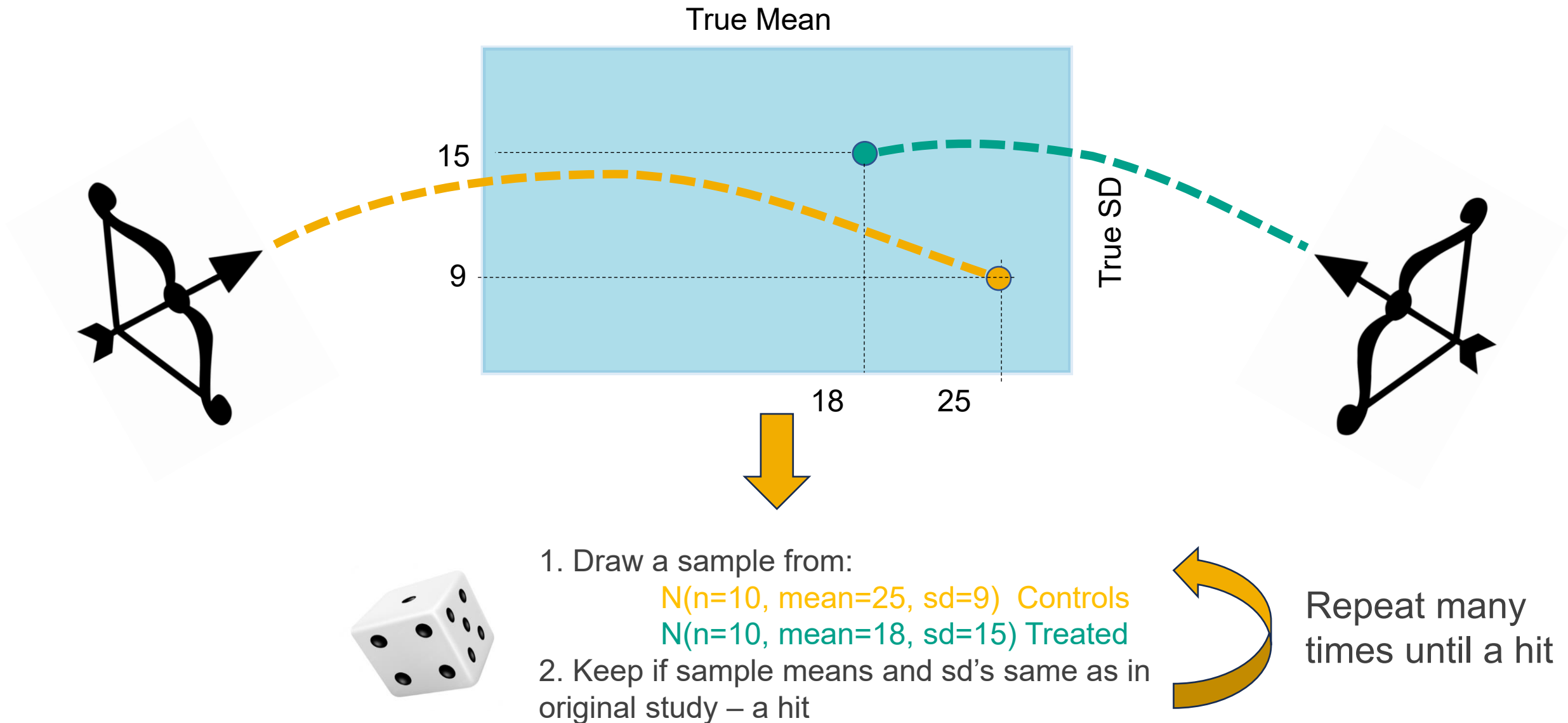


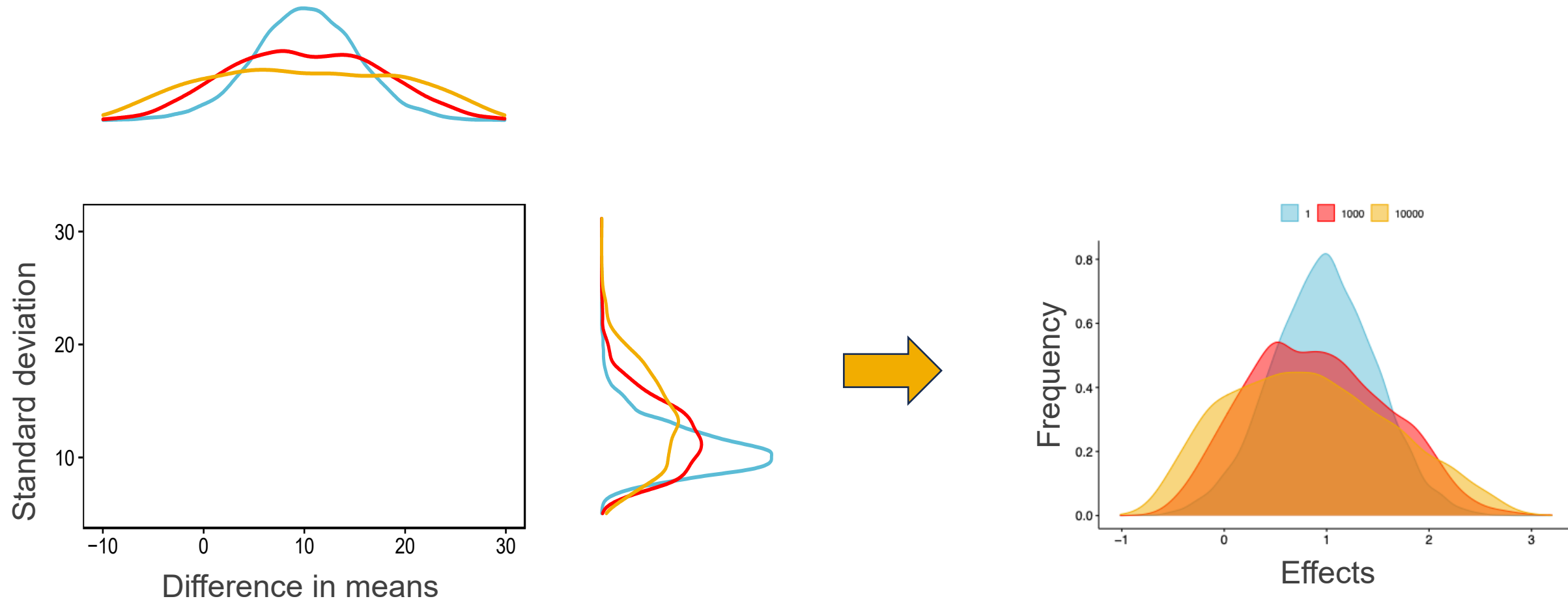
FIGURE 1 The distribution of more than one million z -values from Medline (1976–2019)

