

# The role of replication in Bayesian data analysis

Shravan Vasishth

Cognitive Science / Linguistics

University of Potsdam, Germany

[vasishth.github.io](https://vasishth.github.io)

# The main points of this talk

1. Frequentist and Bayesian methods have different foci
2. Frequentist methods (NHST) only works well when power is high
3. Bayesian methods are useful when data are sparse
4. The replication crisis is partly driven by a “precision crisis” (low powered studies)
5. I will show how I use Bayesian methods to deal with the lack-of-precision problem in my field.

1. What I will say today holds for psycholinguistics; I don't speak for all of cognitive psychology.
2. There are many different approaches to using Bayesian methods; I will present my own perspective, which is influenced by Andrew Gelman's work.
3. This talk is **only** about data that can be replicated in principle. Examples of non-replicable “experiments”: political elections, football match outcomes, earthquakes.

# What is replication?

## **Standard frequentist definition:**

Repeatedly obtaining “significant” results

## **Bayesian (Gelman) perspective:**

“Consistency” across studies

Before discussing the role of replicability in Bayes, it is important to be clear on the differences between frequentist and Bayesian approaches.

# The frequentist procedure

Imagine that you have some independent and identically distributed data:  $x_1, x_2, \dots, x_n$

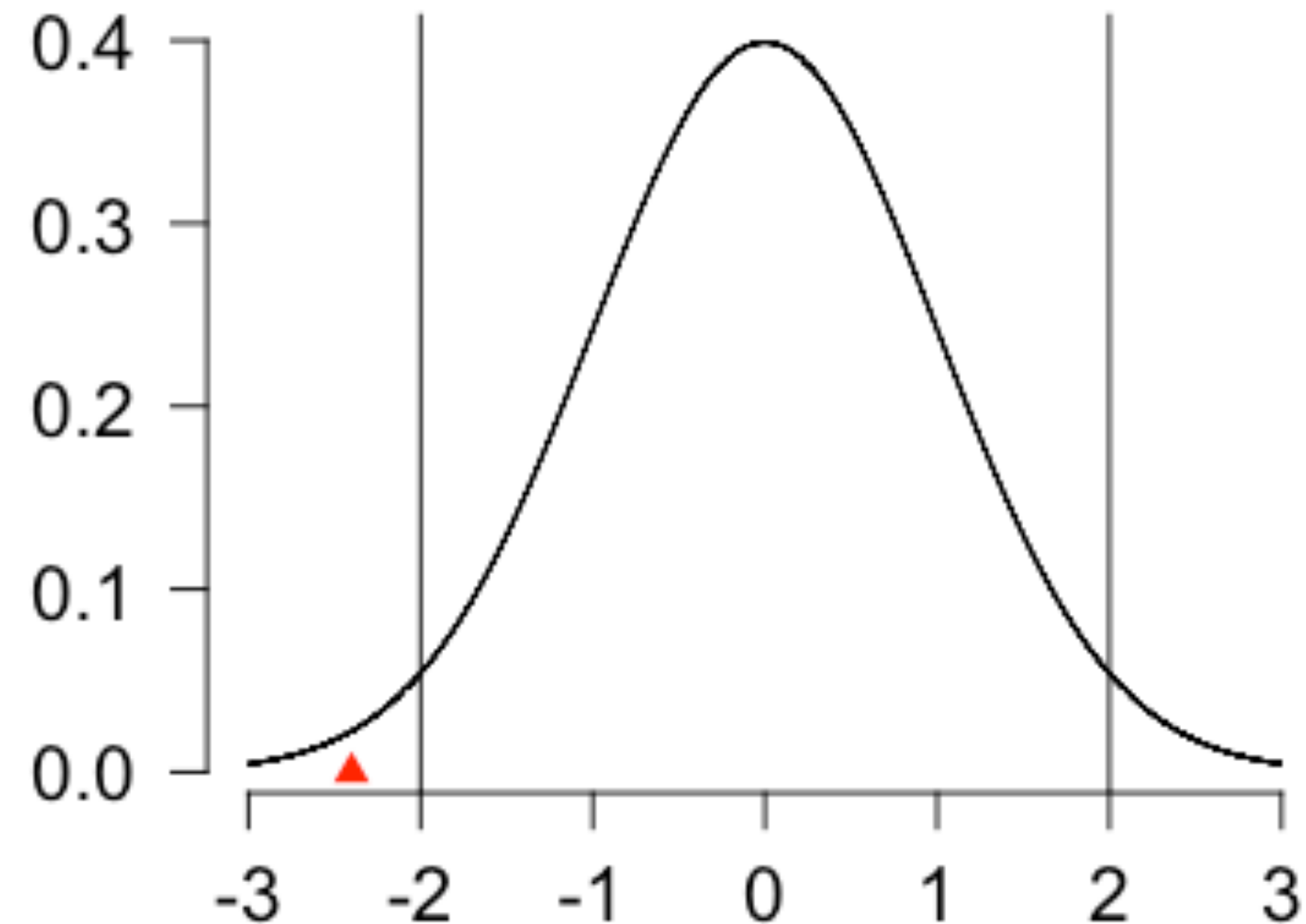
$$X \sim \text{Normal}(\mu, \sigma)$$

1. Set up a null hypothesis:  $H_0 : \mu = \mu_0$
2. Check how far sample mean  $\bar{x}$  is from  $\mu_0$  in SE units:  
$$t_{\text{observed}} \times SE = \bar{x} - \mu_0$$
3. If  $t_{\text{observed}}$  is large enough, reject null hypothesis

Statistical data analysis is reduced to checking for significance (is  $p < 0.05$ ?)

