Replicability, effect sizes, & power

HESP Seminar 3/4/19

Phoebe Gaston & Hanna Muller

Motivation

- We've worried about how to draw conclusions from our studies and through worrying we've learned some (surprising) things.

- This is about small, simple steps that can increase our efficiency in conducting studies and decrease our uncertainty in interpreting them.

- The goal is for every experiment's outcome to be informative about what step to take next.

p-values

- Probability of observing a result this or more extreme given that the null hypothesis is true.

- We reject the null when the p-value is below a predetermined significance threshold (alpha).

Error

- Type 1 error: rejecting the null when it is true
    - false positive
- Type 2 error: failing to reject the null when it is false
    - false negative

Power

- If there is no true effect:
  - “alpha” = false positive rate (Type 1 error)
  - 1- alpha = true negative
- If there is a true effect:
  - “beta” = false negative rate (Type 2 error)
  - 1 - beta = true positive (statistical power)

Effect sizes

- Quantifies the difference between groups
- Cohen's d/Hedges' g = difference in standard deviations
  - (Mean 1 - Mean 2) / Pooled SD
- Many people argue that we should be estimating effect sizes rather than calculating p-values.

http://rpsychologist.com/d3/NHST/

How to calculate power?

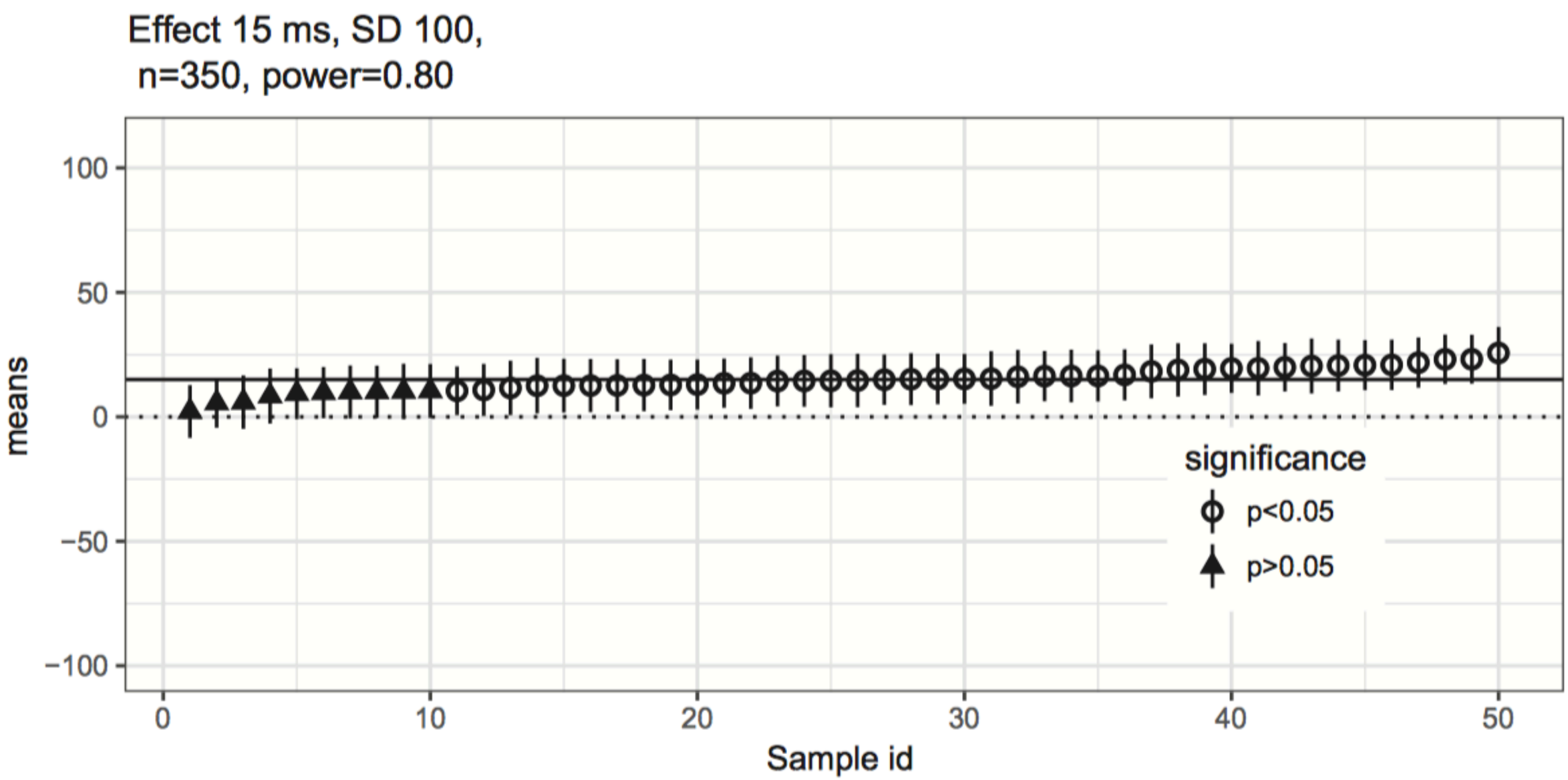
- Effect size requires t-value and sample size.