van Zwet & Cator, 2020; DOI:10.1111/stan.12241

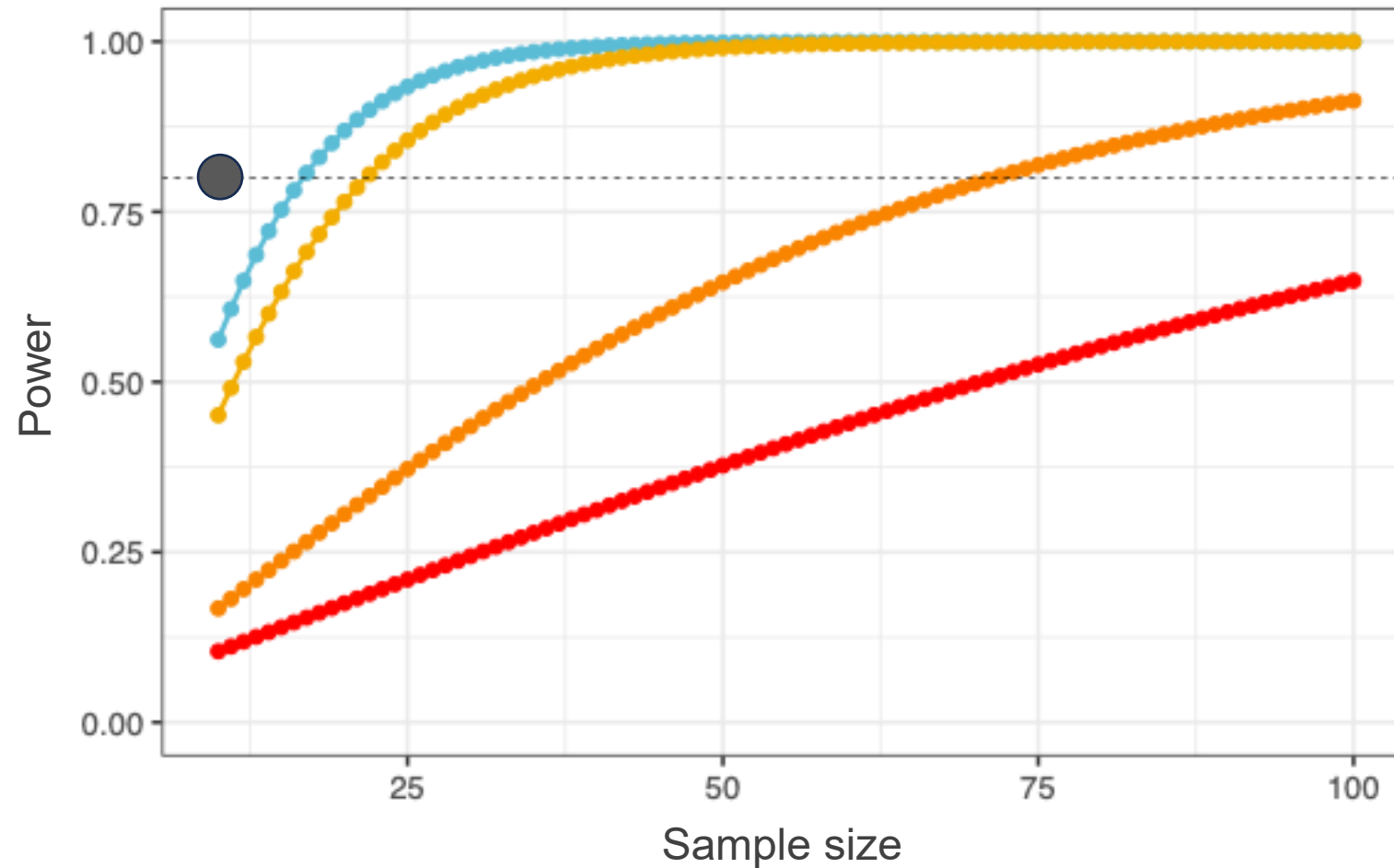
Simulations help generate a distribution of bias-adjusted effects



Bias-adjusted effects shift towards smaller effect sizes

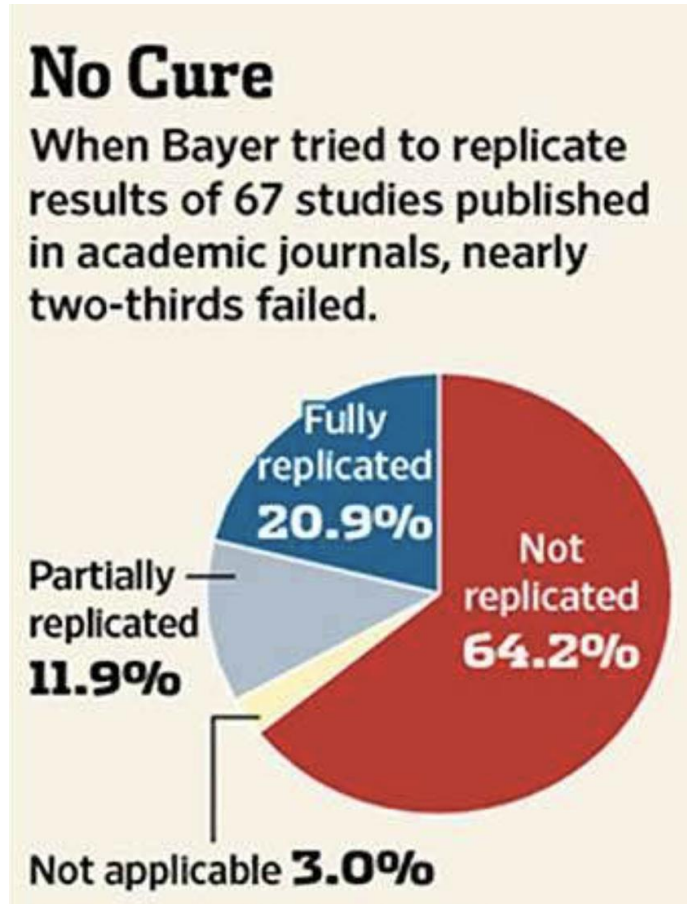


Replications under publication bias require even larger samples



Original study: n=10
Conditional power: n=17
50% assurance of 80%: n=22
80% assurance of 80%: n=72
80%-80% public. bias: n=143

If replication is so difficult, shouldn't we expect some failure?



Prinz et al., 2011 doi.org/10.1038/nrd3439-c1

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

- **Publication bias** (significance filter) Sterling TD, 1959
- **Hidden multiplicity** (investigator deg. of freedom) Gelman A Loken E, 2013)
- **Questionable research practices**
- **Fraud**

What is the probability of replicating a true finding?