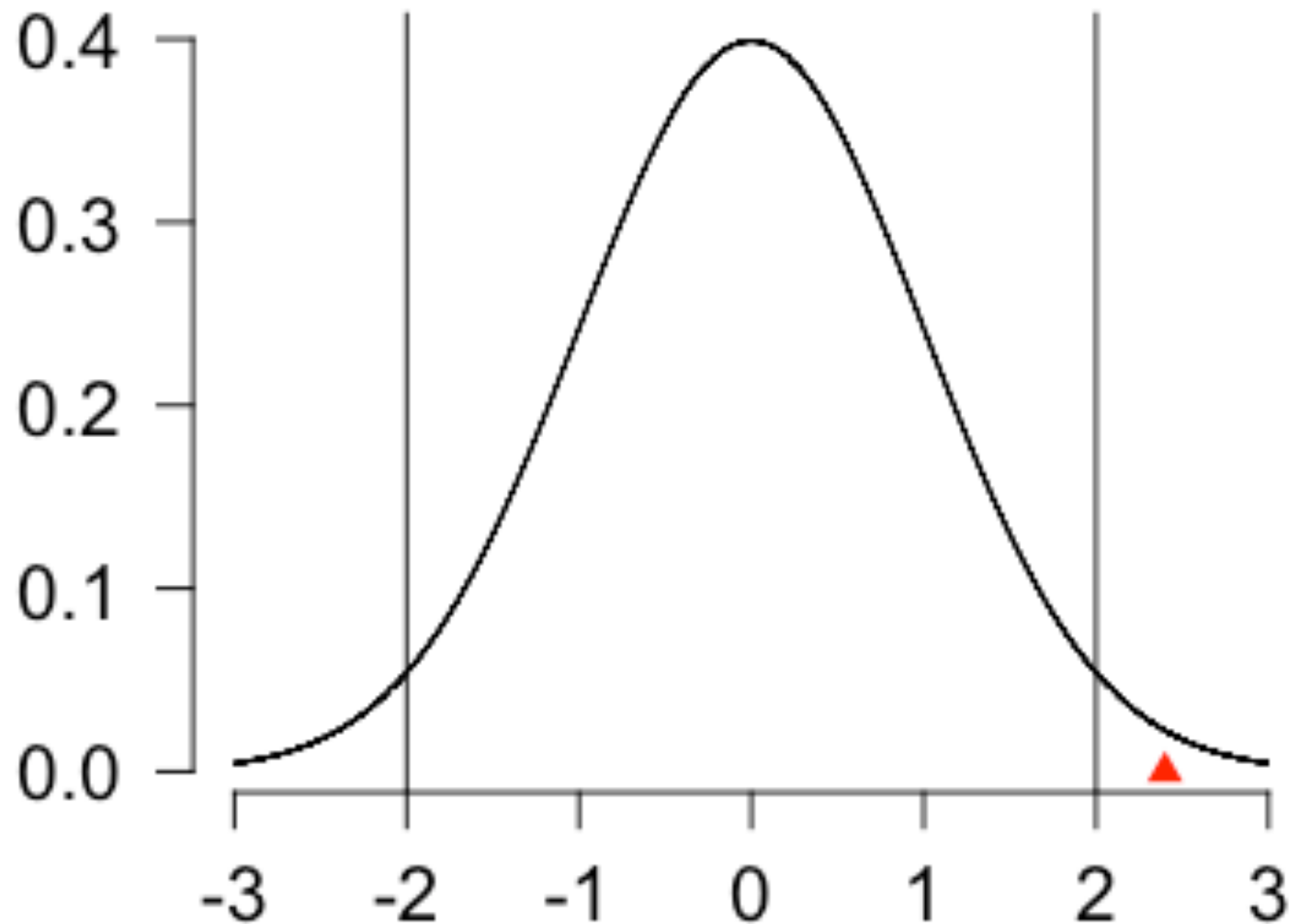
# The frequentist procedure

Decision: Reject null and publish



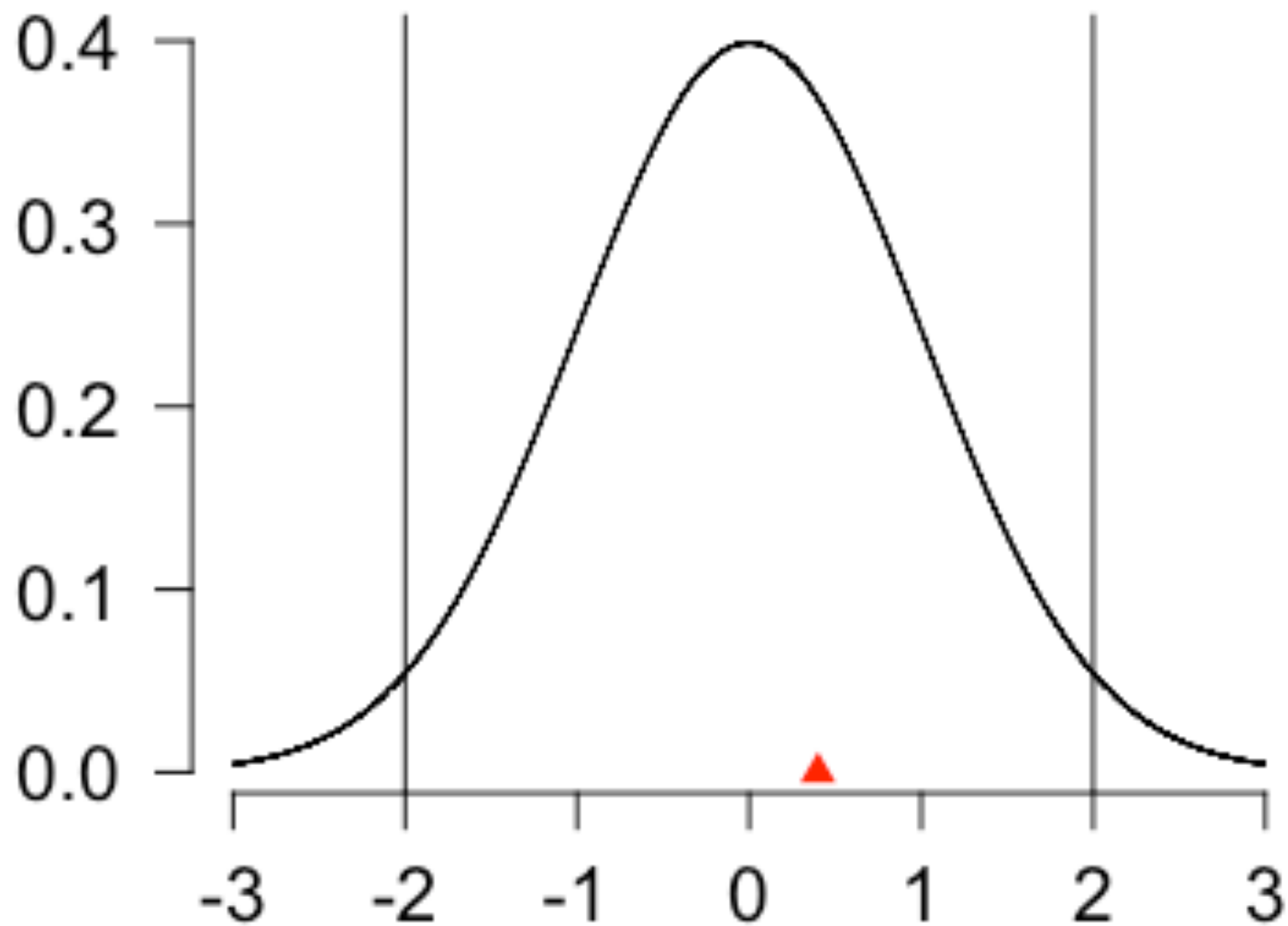
# The frequentist procedure

Decision: Reject null and publish



# The frequentist procedure

Accept null?





# The frequentist procedure

Power: the probability of detecting a particular effect

Power depends on:

- effect size
- standard deviation(s)
- sample size

The frequentist paradigm works well when power is high (80% or higher).

**The frequentist paradigm is not designed to be used in low power situations.**

# Example: agreement attraction

Journal of Memory and Language 111 (2020) 104063



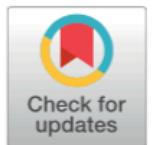
Contents lists available at [ScienceDirect](#)

Journal of Memory and Language

journal homepage: [www.elsevier.com/locate/jml](http://www.elsevier.com/locate/jml)



## Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study<sup>☆</sup>



Lena A. Jäger<sup>a,b</sup>, Daniela Mertzen<sup>b</sup>, Julie A. Van Dyke<sup>c</sup>, Shravan Vasishth<sup>b,\*</sup>

<sup>a</sup> Department of Computer Science, University of Potsdam, Germany

<sup>b</sup> Department of Linguistics, University of Potsdam, Germany

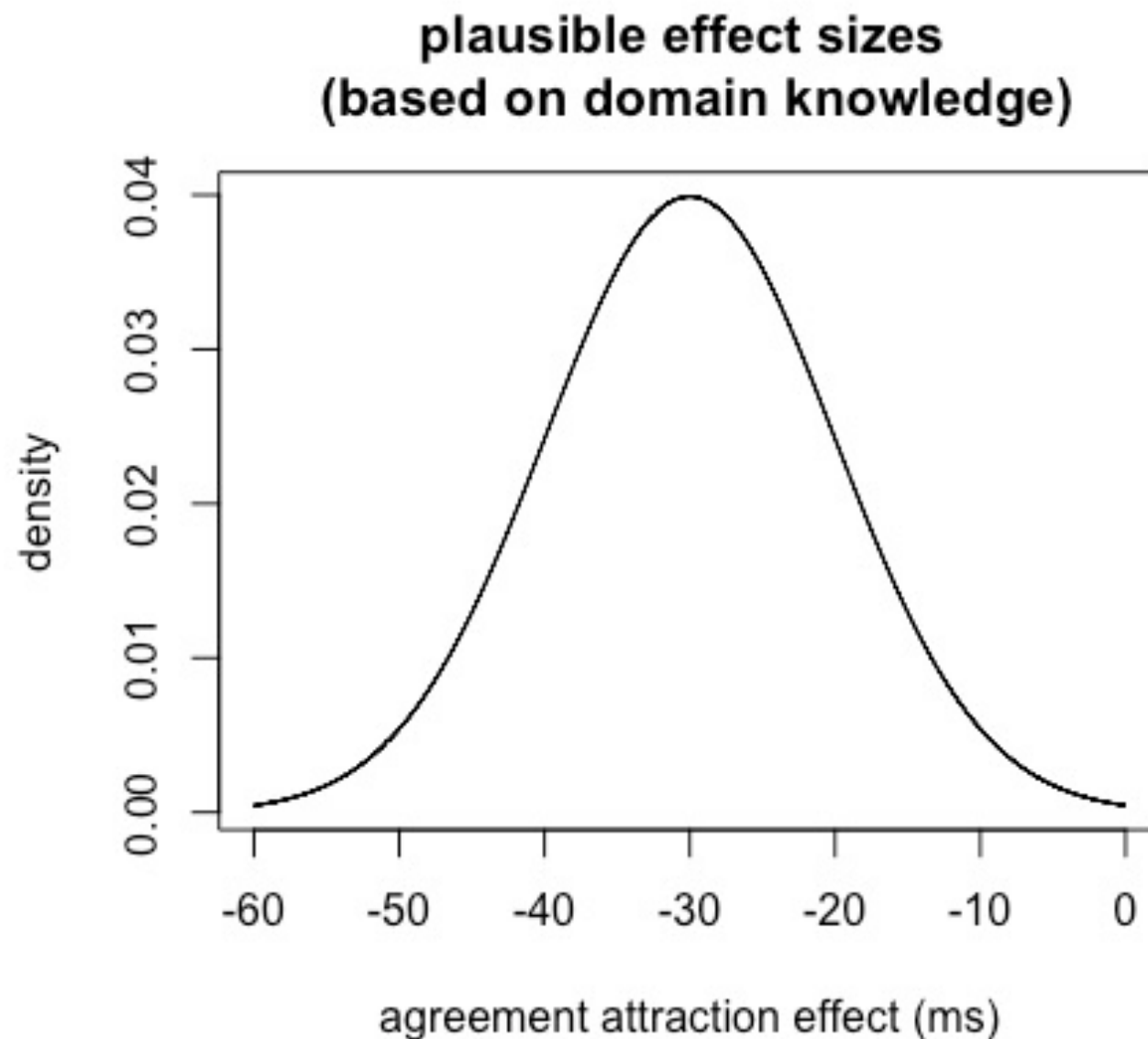
<sup>c</sup> Haskins Laboratories, New Haven, CT, United States

# Example: agreement attraction

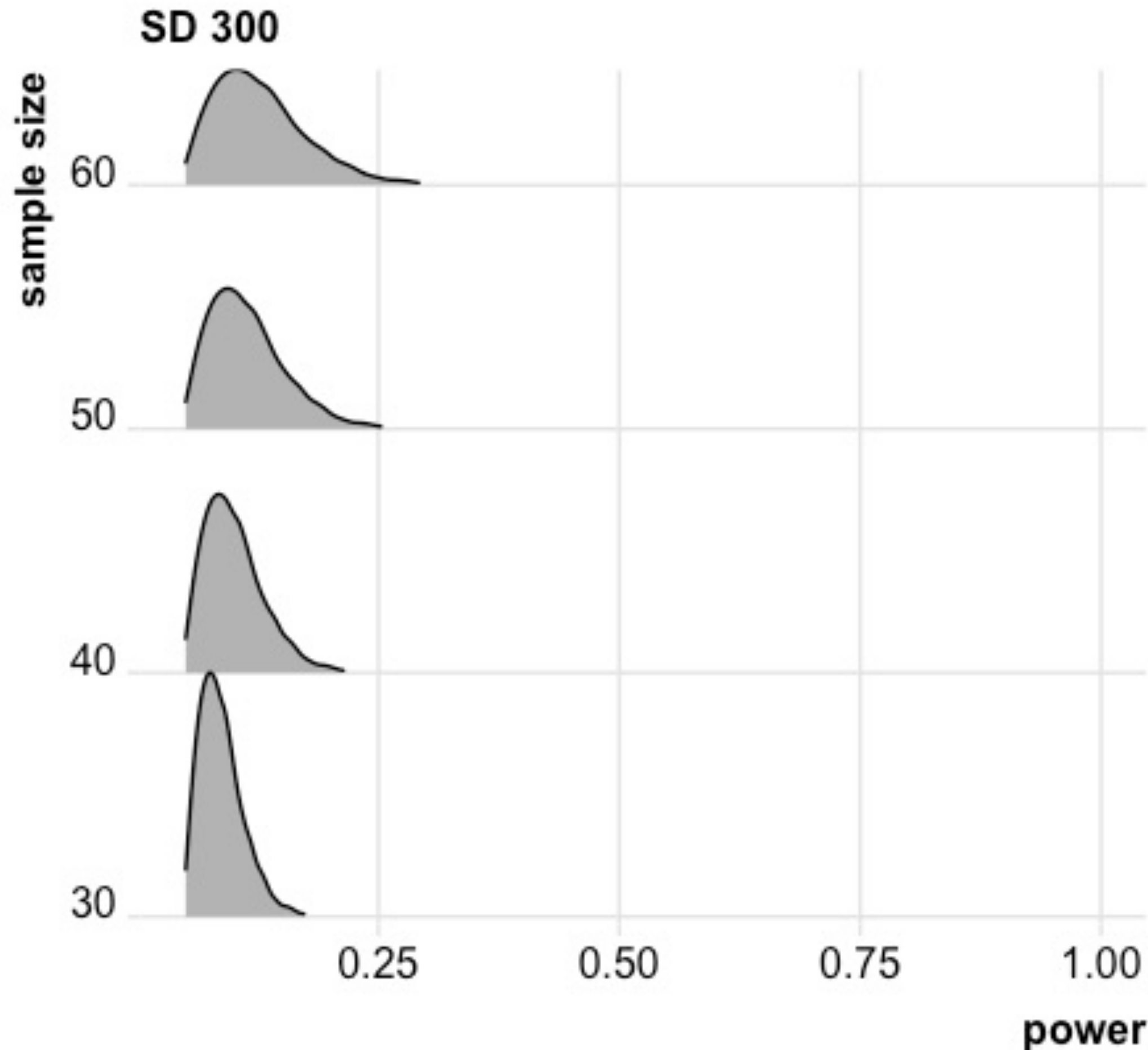
a. \*The bodybuilder<sub>+subject</sub><sup>-plural</sup> who met the trainers<sub>-subject</sub><sup>+plural</sup> were<sub>subject</sub><sup>plural</sup>  
...

b. \*The bodybuilder<sub>+subject</sub><sup>-plural</sup> who met the trainer<sub>-subject</sub><sup>-plural</sup> were<sub>subject</sub><sup>plural</sup>  
...