- Then plug into e.g. G*Power.

Why avoid low power?

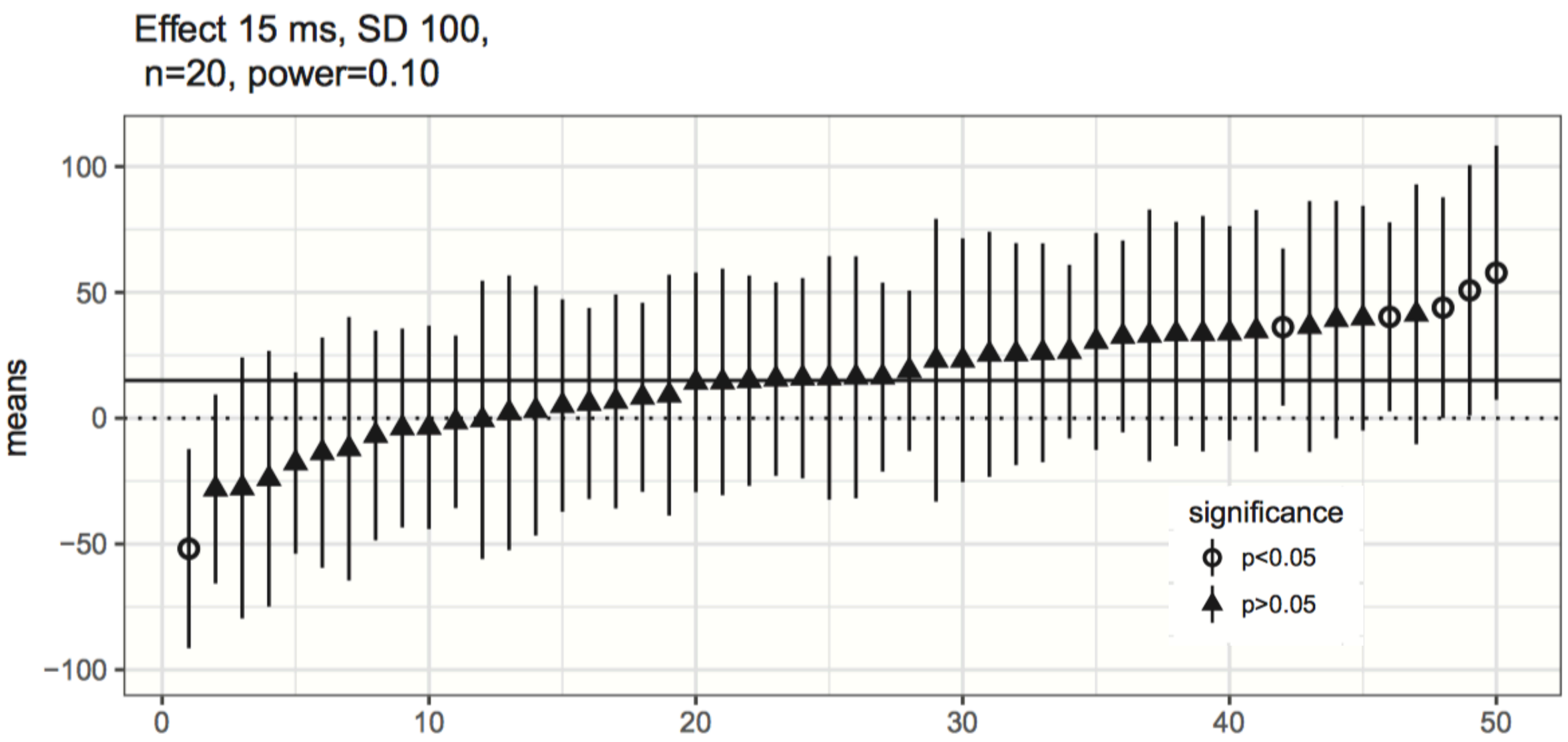
- Reason 1: (potentially) wasted time

- Reason 2: poor estimate of the true effect size

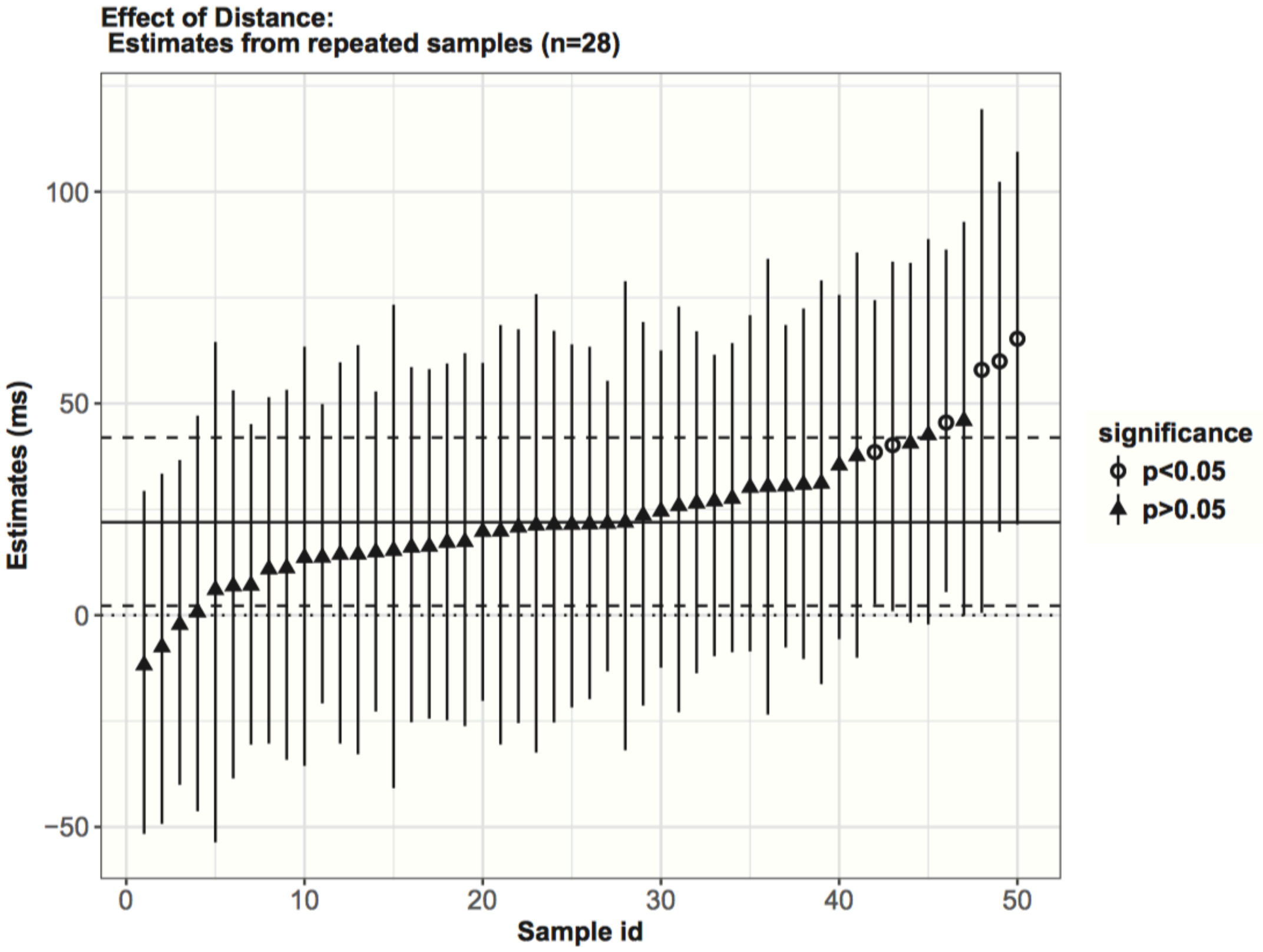


Vasishth et al. 2018

Type M error



Vasishth et al. 2018



Vasishth et al. 2018

Take-away

- If a published result comes from an under-powered study, it is likely an over-estimate.

- Using that effect size for a prospective power analysis will lead to under-estimates of the necessary sample size.

- This leads to more under-powered studies, which are unlikely to replicate the original finding.

(protected) Optional Stopping

- Plan on a large sample, but test every e.g. 20-30 participants.

- Adjust your alpha to account for the number of (predetermined) stops.

- Stop collecting data after first significant result.

See Lakens (2014)

Practical Questions

- What if you can't know the effect size for a prospective power analysis?
- How to decide if a prior result was underpowered?
- What if effect size is meaningless for your theory?

What next?

- Is it better to run an under-powered study than not run a study at all?
- Is running fewer studies the answer? Are high-powered studies also higher risk?
- Should we only study big effects? Avoid interactions?
- Does self-replication solve the problem?

A lot to gain

- We can (sometimes) collect smaller samples.
- We can find effects we wouldn't otherwise be able to detect with a typical sample size.
- We can avoid conducting experiments that are almost guaranteed to lead to null results.
- We can avoid puzzling over (meaningless) null results.