1. Start with a hypothetical *true* effect

$\text{true_mean}_1 \pm \text{true_sd}_1, \text{true_mean}_2 \pm \text{true_sd}_2$

2. Generate a distribution of sample means and sd's

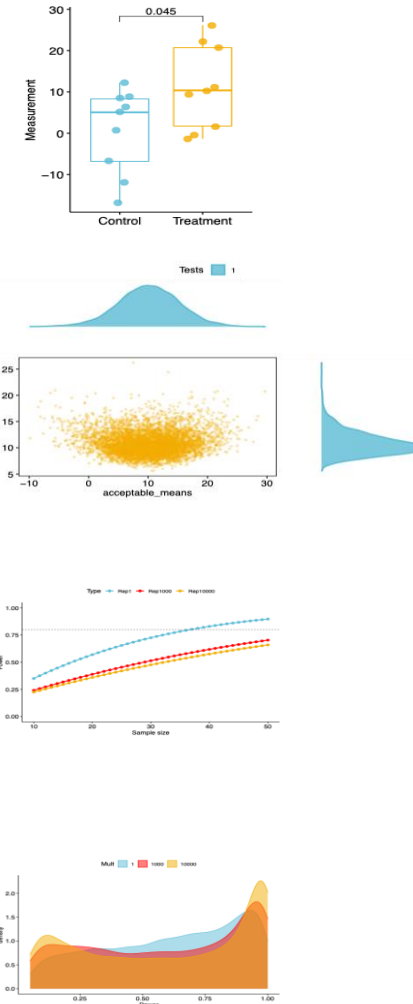
1. No bias (keep all)
2. Directional bias (keep if $\text{sample_mean}_2 > \text{sample_mean}_1$)
3. Publication bias (keep if $\text{sample_mean}_2 > \text{sample_mean}_1$ & $p < 0.05$)

3. Identify n needed to detect a difference in sample means

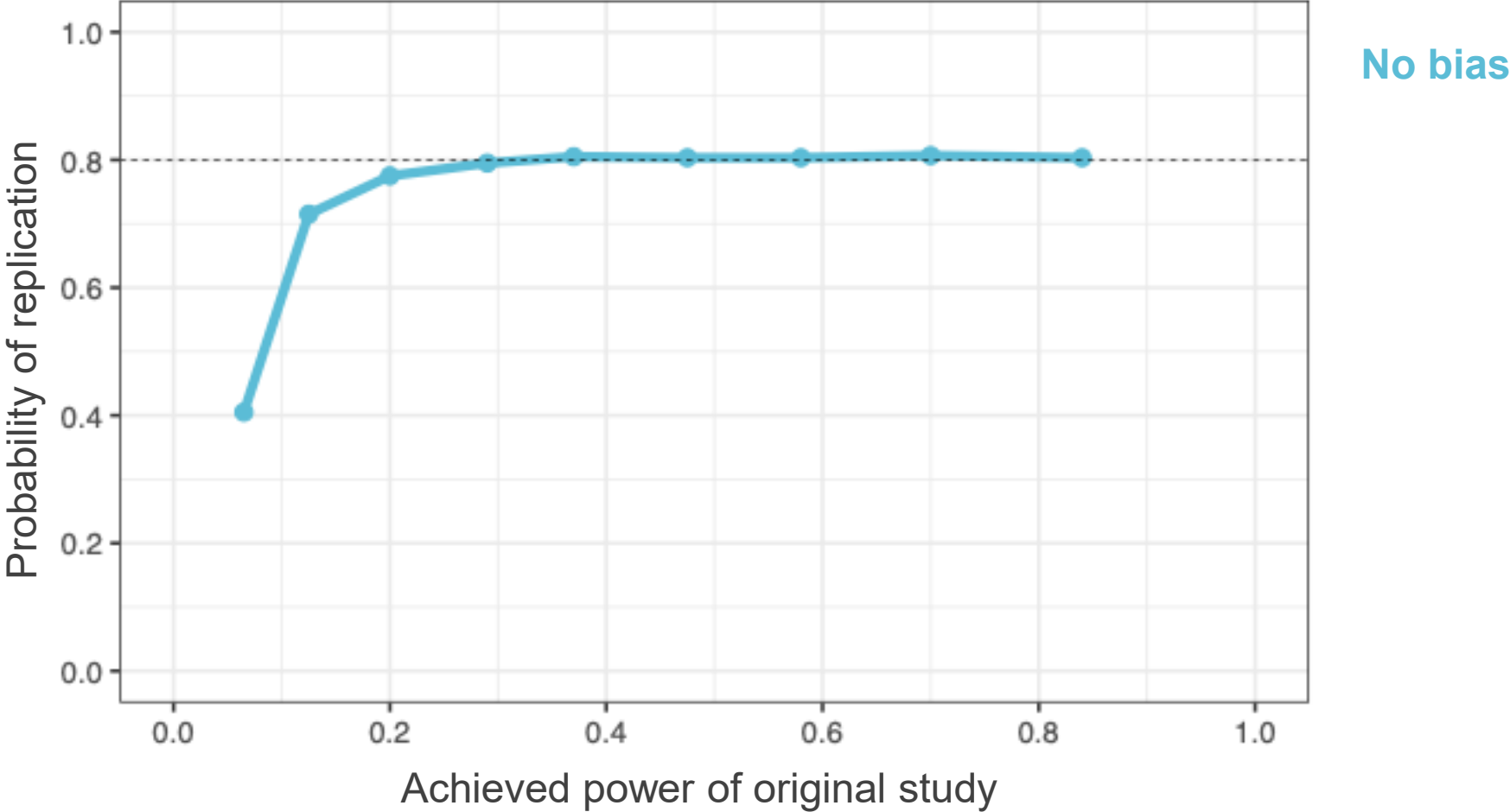
1. Conditional 80% power
2. 50% (80%) assurance of 80% power
3. Bias-adjusted 50% (80%) assurance of 80% power

4. Determine the probability of replicating a true finding

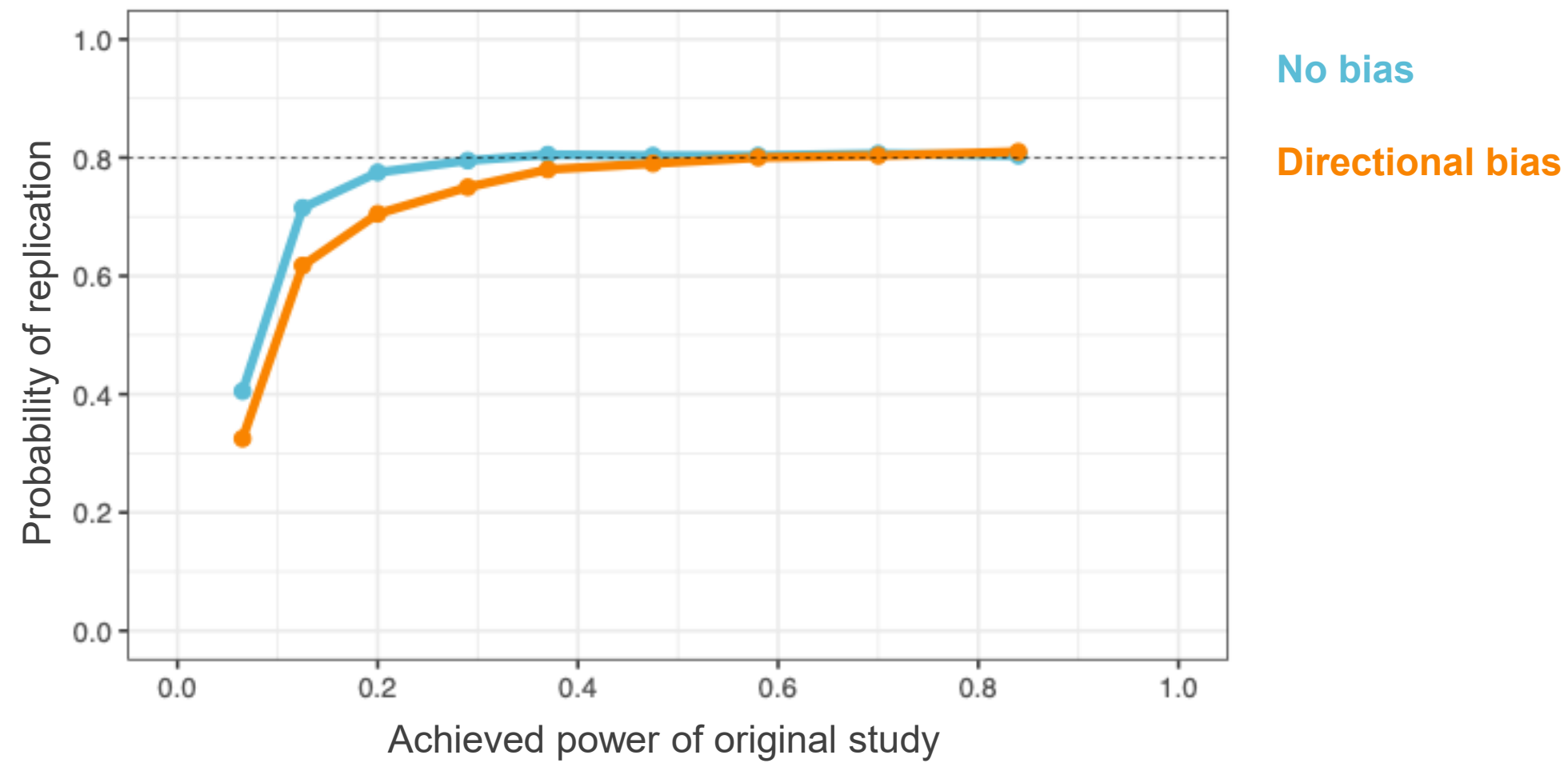
median power of using n to detect a stat. signif. difference between $\text{true_mean}_1 \pm \text{true_sd}_1$ and $\text{true_mean}_2 \pm \text{true_sd}_2$ with $p < 0.05$