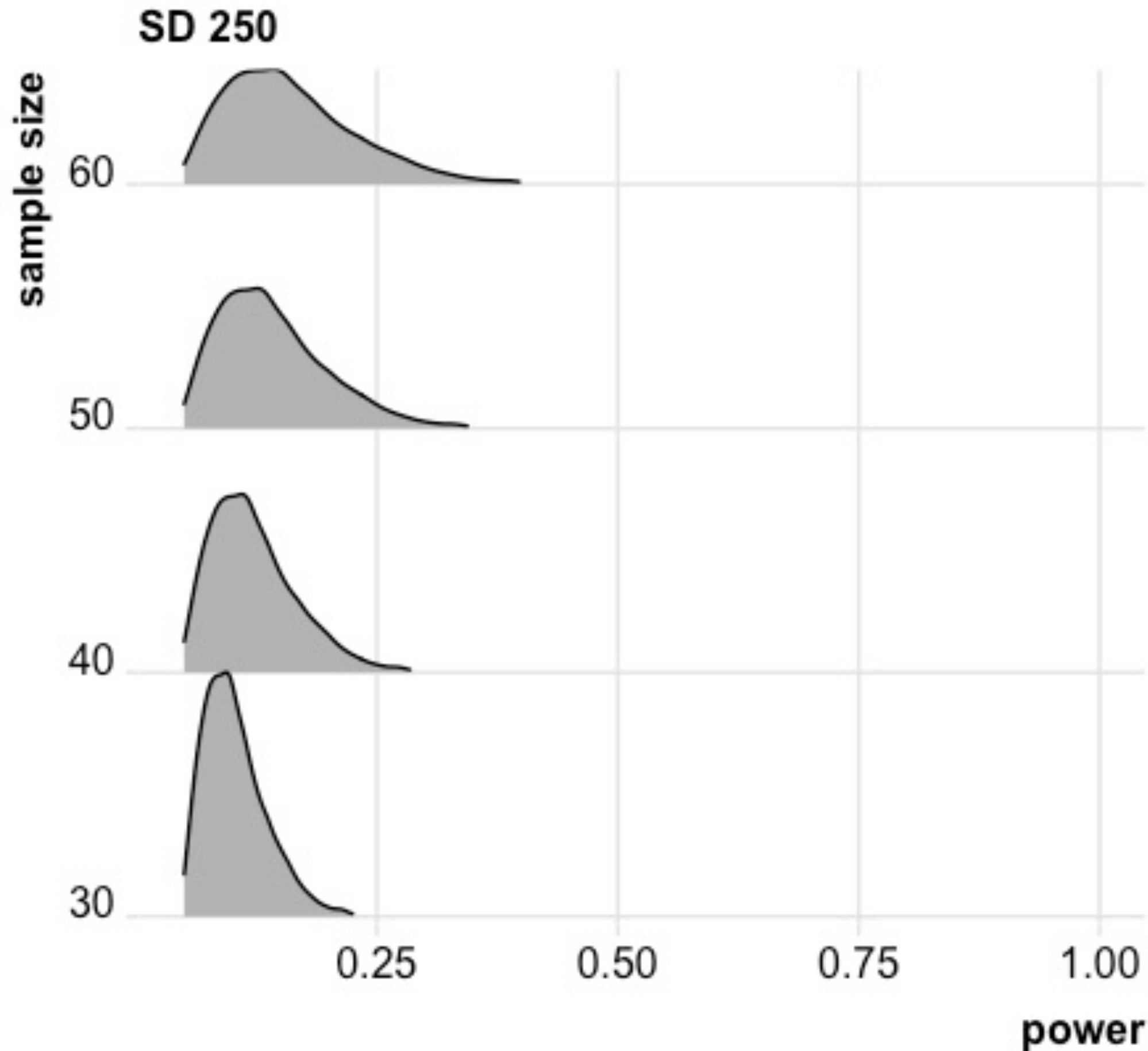
**b read faster than a at *were***



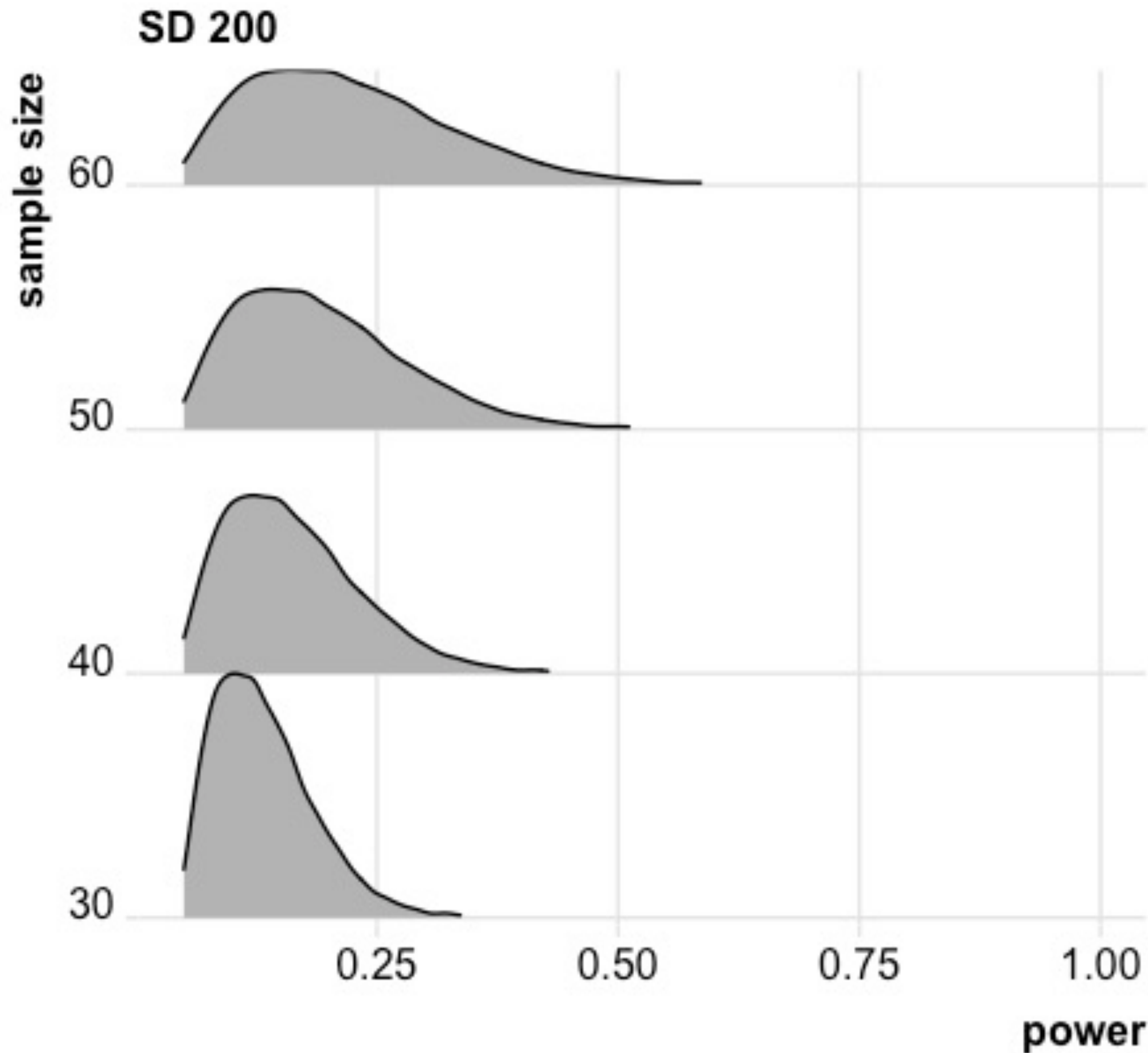
# Power in reading studies on agreement attraction



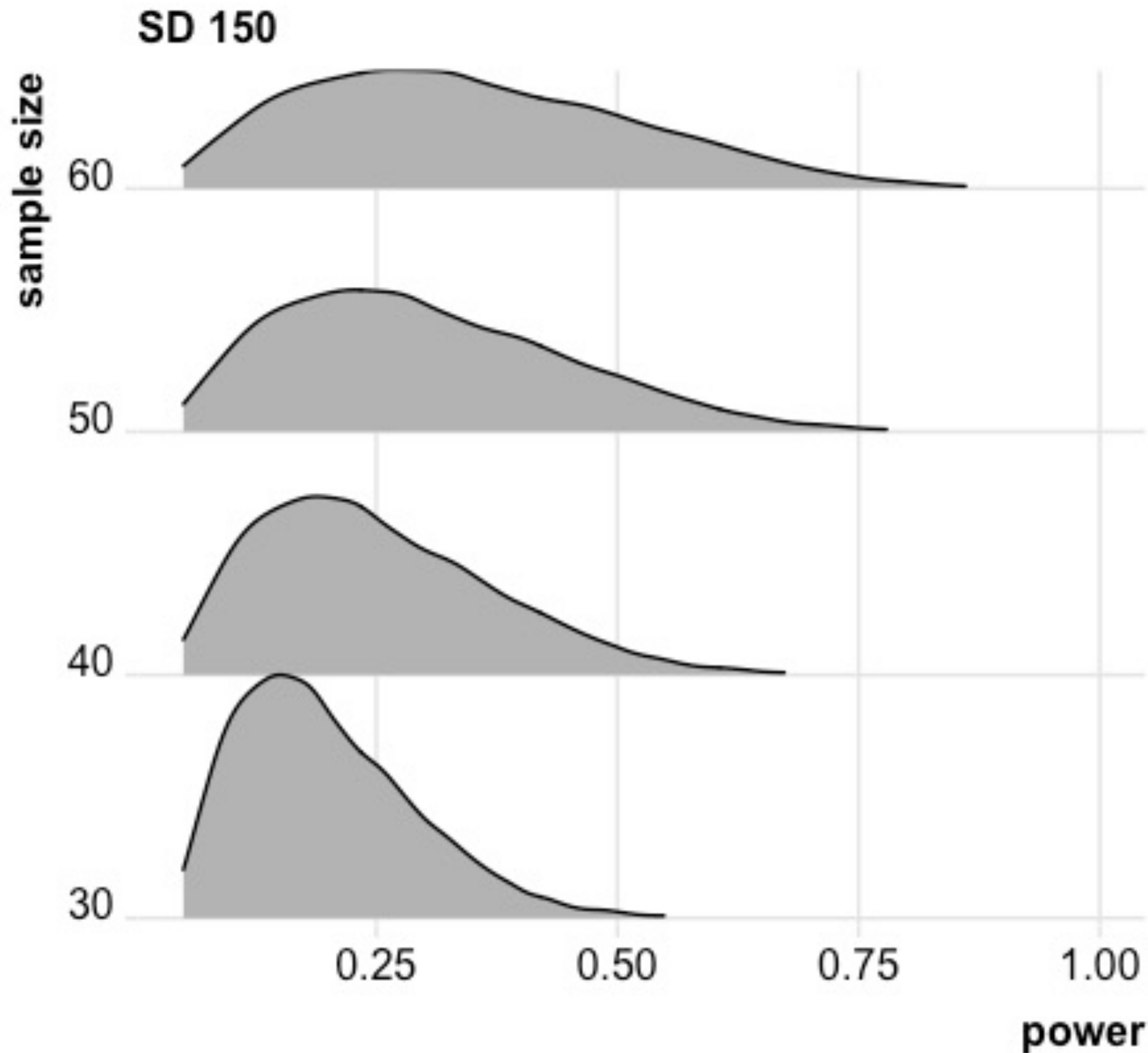
# Power in reading studies on agreement attraction