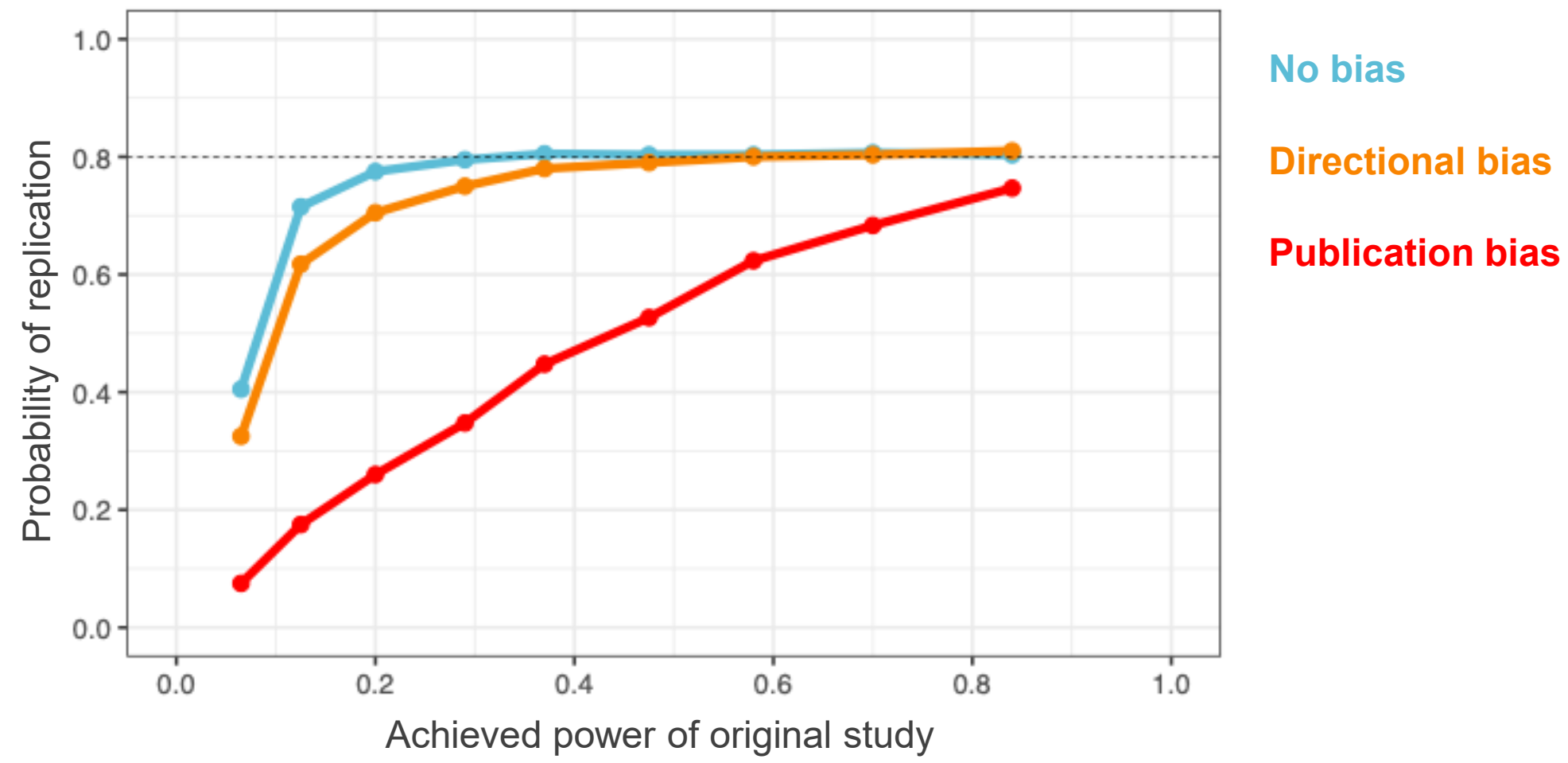
P(replication) using conditional 80% power under no bias



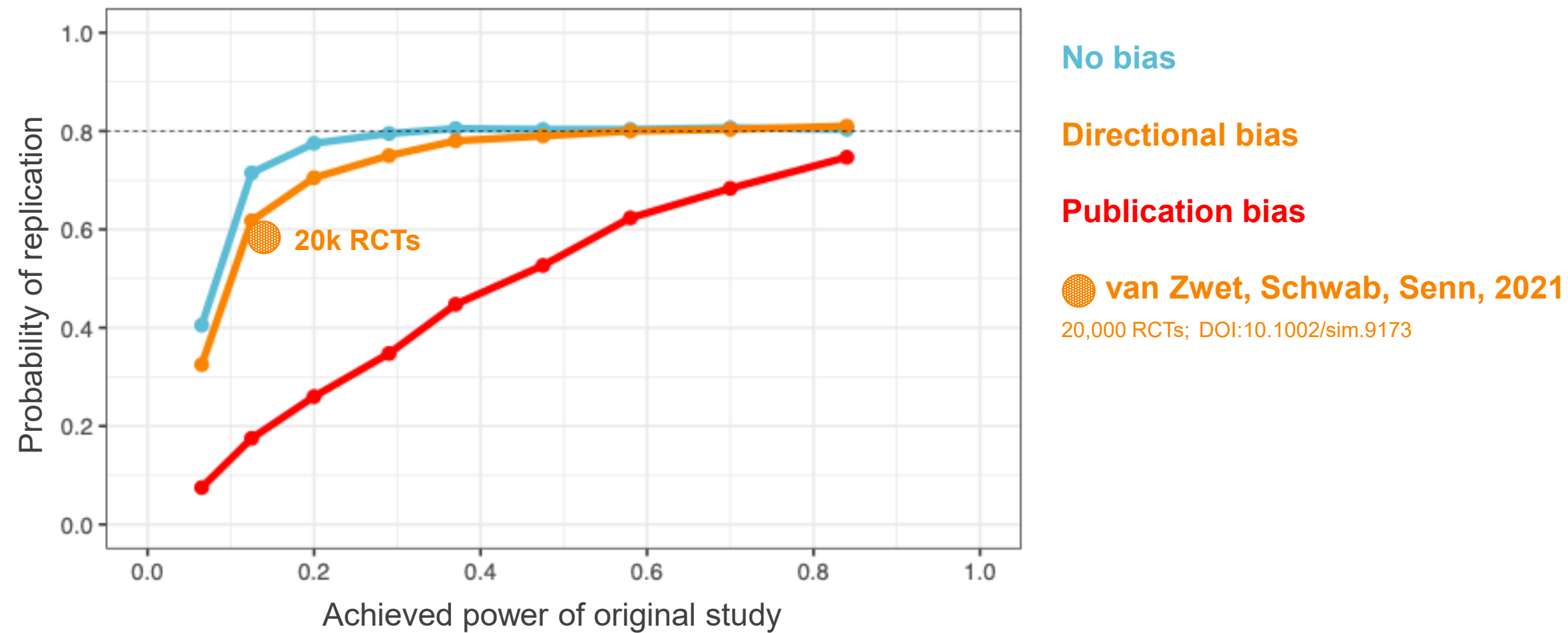
P(replication) using conditional 80% power under directional bias



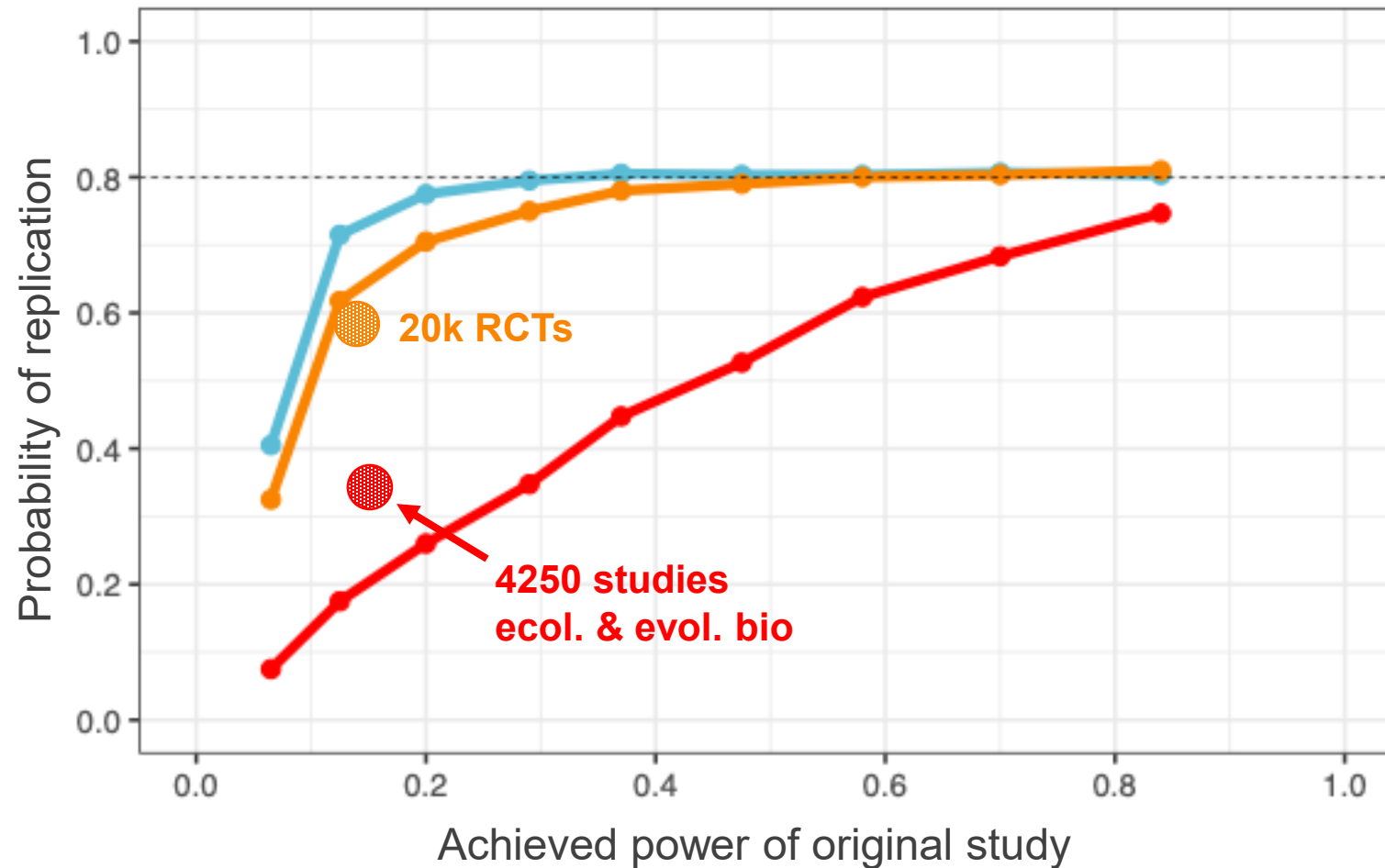
P(replication) using conditional 80% power under publication bias