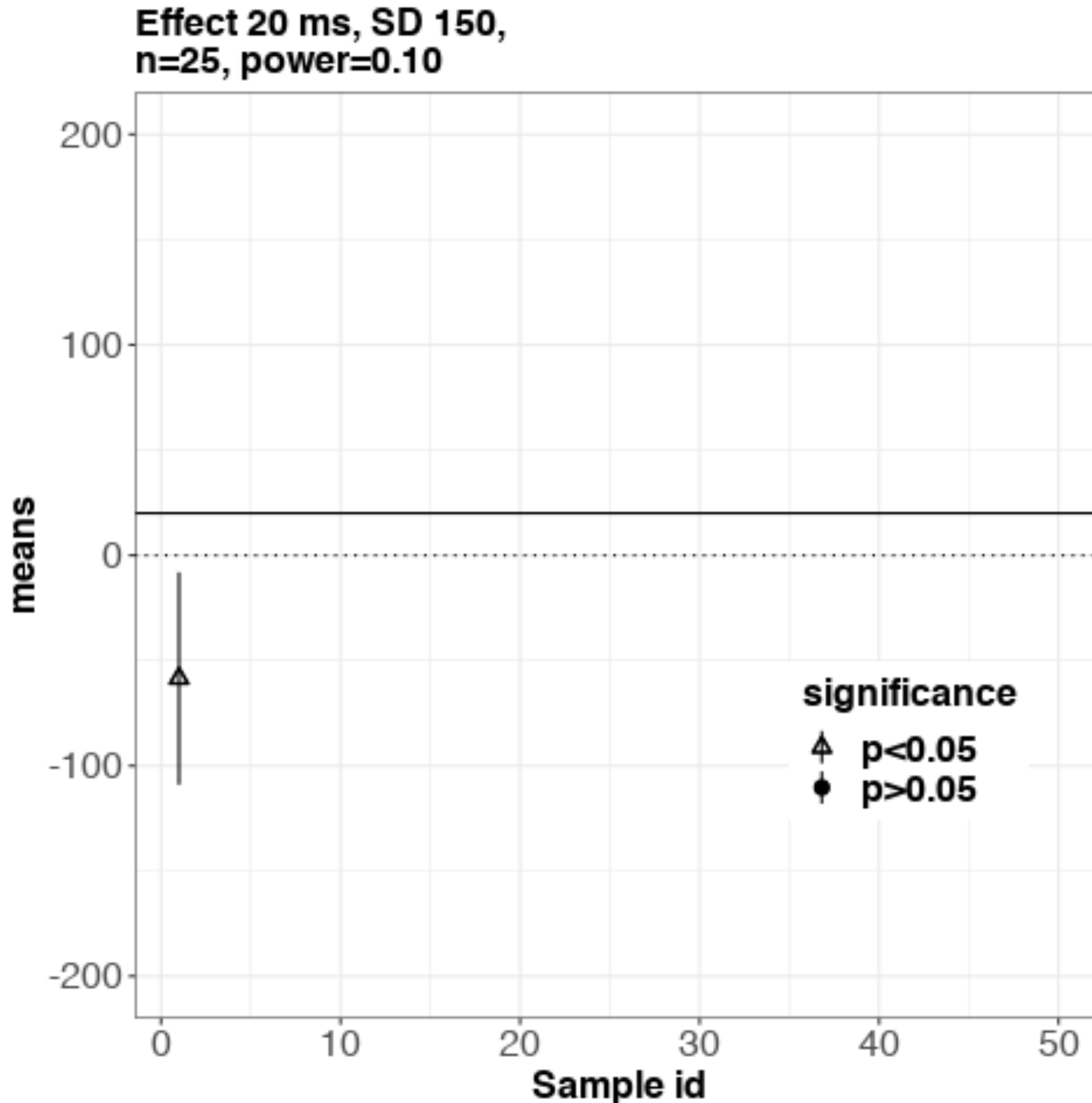
# Power in reading studies on agreement attraction