Simulated probabilities are consistent with observed repl. rates



Simulated probabilities are consistent with observed repl. rates



No bias

Directional bias

Publication bias

● van Zwet, Schwab, Senn, 2021

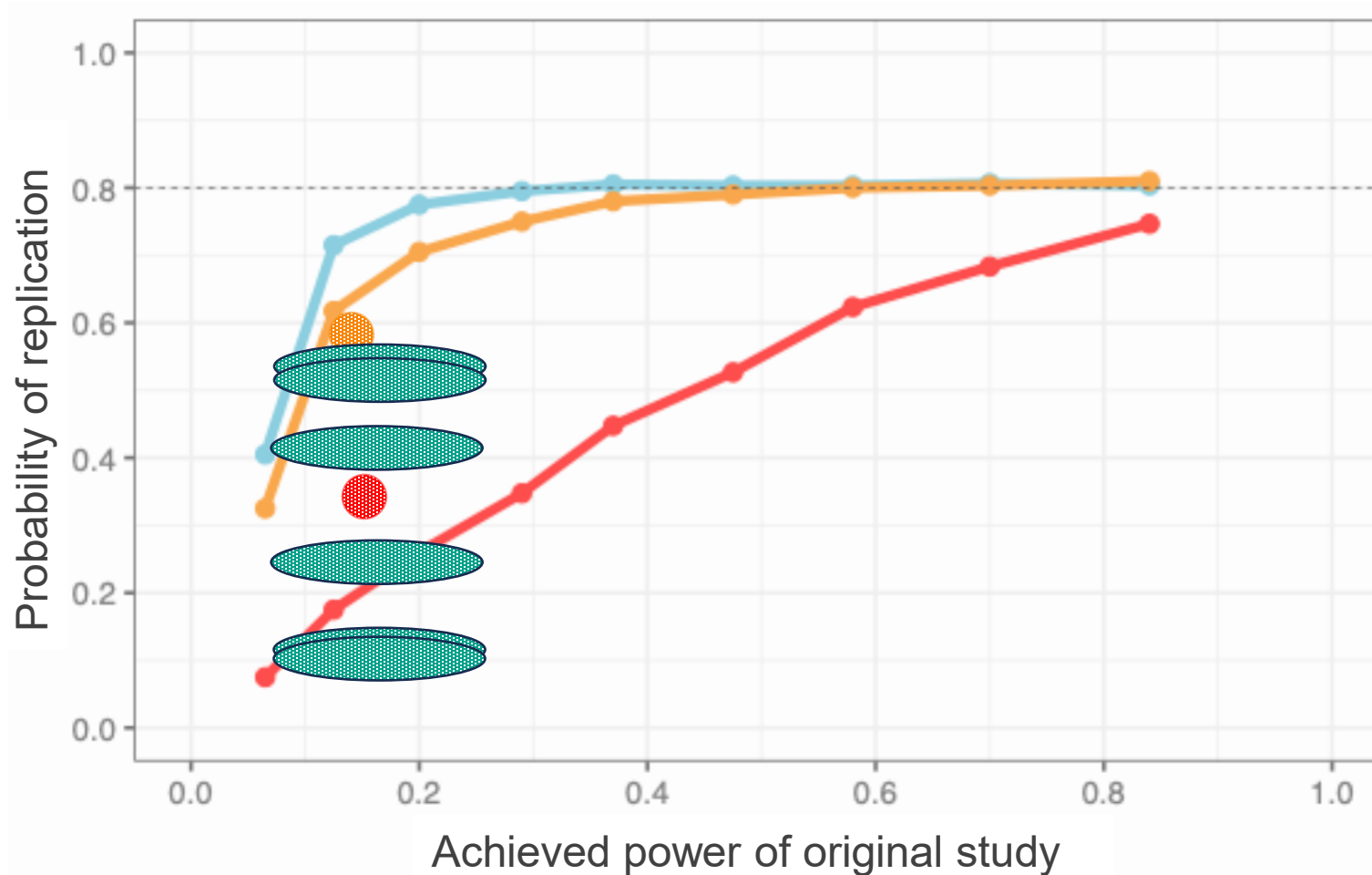
20,000 RCTs; DOI:10.1002/sim.9173

● Yang et al., 2023

4250 studies ecology and evol. biology

<https://doi.org/10.1186/s12915-022-01485-y>

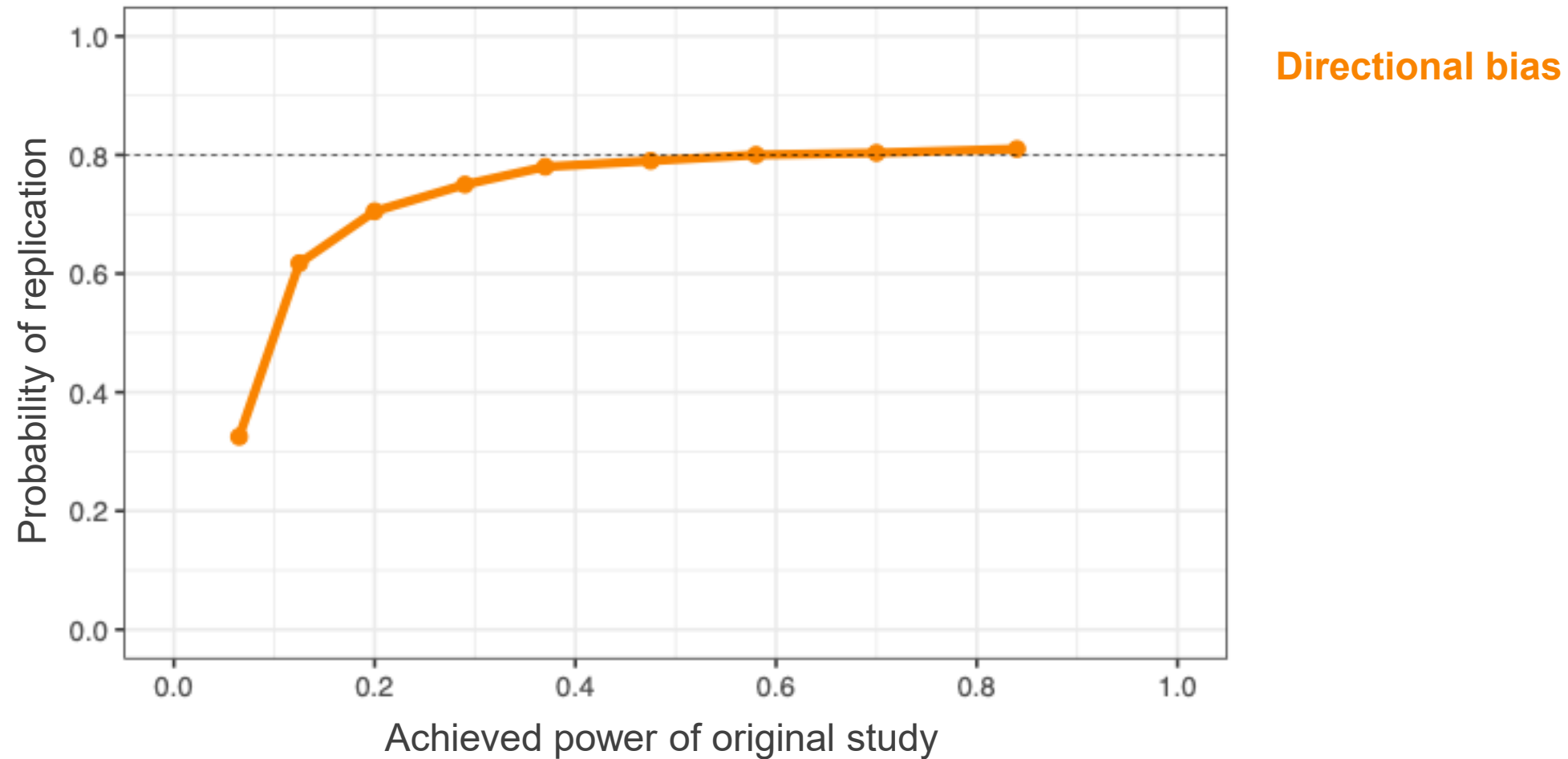
Publication bias is sufficient to explain low observed repl. rates



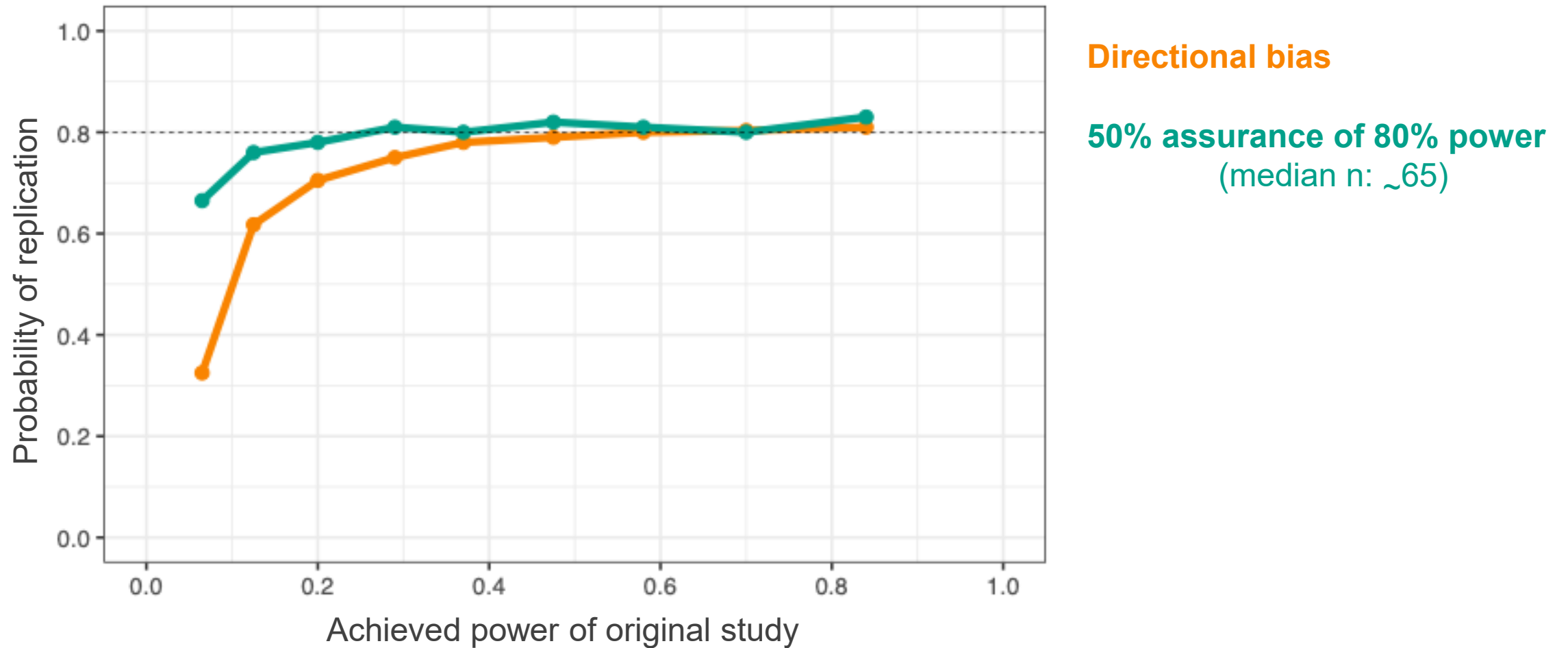
Observed replication rates:

- van Zwet, Schwab, Senn, 2021
doi.org/10.1002/sim.9173
- Yang et al., 2023 eco evo bio
doi.org/10.1186/s12915-022-01485-y
- Bartos et al., 2023 (psych.)
arXiv:2208.12334v3 [stat.AP] 26 Sep 2023
- Tyner et al., 2026 (soc bio)
doi.org/10.1038/s41586-025-10078-y
- Errington et al., 2021(onco)
doi.org/10.7554/eLife.71601
- Errington et al., 2021(animal)
doi.org/10.7554/eLife.71601
- Prinz et al., 2011
doi.org/10.1038/nrd3439-c1
- Begley & Ellis 2012 (onco)
doi.org/10.1038/483531a

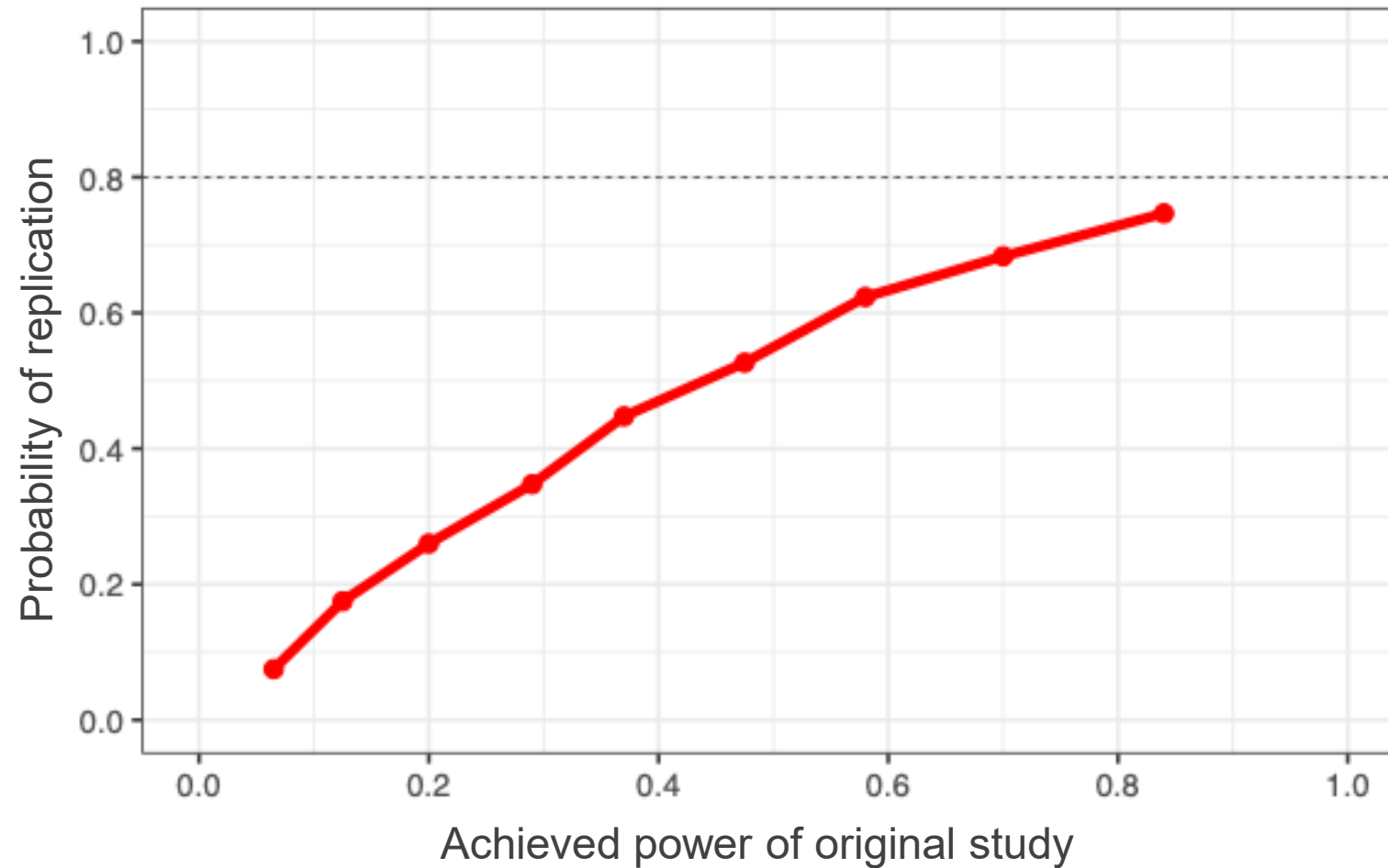
Does assurance improve $P(\text{replication})$ under directional bias?