# Power in reading studies on agreement attraction

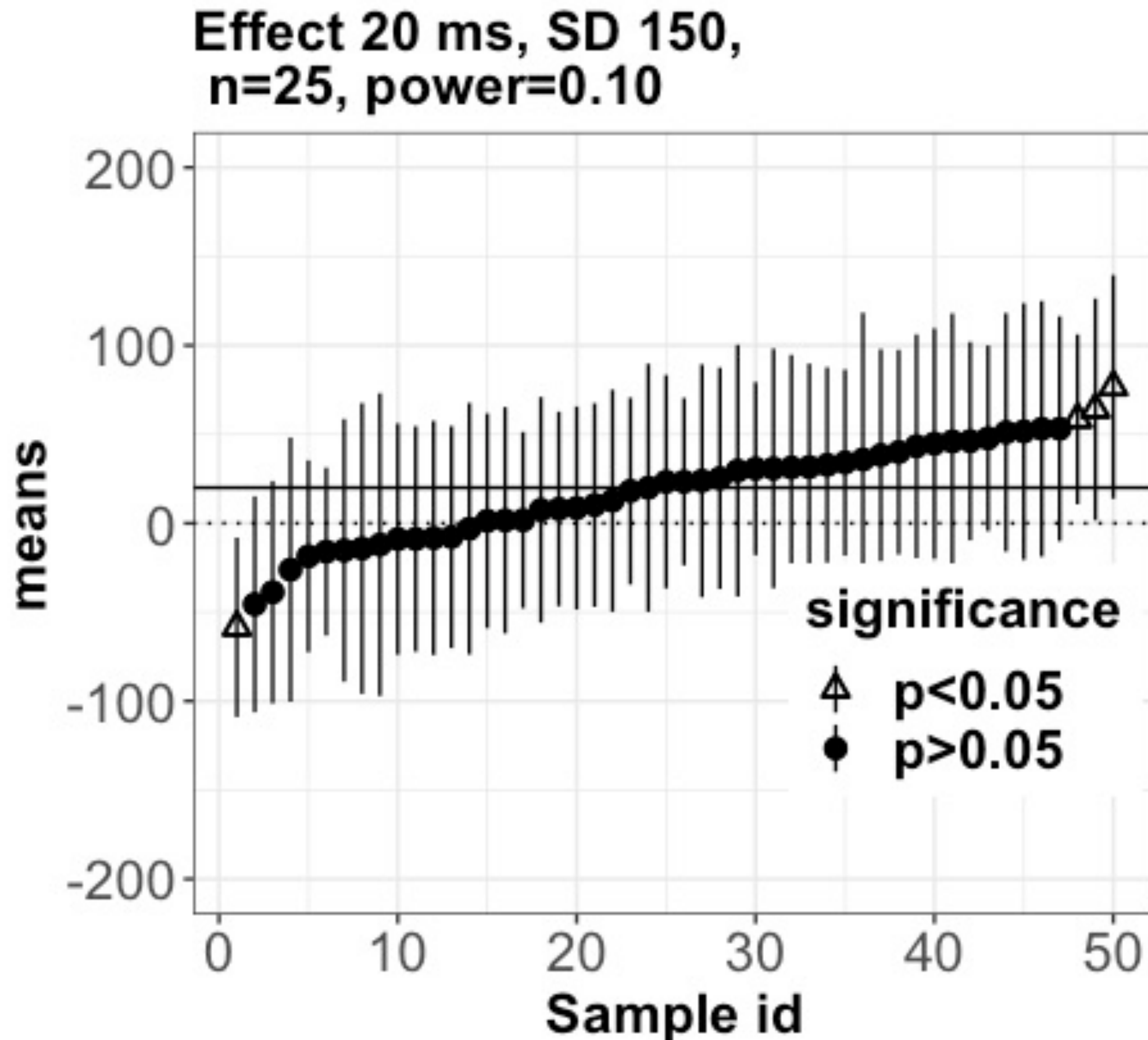


# Low power leads to exaggerated estimates: Type M error

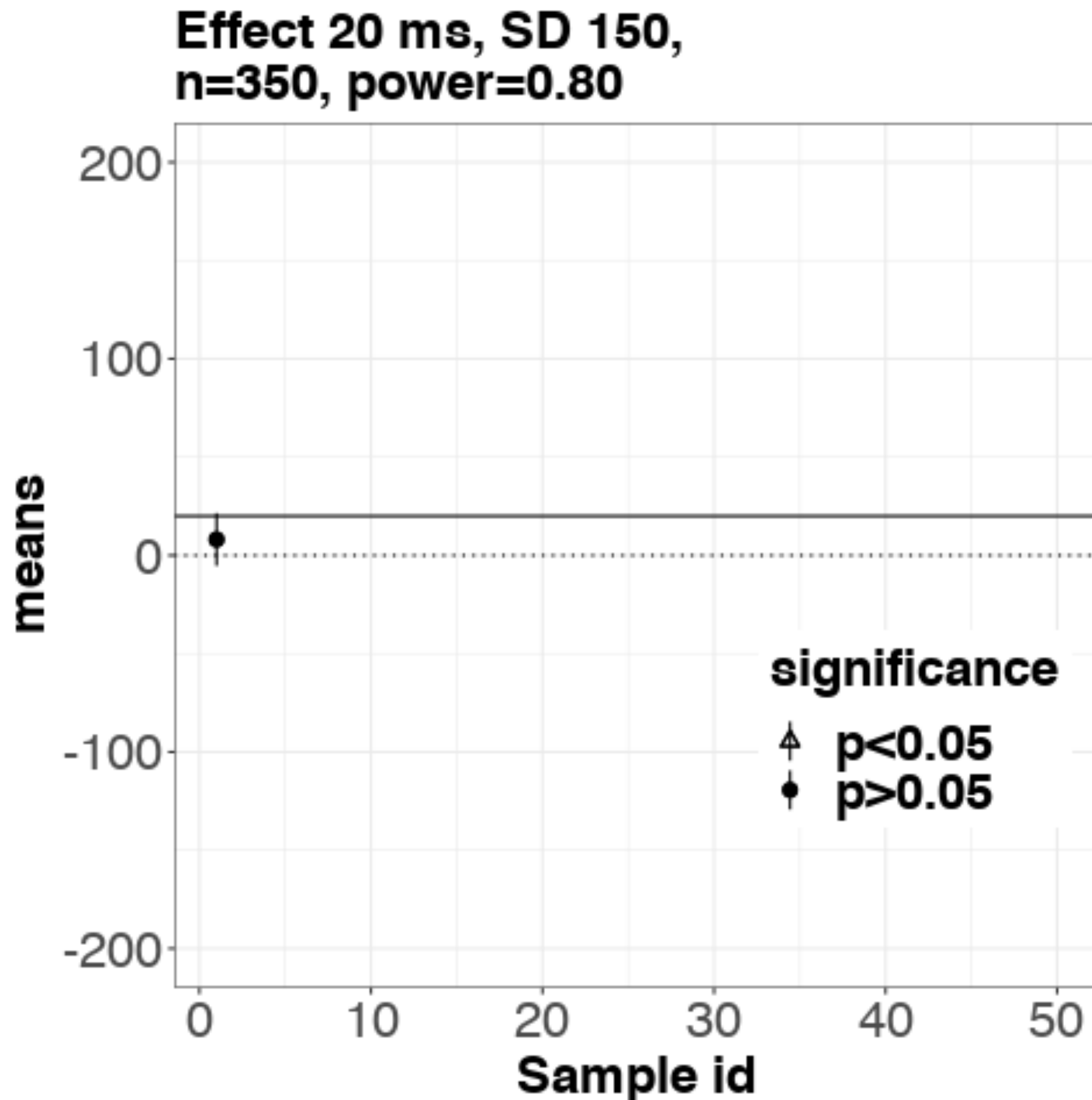




# Low power leads to exaggerated estimates: Type M error

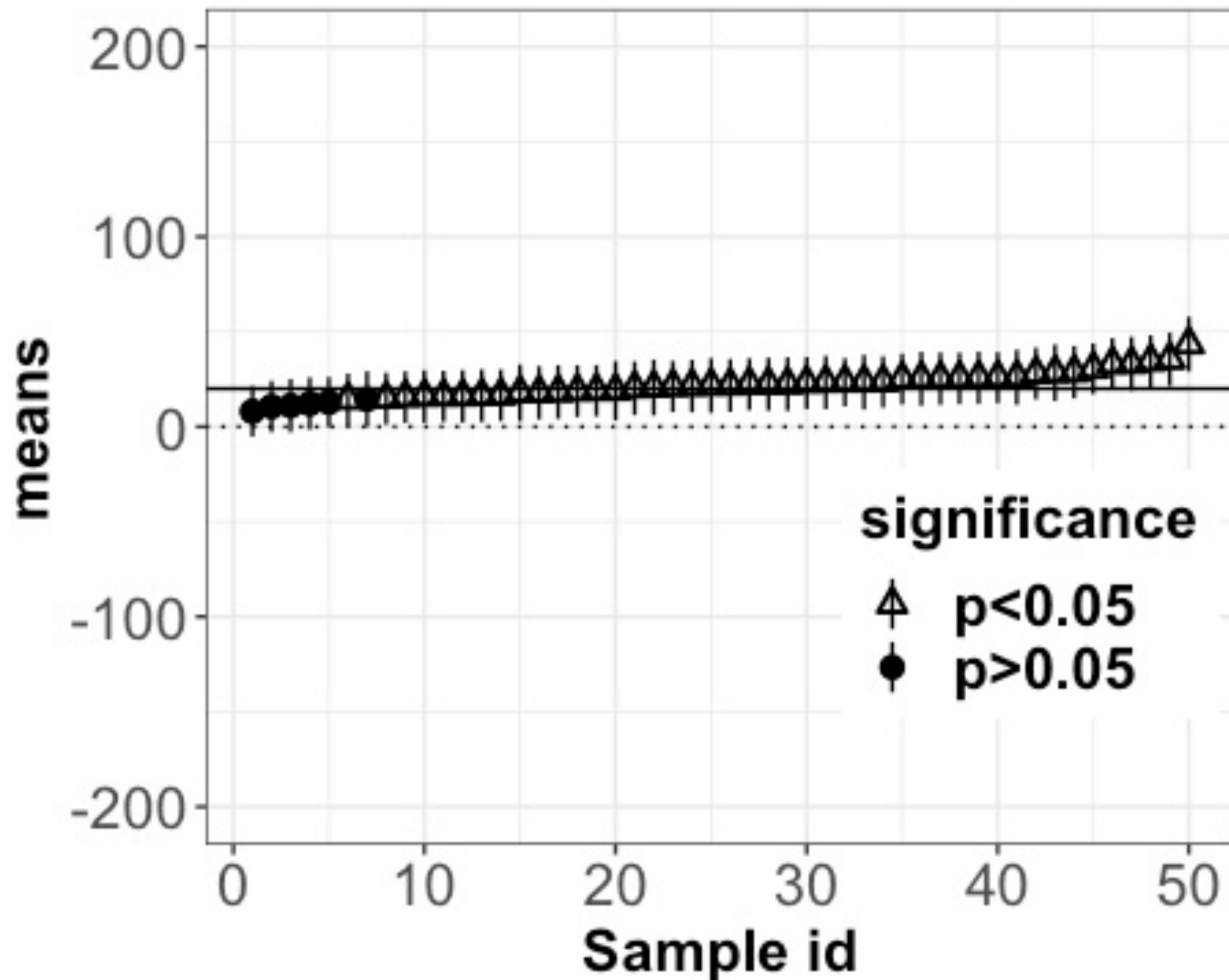


Compare with a high power situation



Compare with a high power situation

**Effect 20 ms, SD 150,  
n=350, power=0.80**



# The frequentist paradigm breaks down when power is low

1. Null results are inconclusive
2. Significant results are based on biased estimates  
(Type M error)

Consequences:

1. Non-replicable results
2. Incorrect inferences

# The Bayesian approach

Imagine that you have some independent and identically distributed data:  $x_1, x_2, \dots, x_n$

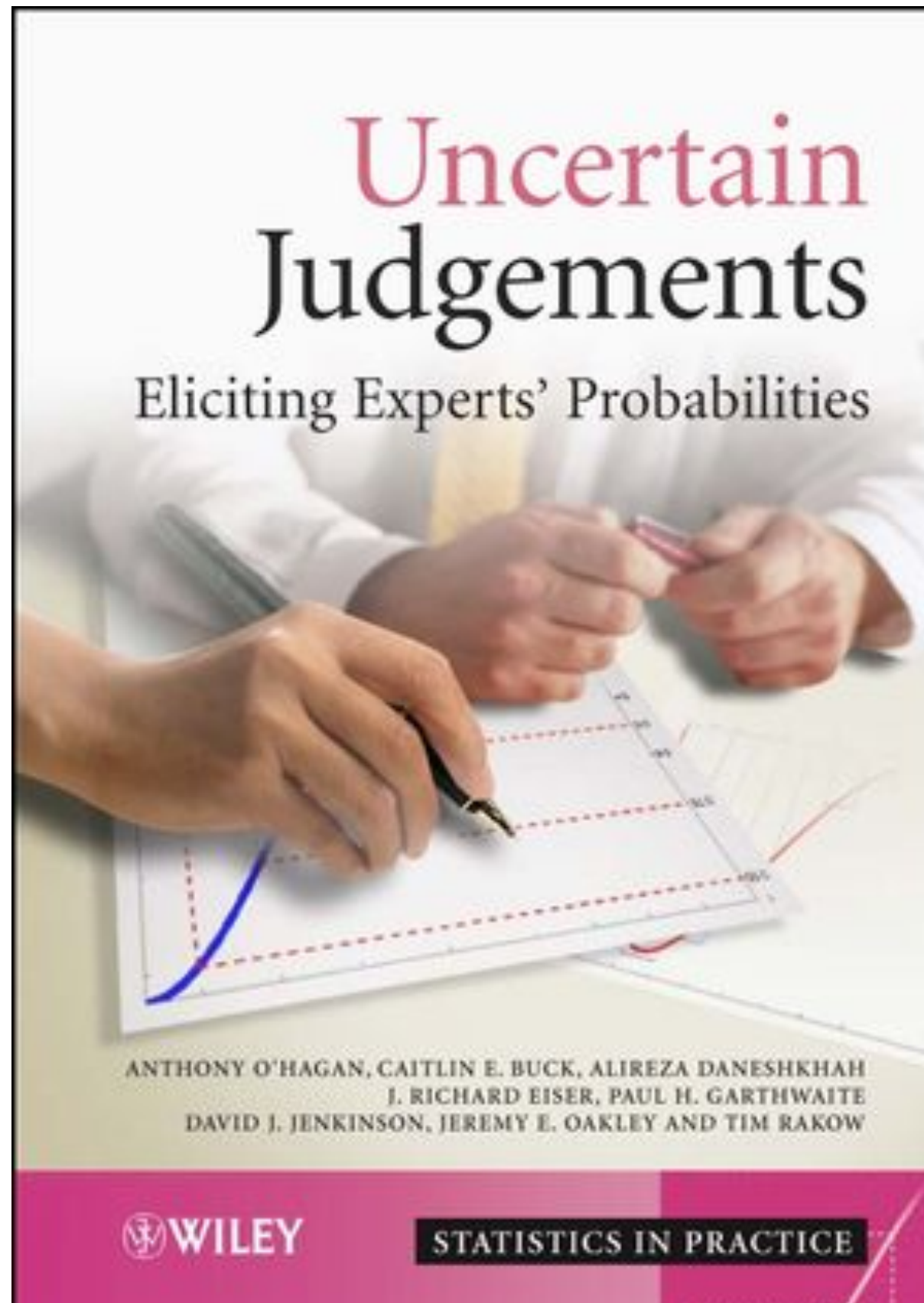
$$X \sim \text{Normal}(\mu, \sigma)$$

1. Define **priors** for the parameters (here,  $\mu, \sigma$ )
2. Derive **posterior distribution** of the parameter(s) of interest using Bayes' rule:

$$\underset{\text{posterior}}{f(\mu \mid \text{data})} \propto \underset{\text{likelihood}}{f(\text{data} \mid \mu)} \times \underset{\text{prior}}{f(\mu)}$$

3. Carry out inference based on the posterior

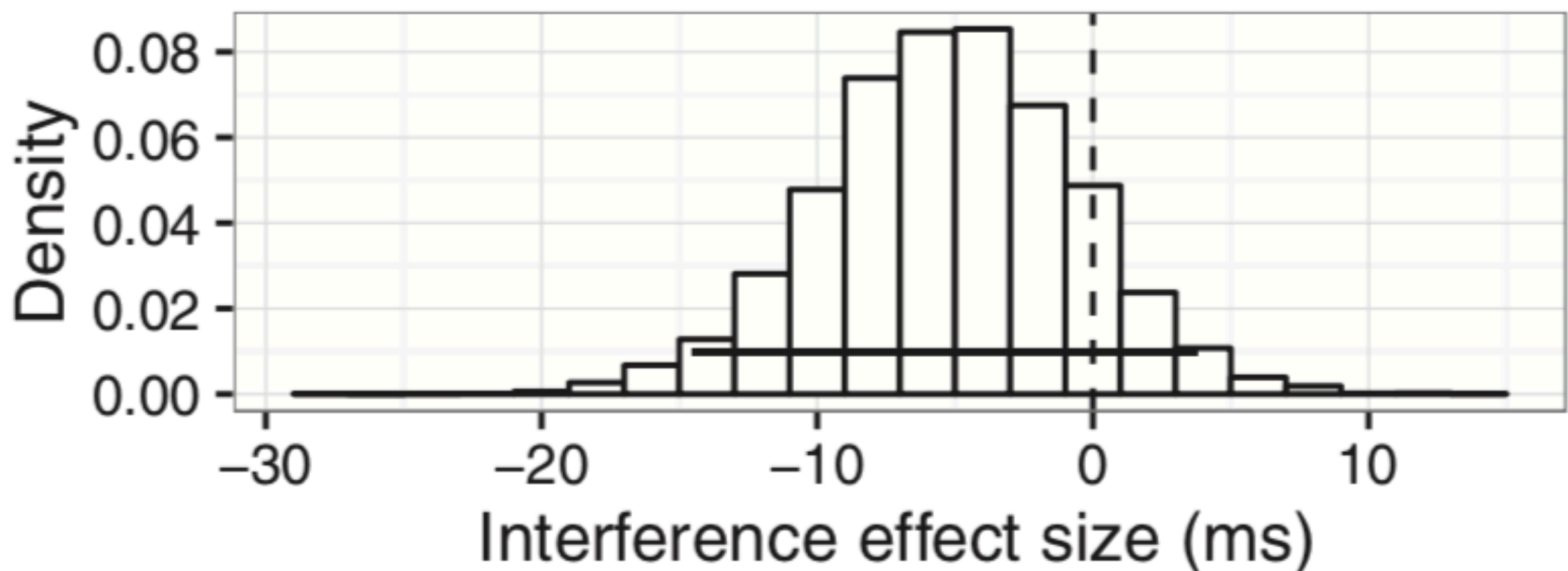
Deriving prior distributions requires domain knowledge



# The Bayesian approach

The end result of a Bayesian analysis is a posterior distribution of the parameter of interest

**Agreement attraction effect  
(meta-analysis estimate)**





# Comparison of Frequentist vs Bayesian approaches

	Frequentist	Bayesian
Parameters	Fixed	Random*
Data	Random	Fixed
Prior knowledge used	No	Yes
Type I, II error	relevant	irrelevant**
Hypothesis testing	reject null	Bayes factor
Uncertainty quantification	No***	Yes

\* Random variables

\*\* Type I, II error could be seen as relevant for Bayes

\*\*\* Confidence intervals can be a proxy



# ANOVA vs Bayes factors

$$BF_{12} = \frac{\textit{Likelihood}(\textit{Data} \mid \textit{Model}_1)}{\textit{Likelihood}(\textit{Data} \mid \textit{Model}_2)}$$

The Bayes factor is similar to the frequentist likelihood ratio test (or ANOVA), with the difference that in the Bayes factor, the likelihood is integrated over the parameter space.

The BF can be highly sensitive to the priors.  
**A Bayes factor analysis must come with a sensitivity analysis.**

# Why is the Bayesian approach useful?

1. Handles sparse data without any problems
2. Highly customised models can be defined
3. The focus is on **uncertainty quantification**

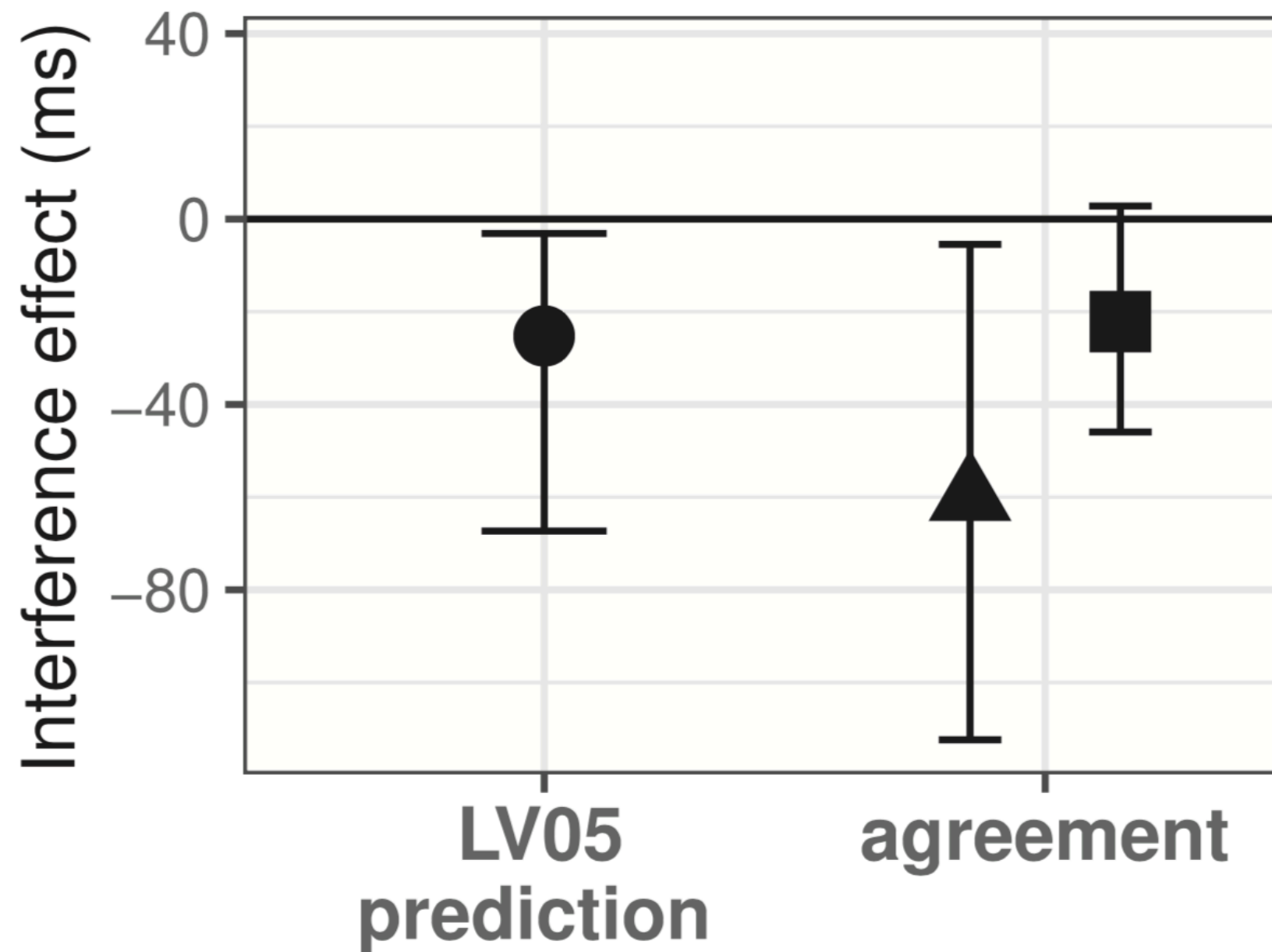
## But Bayes comes with a cost

1. You have to think about your prior knowledge/belief
2. There is no one answer corresponding to  $p < 0.05$  or  $p > 0.05$ .
3. You have to learn to think about uncertainty

**Is replication important  
in Bayesian data analysis?**

Let's look at a concrete example.

# Example: a high-powered replication attempt of the agreement attraction effect



● LV05    ▲ Dillon et al., 2013 (N=40)    ■ Replication (N=181)

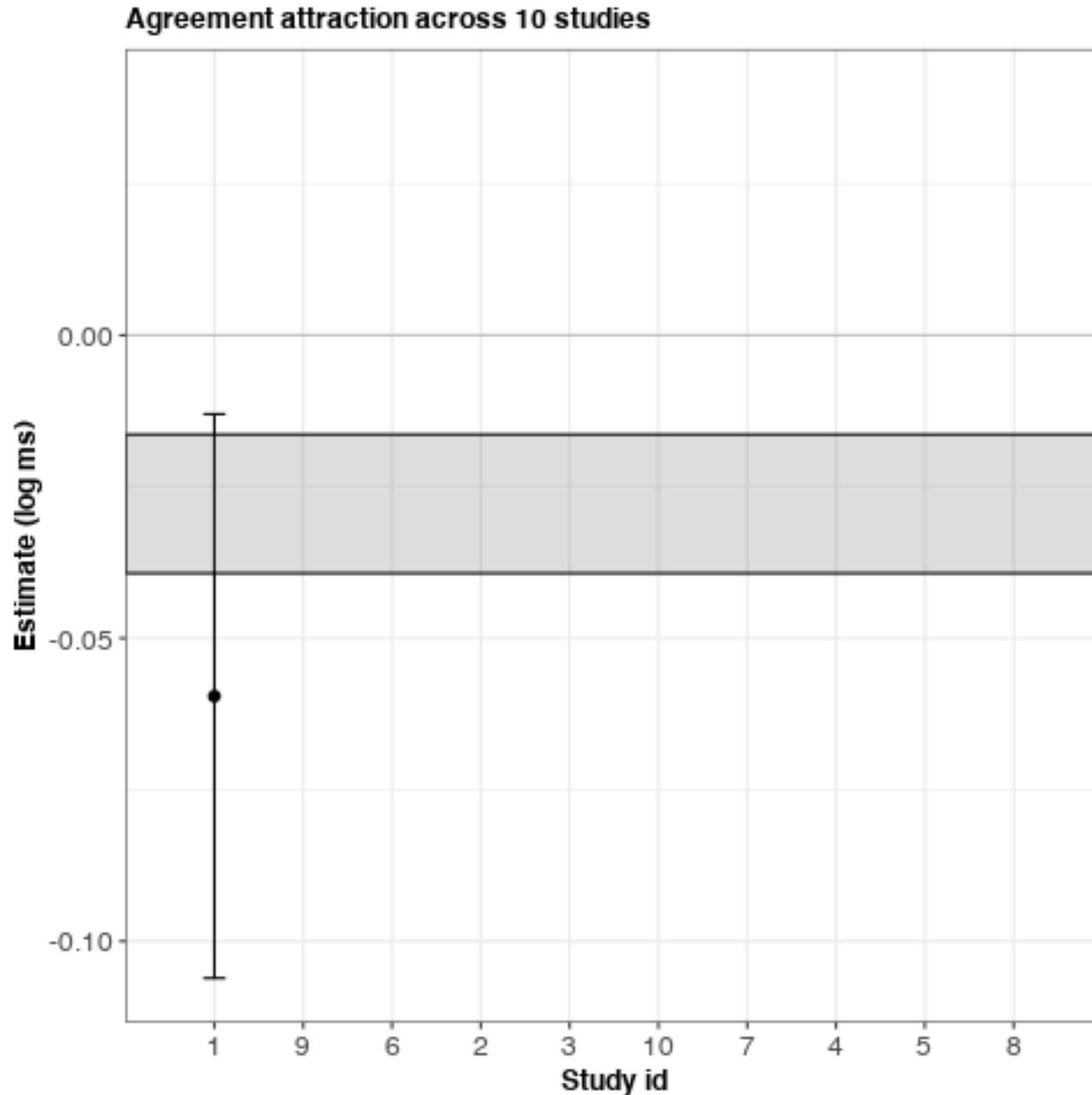
# Example of a replication attempt of a low-powered study

1. The model prediction (LV05) is quite constrained
2. The  $N=40$  agreement estimate has very wide uncertainty but shows a significant effect
3. **Question:** Does the  $N=181$  study replicate the significant agreement effect from  $N=40$ ?