Yes! Using assurance significantly increases P(replication)

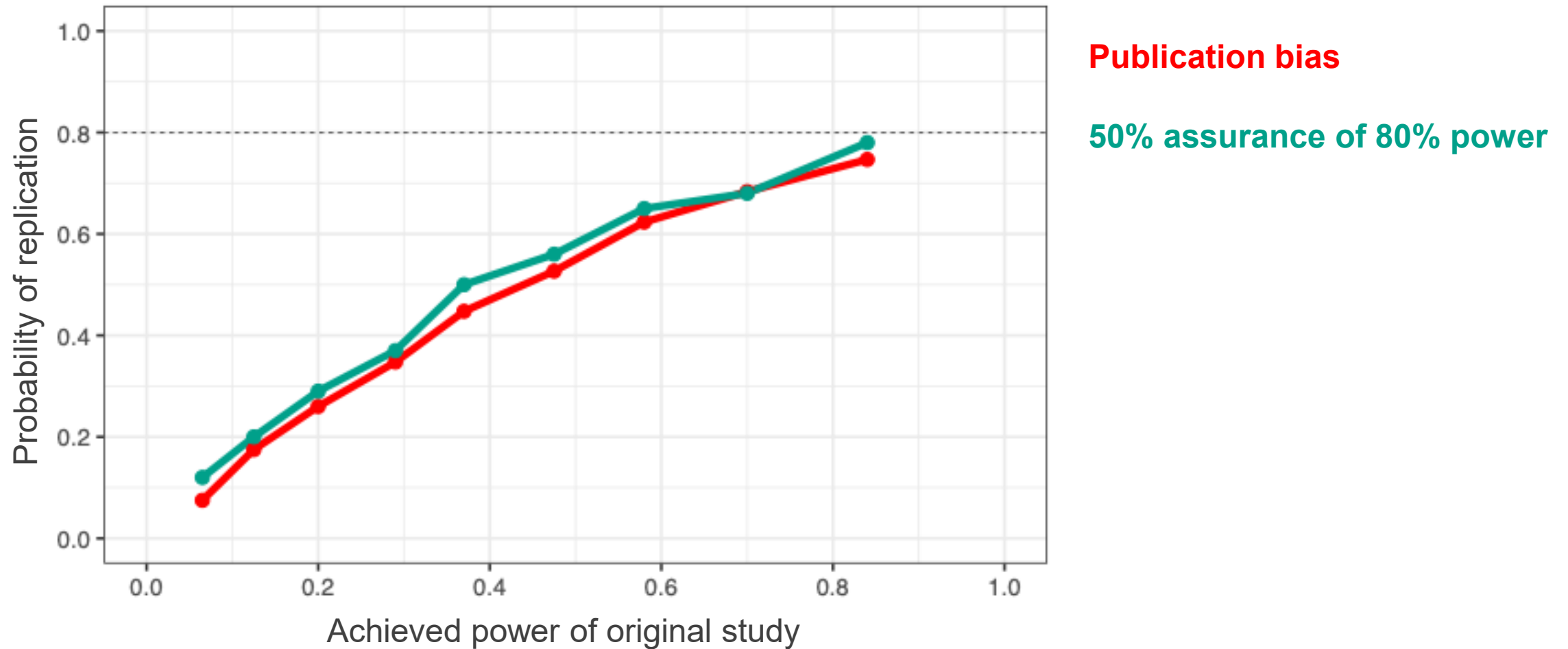


Does assurance improve $P(\text{replication})$ under publication bias?

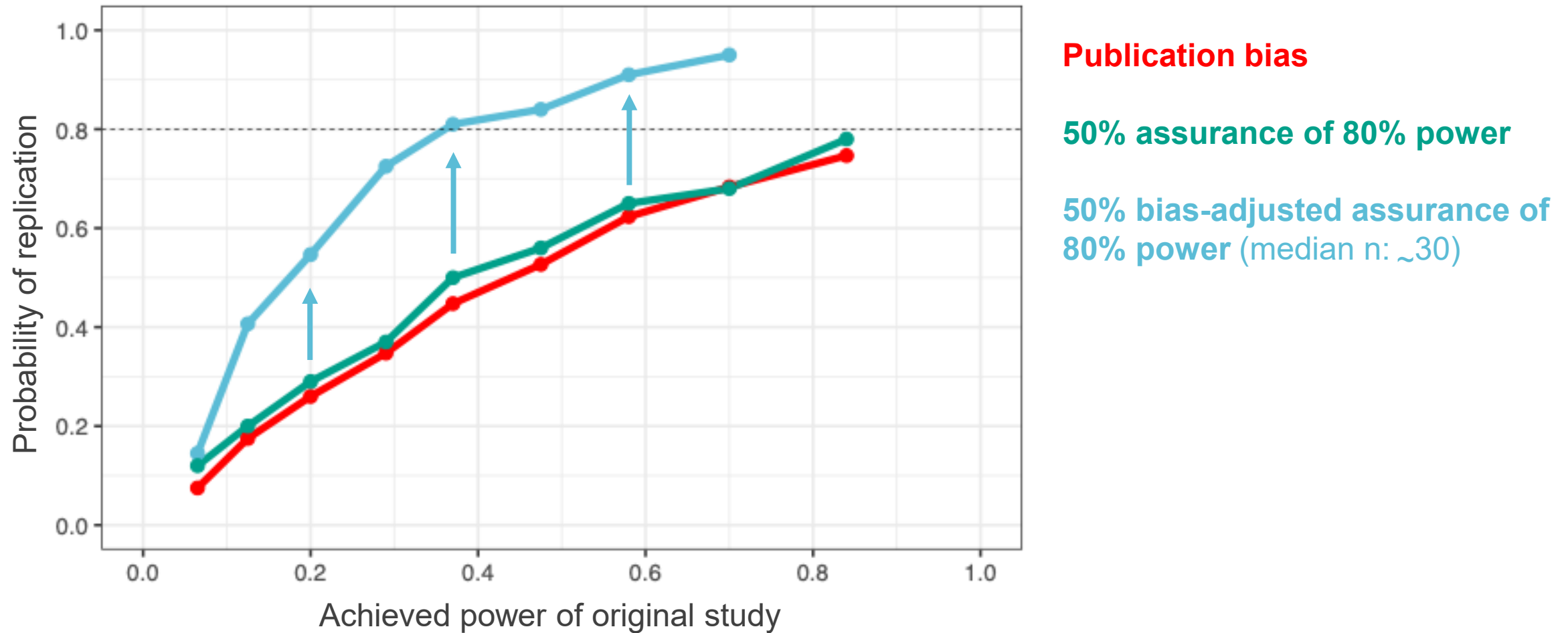


Publication bias

No. Assurance alone doesn't correct for publication bias, but...



Publication bias-adjusted assurance increases P(replication)!



Conclusions

- Failure to replicate **DOES NOT** imply that most published findings are false
- Replication often requires large sample sizes, especially for published studies
- Under-powered studies * publication bias → exaggerated effect sizes → under-powered replication studies → failure to replicate