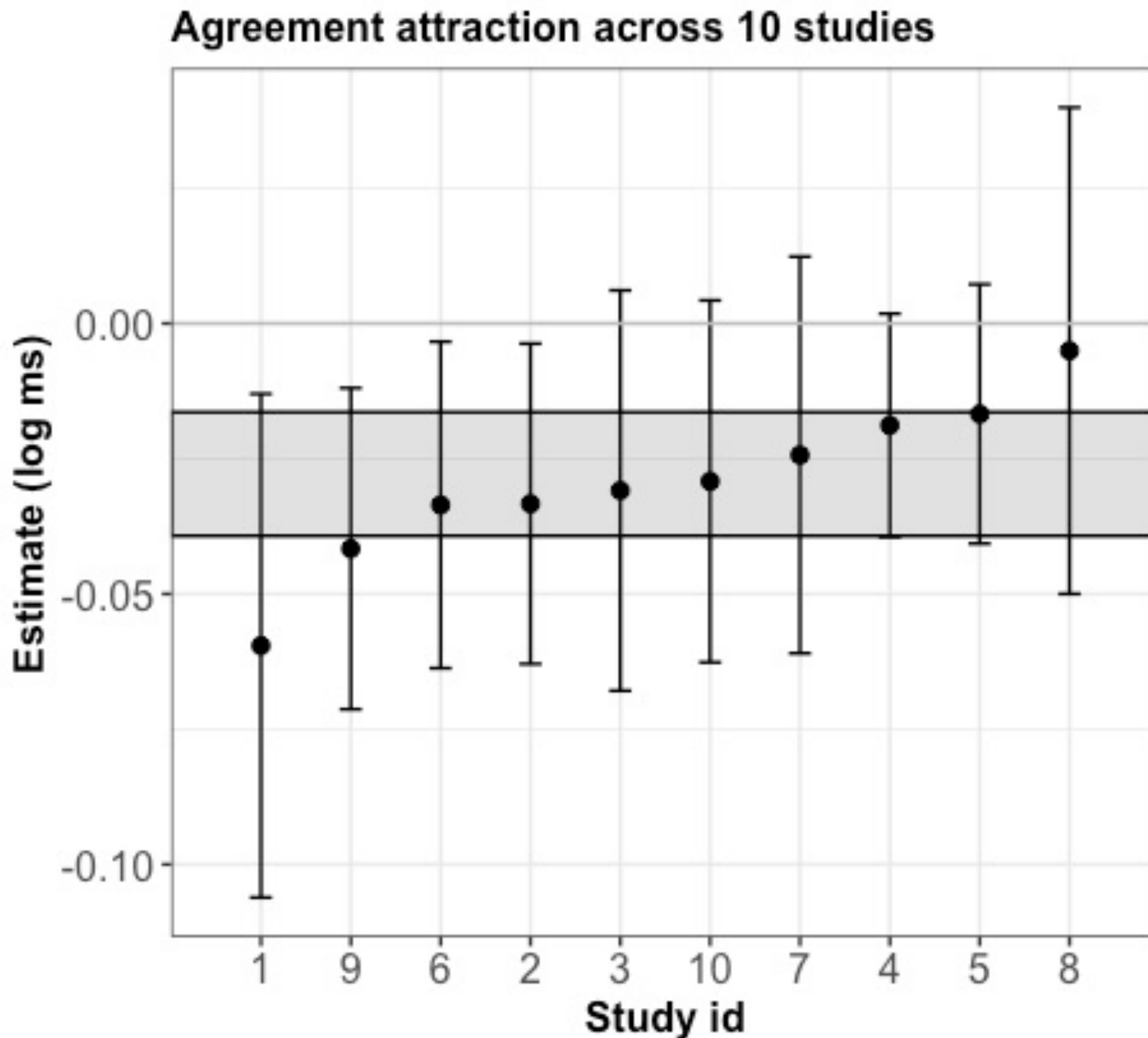
**Frequentist answer:** no

**Bayesian answer:** yes, because the original estimates have a very wide 95% confidence interval!

We can use frequentist confidence intervals to quantify uncertainty

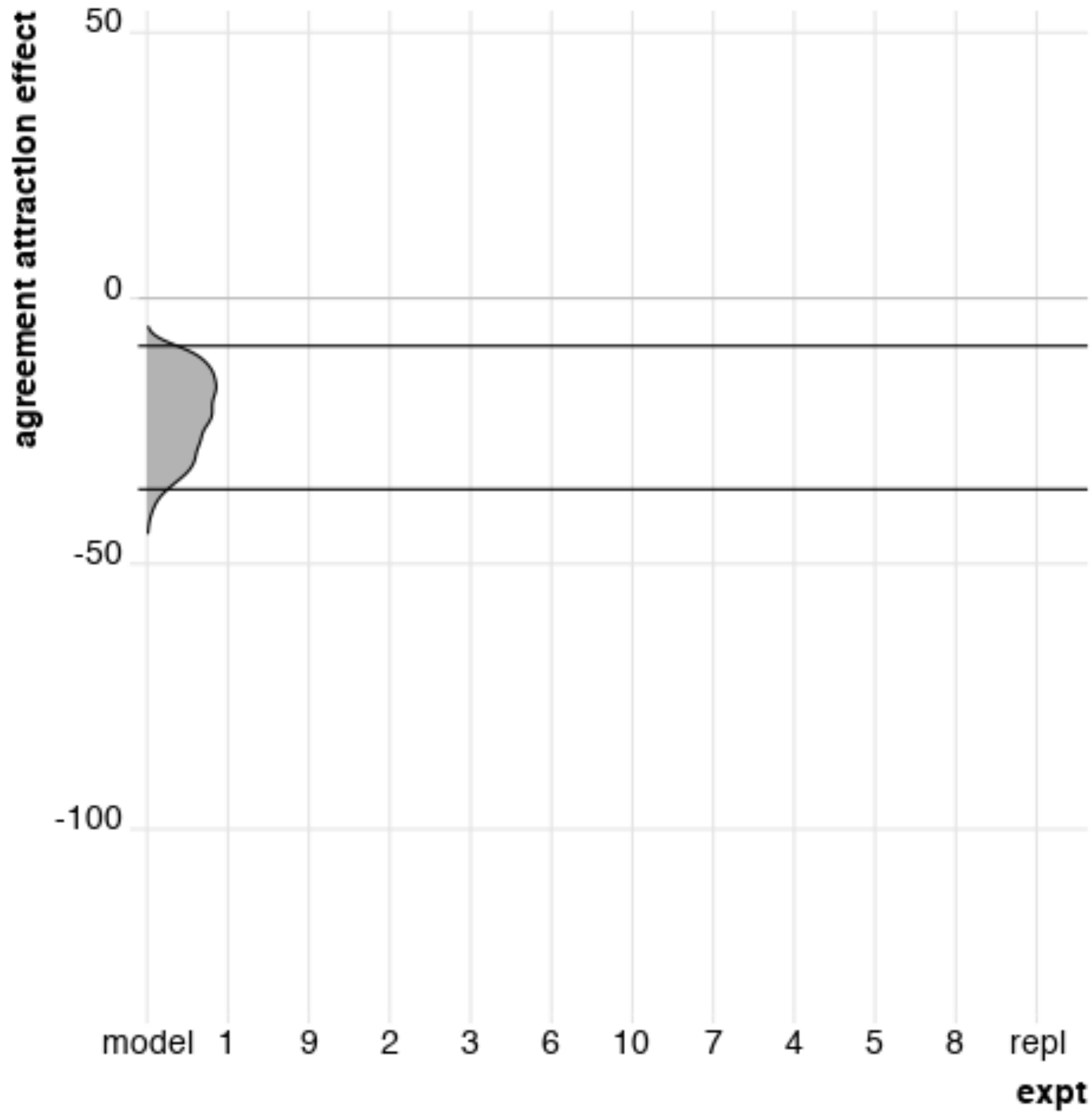


We can use frequentist confidence intervals to quantify uncertainty

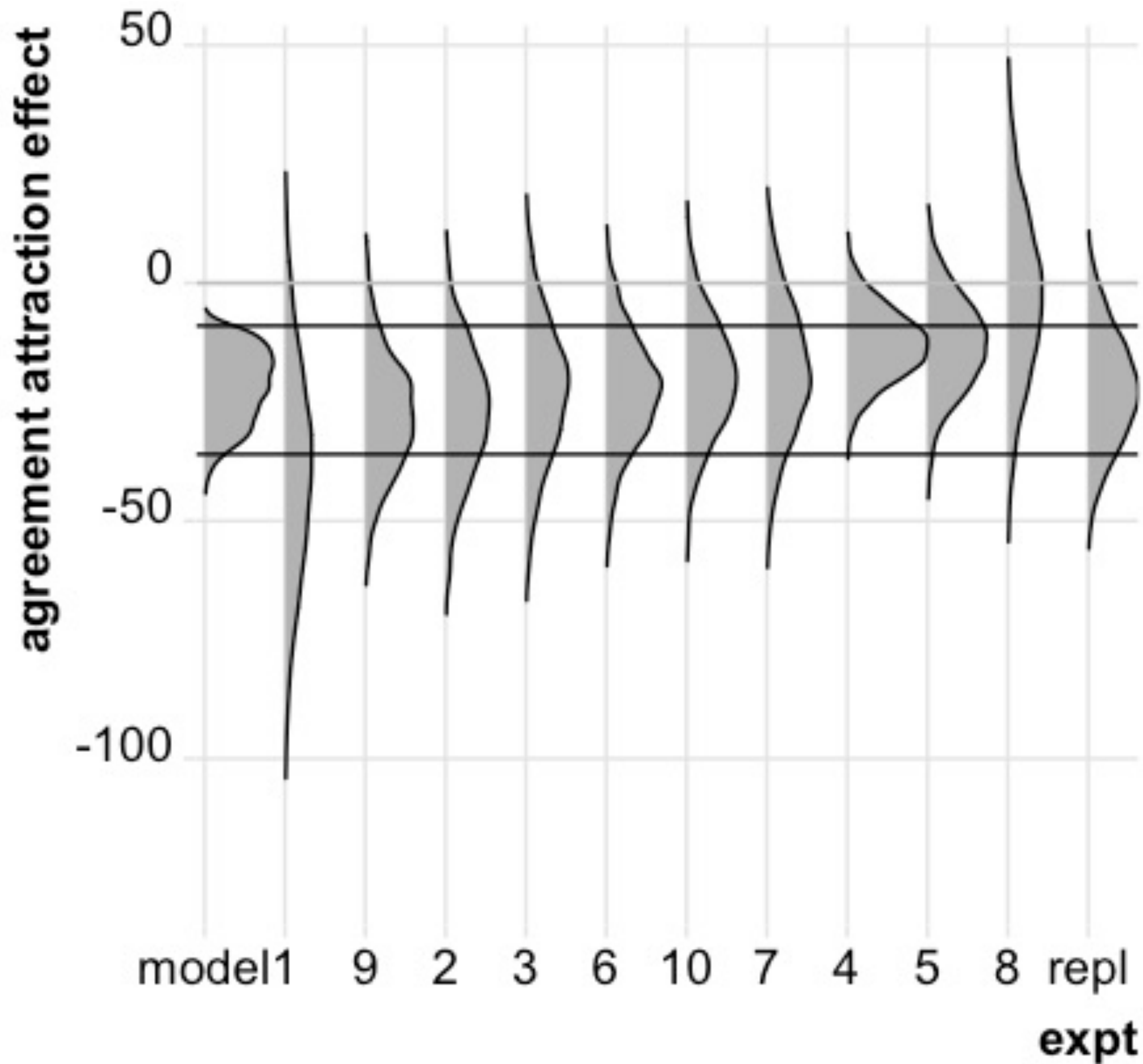




# Bayesian approach to studying replicability



# Bayesian approach to studying replicability



## Concluding remarks

1. Replicability is equally important in Bayes.
2. Low-powered designs will be misleading (NHST) or uninformative (Bayes).
3. With Bayes, one can accumulate evidence and use it in future studies.
4. We can establish replicability by (a) showing consistency across studies, (b) accumulating evidence through meta-analyses, (c) running higher-precision studies **informed by accumulated knowledge**.
5. Bayes provides the machinery for achieving these goals.

# Example of accumulating knowledge

Nicenboim, Vasishth, Rösler (under review)

It was a windy day.

The boy went out to fly a/an \_\_\_\_\_

DeLong et al 2005, *Nature Neuroscience*

Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data

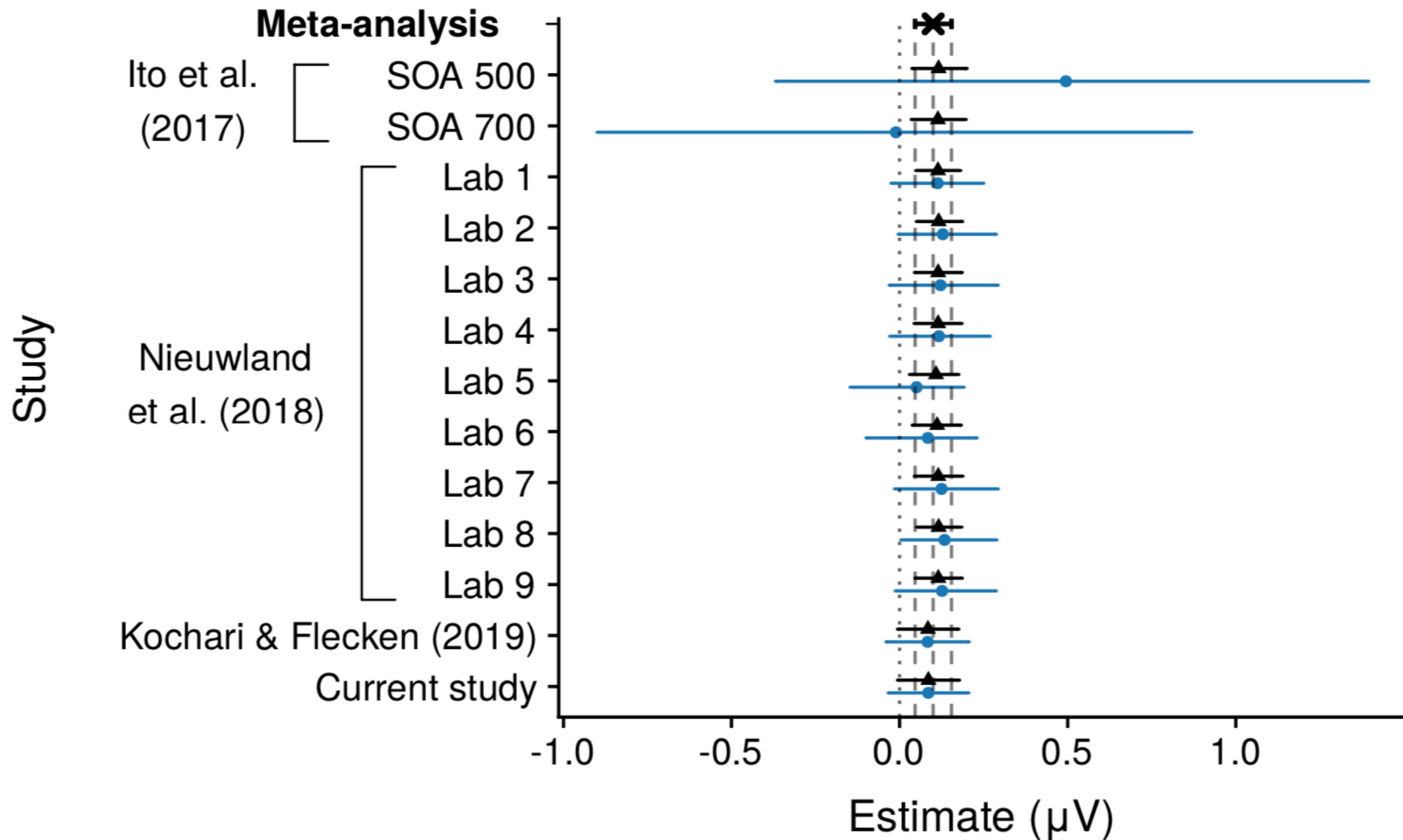
Bruno Nicenboim<sup>1</sup>, Shravan Vasishth<sup>1</sup>, & Frank Rösler<sup>2</sup>

<sup>1</sup> University of Potsdam

<sup>2</sup> University of Hamburg

# Example of accumulating knowledge

Nicenboim, Vasisht, Rösler (under review)

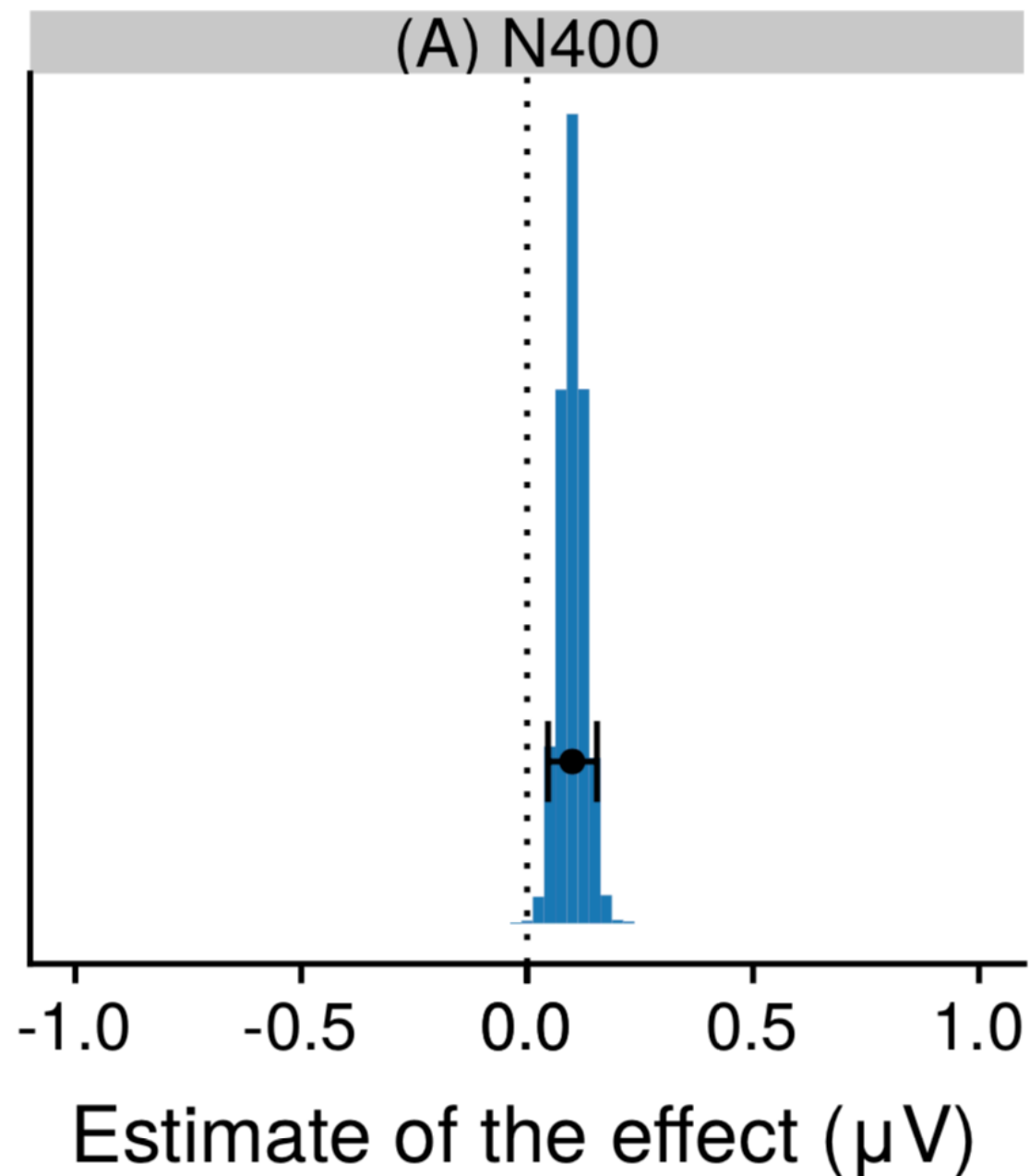


# Example of accumulating knowledge

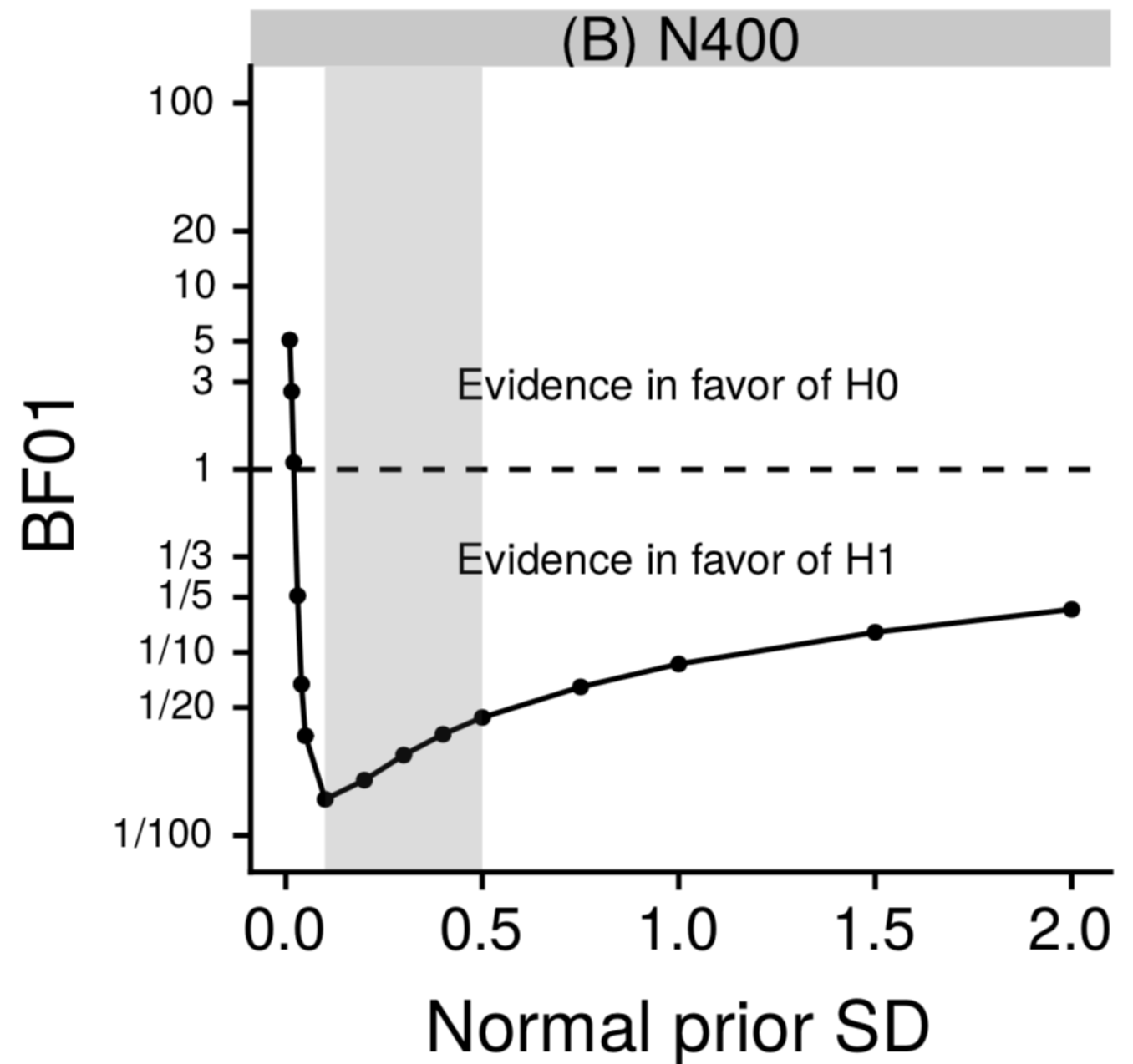
Nicenboim, Vasishth, Rösler (under review)

## Meta-analytic estimate

### Posterior distribution



### Bayes factors



All code and data available from:

<https://osf.io/p8amv/>